Multiple Person Tracking based on Spatial-temporal Information by Global Graph Clustering

Yu-ting Su[1], Xiao-rong Zhu[1] and Wei-Zhi Nie [1]*
[1] School of Electronic Information Engineering, Tianjin University
Tianjin, 300072, China
[e-mail: weizhinie@tju.edu.cn]
*Corresponding author: weizhinie@tju.edu.cn

Received December 15, 2015; revised April 17, 2015; accepted May 7, 2015;
published June 30, 2015

Abstract

Since the variations of illumination, the irregular changes of human shapes, and the partial occlusions, multiple person tracking is a challenging work in computer vision. In this paper, we propose a graph clustering method based on spatio-temporal information of moving objects for multiple person tracking. First, the part-based model is utilized to localize individual foreground regions in each frame. Then, we heuristically leverage the spatio-temporal constraints to generate a set of reliable tracklets. Finally, the graph shift method is applied to handle tracklet association problem and consequently generate the completed trajectory for individual object. The extensive comparison experiments demonstrate the superiority of the proposed method.

Keywords: Multiple person tracking, object tracking, multiple cameras tracking, graph clustering

This work was supported in part by the National Natural Science Foundation of China (61472275, 61170239), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

http://dx.doi.org/10.3837/tiis.2015.06.014 **ISSN : 1976-7277**

1. Introduction

Multiple object tracking is a popular research point in recent years and it is widely used in many applications such as intelligent video surveillance, gesture and action recognition [1, 2, 3, 4, 5], and events recognition. However, it is hard to guarantee the reliability of tracking by the irregular changes of human shapes, the partical occlusions and the variations of scene illumination [6].

Over the past two decades, a large number of different tracking algorithms [7, 8] have been proposed to handle tracking problems. Traditional trackers such as mean-shift, Kalman, particle filter [9, 10, 11] can be seen as a process of optimization. They need multiple iterations to find the best location of each tracking object. However, these traditional tracking methods are more suitable for single person tracking. In recent years, tracking-by-detection methods have became increasingly popular [12, 13, 14, 15]. Shu et al [16] applied an online learning method to track objects, where they trained a decision model by SVM for each tracking object. Their method could judge the disappearance condition of object, but they need to update the decision model in the tracking process, which leads to the additional computational overhead. Nie et al [8] proposed to apply DPM [17] to detect moving people in each frame and utilized Baysian model to handle tracklet association problem, which can address occlusion problem in tracking process. However, this kind of method relies too much on the performance of detector and dataset association. For the detection results, several existing approaches address this issue by linking detections with confidence to build tracklet in order to improve the performance of detection [18]. For the dataset association problem, some existing methods apply Conditional Random Field (CRF), Byasian Model, and Convex optimization to handle. However, in these methods, the spatio-temporal information of tracked object is not fully utilized.

In order to solve these problems, we propose a graph model based on spatio-temporal information of tracking objects for object tracking. First, we employ DPM [17] to detect people in each frame. Then, spatial and temporal constraints help us to generate reliable tracklets through these detection results. Each tracklet includes a set of detection results. We apply mean pooling method to fuse all of detection results and extract Hog and HSV features for tracklet representation. Finally, we utilize spatial and temporal information of tracklets to structure graph model, and leverage graph clustering method to handle data association problem for trajectory generation.

The main contributions are two-fold:

- We formulated tracklet association problem into one clustering problem and applied graph clustering method to address it;
- The proposed method is suitable for different scenes. It can also be used in multiple viewing scenes for object tracking.

The remainder of this paper is organized as follows. Section 2 will introduce the state-of-the-art methods in visual tracking related to this paper. System overview and some details of the proposed method will be introduced in Section 3. The experimental results are showed in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

A vast amount of work has been published on multiple person tracking [19, 20, 21, 22]. Tracking-by-detection is a popular tracking algorithm in the last decade. It associates the detection candidates with spatio-temporal constraints to generate a set of tracklets, then applies data association algorithms to get the final trajectory. Shu et al [16] proposed an online learning method to track multiple people and applied greedy algorithm to handle the data association problem. However, online learning did not provide a stable detected result, and false samples will bring catastrophic effects. Mohamed et al. used HoG [23] feature as the detector in the tracking system [24]. However, the performance of HoG feature is not good in low resolution videos. A large number of works [25, 26, 27, 28] have showed that good detected results will greatly help the tracking process. So we apply part-based model to detect person in each video frame aiming to improve the detected accuracy. Yang et al's [29] approach was similar to ours, but they used a classifier to predict the potential positions of the tracking person, while we first use optical flow to predict motion regions, then apply spatial and physics information to get the final predicted region as the detected result. Zdenek et al [7] proposed a TLD tracking algorithm, which used detection+tracker model to detect objects. However, in the tracking process, Zdenek et al. applied online learning classifier to judge the detected results. Inspired by this work, features were extracted from the average image of detection windows as the whole feature of tracking object aiming to find a kind of more stable feature to represent the tracked object.

Classic multiple object trackers such as multi-hypothesis tracking [30] and joint probabilistic data association filters [31] jointly consider the data association from sensor measurements to multiple overlapping tracks. While not restricted to Markov chains, they are able to only keep few steps in memory due to the exponential task complexity and they do not take physical exclusion constraints between object volumes into account. Jiang et al. [32] employed integer linear programming to handle the data association problem. However, the number of objects need to be known priori. To overcome this limitation, Berclax et al. [33] introduced virtual source and sinked locations to initiate and terminate trajectories. A common trait of these works is that they lead to global optimization problems, which are usually solved by linear programming. However, the global optimization problem must consider the kinds of conditions such as entering, leaving and occlusion of tracking objects. These conditions will increase the complexity and calculation of the algorithm. We draw on the idea to handle occlusion problem by linear programming in this paper. Our method is different from [33] in a way that we do not apply global optimization to handle the trajectory problem, while we use greedy algorithm to generate the final trajectory for each tracking object.

3. System Overview

Our tracking system includes two steps. 1) Tracklet generation: in this stage, we apply DPM method to detect each person in each frame from video sequence. DPM is a popular detection model, which can localize the body parts of one person by dynamic programming with the visual features. This advantage is very useful to handle the occlusion problem. Then, we utilize spatial and temporal constrains to generate a set of reliable tracklets [18]. 2) Data association: each tracklet is constituted by a set of detection results. These detection results represent the same person. We apply mean pooling method to generate average image from these detection results, and extract color [34] and HoG [35] feature to represent tracklet.

Finally, clustering method is used to handle data association problem and generate the final trajectory for each tracking object.

Data association is a key step in this work. Thus, we will further explain this process in section 3.1. First, we will introduce the process of feature extraction from average image. Then, the process of graph building can be introduced. Finally, we will show the graph clustering method.

3.1. Data Association

3.1.1. Feature Extraction

After the tracklet generation, we have generated a set of reliable tracklets. Each tracklet is represented by a set of detected results. Obviously, these detected results belong to one same person.

We assume that each tracklet has N detected results $T = \{t_1, \ldots t_N\}$. We are able to get the average image from these detected results shown in **Fig. 1**. The average image shows that these detected results have similar feature representations such as color or edge feature. The trajectory of each person in a video sequence must be formed by a set of tracklets, while these tracklets must have similar feature representations since one moving person should represent an average motion specially in a long time movement. So we extract color and Hog features from the average image as the feature of tracklet. If some tracklets belong to one person, these similarities of these tracklets should be higher than other tracklets.

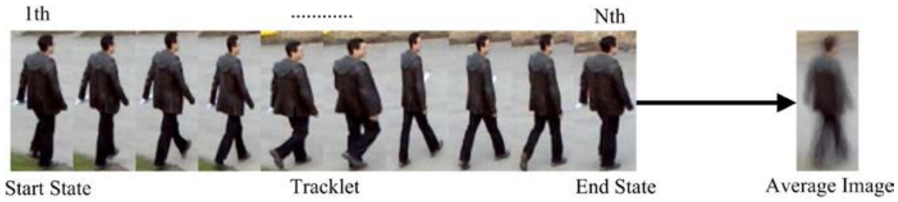

Fig. 1. The information of one tracklet.

3.1.2. Graph Structure

In **Fig. 2**, we record the initial and the terminational state of one tracklet including the positions and spatio-temporal information. We then use $t_i = \{n_s, n_e, f, x_s, y_s, x_e, y_e\}$ to represent each tracklet. Here n_s is the initial time of tracklet, n_e is the terminational time of tracklet, f is the feature of tracklet, and (x_s, y_s) is the initial position of tracklet. (x_e, y_e) is the terminational position of tracklet. These information will be used to compute similarity between two different tracklets as follow:

$$S(i,j) = p(n_{ie}, n_{js})(\varepsilon d(i,j) + \tau h(i,j)). \quad (1)$$

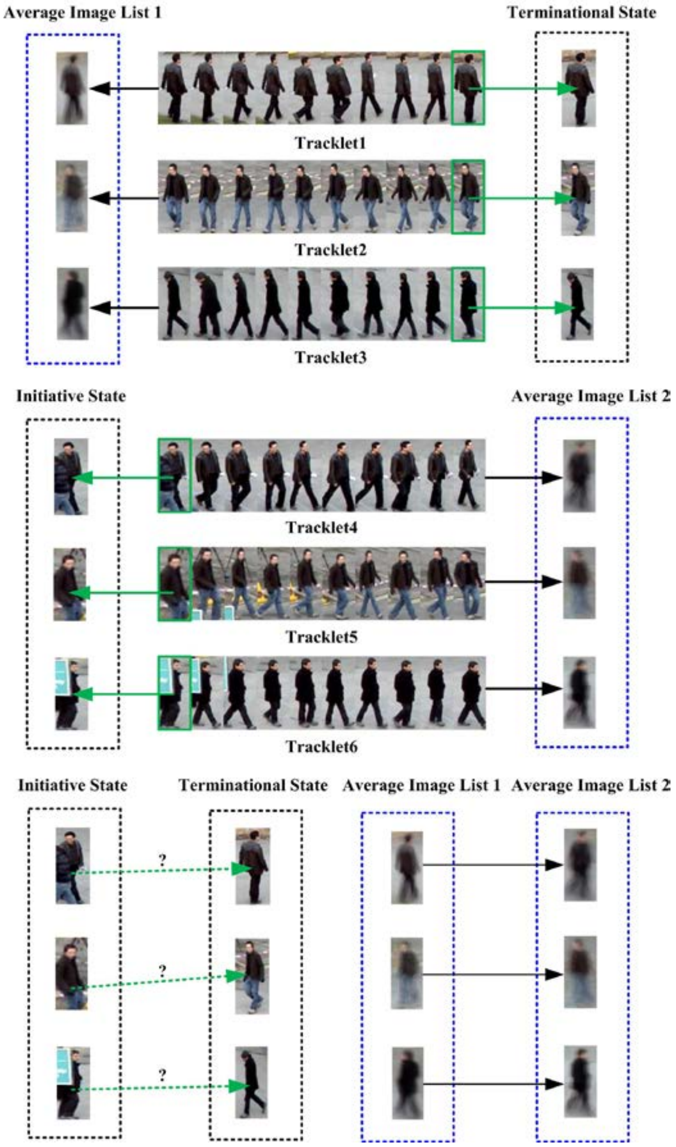

Fig. 2. The sample of matching tracklets. The traditional methods often use matching result between initial and terminational state as the matching result between two tracklets. Our approach applies matching result between two average image from different tracklets as the final matching result.

where $S(i, j)$ is the similarity between tracklet i and tracklet j, $p(n_{ie}, n_{js})$ is the relationship between two different tracklets, and it is computed by Eq 2, $d(i, j)$ and $h(i, j)$ are the similarities in distance and feature, which are computed by the Eq.3 and Eq.4, ε and τ are the weights of $d(i, j)$ and $h(i, j)$. In this work, we set $\varepsilon = \tau = 0.5$.

$$p(n_{ie}, n_{js}) = \begin{Bmatrix} 1, & if & n_{ie} \leq n_{js} \\ 0, & if & n_{ie} > n_{js} \end{Bmatrix}. \tag{2}$$

$$d(i,j)=\frac{\sqrt{(x_{ie}-x_{js})^2+(y_{ie}-y_{js})^2}}{\sum_{j=1}^{N}\sqrt{(x_{ie}-x_{js})^2+(y_{ie}-y_{js})^2}}\ . \tag{3}$$

where (x_{ie}, y_{ie}) is the terminational state position of tracklet i in video frame, (x_{js}, y_{js}) is the initial state position of tracklet j in video frame. We use Euclidean distance to represent the similarity between tracklet i and tracklet j in spatial terms.

$$h(i,j)=\frac{\sqrt{(f_i-f_j)\cdot(f_i-f_j)'}}{\sum_{j=1}^{N}\sqrt{(f_i-f_j)\cdot(f_i-f_j)'}}\ . \tag{4}$$

where f_i is the feature of tracklet i, f_j is the feature of tracklet j. We apply Eq.4 to compute the similarity between tracklet i and tracklet j in feature space. Finally, we fuse both similarities in feature and spatial by Eq.1 to obtain the final similarity between the tracklet i and the tracklet j. The similarity can be used to build graph $G=(V,E)$, where, V is the node set, E is the edge set. Each node V_i denotes one tracklet. Each edge e_j denotes relevance between different tracklets. The weight of edge is computed by Eq.1.

3.1.3. Tracklet Association

Based on the graph model, the dense subgraphs are considered as the trajectory for one person. To detect the dense subgraphs, we applied the pair-wise clustering method of graph shift (GS) [35], which is a popular method in detecting dense subgraph. The input of the GS method is the adjacency matrix A of the multimedia. A subgraph is represented by a probabilistic cluster $x \in \Delta^n$, where $\Delta^n=\{x \mid x\in R^n, x\geq 0, |x|_1=1\}$. In fact, x is a unit mapping vector, which is the probability that the subgraph contains each vertex. Particularly, $x_i=0$ means that the i th vertex is not included in the dense subgraph. The GS method measures the average connection strength of subgraph x^* by $g(x)$ in Eq.5 and efficiently finds all the local maximums x^* of $g(x)$. Each local maximum indicates a dense subgraph of the graph, which is defined as a trajectory for each person.

$$g(x)=x^T Ax\,. \tag{5}$$

Finding the dense subgraph is equivalent to maximizing $g(x)$ s.t. $\sum x_i=1$. Since our aim is to find the dense subgraph, each time we only consider one subcluster. Without loss of generality, we denote the sub-cluster as x_i. Since the problem is a constrained optimization problem, by adding Lagrangian multipliers λ for all $i=2,\cdots,n$, we obtain its Lagrangian function:

$$L(x) = g(x) - \lambda(\sum_{i=1}^{m} -1) . \tag{6}$$

Any local maximizer x^* must satisfy the Karush-Kuhn-Tucker (KKT) condition. That is:

$$\begin{gathered} f_i(x^*) - \lambda = 0 \\ \sum_{i=1}^{m} \lambda(x_i^* - 1) \geq 0 \end{gathered} \tag{7}$$

where $f_i(x^*) = \dfrac{\partial g(x^*)}{\partial x_i}$.

As pointed out by [36], we can optimize the problem in the pairwise way by Algorithm 1.

```
**Algorithm 1** Dense subgraph
**Input:**
Weighted adjacency array A and an initialization $x(0)$;
Set $x = x(0)$;
**Repeat:**
1: Update the partial derivative $g(x)$ with respect to each variable $x_i$;
2: Find a pair $(v_i, v_j)$
3: Compute the best step size to maximize Eq.5;
4: **Until:** $x$ is a local maximizer;
5: **Output**: A KKT point $x^*$
```

After this process, we have a set of sub-graphs or sub-clusters. Each sub-graph can be seen as one trajectory for the tracked object.

4. Experimental

4.1. Datasets

We show experiments on four video sets to demonstrate the effectiveness of the proposed method. The first set is a selection from the *Pets* 2012 *video datasets*, which is captured with a stationary camera. In this dataset, 7 cameras record several pedestrians under various angles. Among these cameras, 4 of them are located relatively close to the scene, and captured clear pictures of people. The other 3 cameras are located further from the scene and about 4-5m above the ground, giving a wide angle view of the situation. The frame rate for all cameras is about 7 fps.

The second set is the *Town Center Datasets*, the frame resolution is 1920×1080, and the frame rate is 25 fps. In this dataset, the long-term occlusion often appears and the motion of pedestrians is often linear and predictable. The biggest advantage of this dataset is high resolution, which is useful for our detection algorithm.

The third set is *Caviar video sequence*, which includes 26 video sequences of a walkway in a shopping center taken by a single camera with frame size of 385×288 and the frame rate of 25 fps. At the same time, the video datasets include two angles (front and corner). We selected corner angle to test the effectiveness of our method because more occlusion happen in this angle.

4.2. Experimental Setting

In our implementation, we used the color and Hog features to represent each tracking object. We averagely segmented the detected window into eight parts to extract the features. The feature vector for each part consists of 256-bin RGB color histogram using 10 bins for each channel, 36-D HoG feature and 36-D color feature for each part region. We also applied normalization for each part and concatenated features from all eight parts into one feature vector. To improve the accuracy of human detection, we first implemented Gaussian Mixture Model [37] to get motion regions in each video frame and then applied part-based human detector to detect human body in the foreground. We will detail these experiment results in the next sub-section.

We evaluated our tracking results using the standard CLEAR MOT metrics [38]. TA (tracking accuracy), DP (detection precision) and DA (detection accuracy). DA and DP are used to test the detection result. Higher score means better results. TA is used to evaluate the experimental results of tracking. The the higher score the better. However, TA is computed by the number of lost targets. The difference between TA and DA can be used to evaluate the matching result. The lower score the better.

4.3. Experimental Results

PETS 2012 Datasets: The PETS 2012 datasets include 8 video sequences. The resolution of these video sequences is 768×576. The experimental datasets are challenging due to the existence of occlusions, crowded scenes, and cluttered backgrounds. All of video sequences are from one same scene in different angles, so the dataset also can be seen as a multiple cameras tracking dataset. We apply the dataset to test the performance of our new feature extraction. In the evaluation step, we use MOTA, MODA and MODP to evaluate the performance of tracking results. We compare our method with [8]. Nie's method is similar with ours in detection step. So we apply these comparative results to demonstrate the performance of our approach. The tracking results are showed in the **Table 1**. **Fig. 3** shows some tracking results in single camera scene.

Table 1. Tracking Results on PETS 2012 View001 Dataset

Method	MOTA	MODA	MODP
Nie et al[28](%)	72.4	72.8	75.8
Our approach (%)	75.1	75.6	75.6

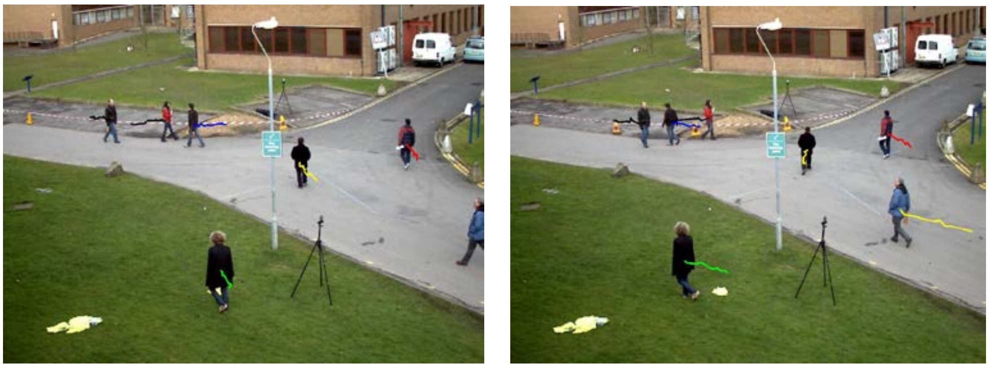

Fig. 3. Tracking results in PETS 2012 View001 dataset

From these experimental results, we could find that the tracking results are better than the prior works. The frame rate of the PETS 2012 dataset is 7 fps. So the change of human shapes is bigger in the consecutive frames than other video sequences. This condition will lead that the detection results disappear. Finally, we will get a lot of tracklets. The experimental results prove better performance of our approach.

Town Center Dataset: We also test our approach in the Town Center dataset. The resolution of this dataset is 1920×1080 and the frame rate is 25 fps. This dataset contains the street scene with long-term occlusions. At the same time, the high resolution of this dataset is very useful for the Part-based model. We compared the results with the recently proposed methods [39, 40, 36, 41, 42, 16]. With the same experimental setting, the performance of our matching method is better than others in MOTA and MODA. The experiment results are shown in the **Table 2**. **Fig. 4** shows some tracking results in single camera scene.

Table 2. Tracking Results on Town Center Dataset

Method	MOTA	MODA	MODP
Benfold et al.[3](%)	64.8	64.9	80.5
Zhang et al.[38](%)	65.7	66.1	71.5
Pellegrini et al.[31](%)	63.4	64.1	70.8
Yamaguchi et al.[36](%)	63.3	64.0	71.1
Leal-Taixe et al.[21](%)	67.3	67.6	71.6
Shu et al.[34](%)	72.9	73.5	71.4
Our approach(%)	74.3	74.8	71.6

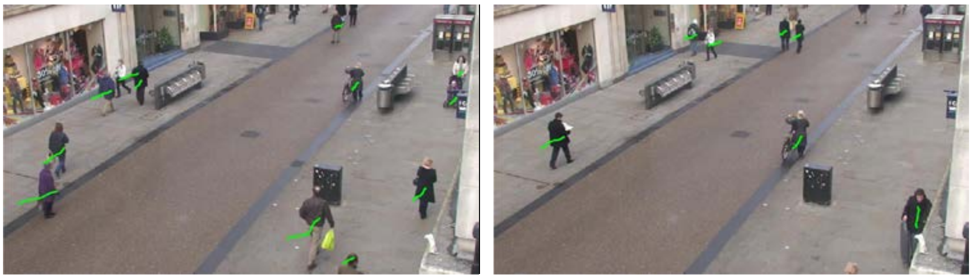

Fig. 4. Tracking results in Town Center dataset

From these results, it can be observed that the detection results are similar with the results of [8]. Meanwhile, our results are also better then HoG detector. We also observe that our tracking results are better than the prior work. Our feature extraction method could provide the better discrimination between different tracking objects. The experimental results also prove the performance of the proposed method.

The CAVIAR Datasets: We also tested our approach on CAVIAR datasets, which includes 26 video sequences and every video sequence often express one events. The resolution of the dataset is 384×288 and the frame rate is 25 fps. In application, we selected 18 video sequences from these 26 video sequences to test our method. The CAVIAR dataset is a multiple cameras datasets, and every video sequence includes two view angles videos. One view angle is a corner angle, and the other angle is front angle. We selected 16 corner angle video sequences because there are more occlusion conditions in these video sequences. The tracking results are showed in **Table 3**.

Table 3. Tracking Results in CAVIAR Dataset

Method	MOTA	MODA	MODP
Nie et al[28](%)	73.3	74.9	66.1
Our approach (%)	75.5	77.1	66.9

In this dataset, occlusion often appears in these video sequences. The resolution of this dataset is low, so the detection results did not have an obviously improvement. However, the feature of an average image brings an improvement in tracking result. However, this dataset also has some disadvantages such as occlusion often appears between two people and the time of occlusion is not long. In other words, this dataset is designed for events detection. In most video sequences, there are only 1 to 3 people appear. In these special dataset, our approach could get a good tracking result.

5. Conclusion

In this paper, we proposed a data association method based on graph clustering to handle object tracking problem. We leverage visual, spatial, and temporal information of detection results to generate the reliable tracklets. Based on these tracklets, we built the graph model and successfully formulated the data association problem into one graph clustering problem. Then, graph shift method is utilized to handle clustering problem and generate the final trajectory for each tracking object. Experiment results demonstrate the superiority of this method.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61472275, 61170239), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

References

[1] An-An Liu, Y-T Su, P-P Jia, Zan Gao, Tong Hao, and Z-X Yang, "Multipe/single-view human action recognition via part-induced multitask structural learning," *Cybernetics, IEEE Transactions on* , pp. 1, 2014. Article (CrossRef Link)

[2] Zan Gao, Hua Zhang, An-An Liu, Yan-bing Xue, and Guang-ping Xu, "Human action recognition using pyramid histograms of oriented gradients and collaborative multi-task learning," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 8, no. 2, pp. 483–503, 2014. Article (CrossRef Link)

[3] An-An Liu, "Bidirectional integrated random fields for human behaviour understanding," *Electronics letters*, vol. 48, no. 5, pp. 262–264, 2012. Article (CrossRef Link)

[4] An-An Liu, "Human action recognition with structured discriminative random fields," *Electronics Letters*, vol. 47, no. 11, pp. 651–653, 2011. Article (CrossRef Link)

[5] An-An Liu, Kang Li, and Takeo Kanade, "A semi-markov model for mitosis segmentation in time-lapse phase contrast microscopy image sequences of stem cell populations," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 359– 369, 2012. Article (CrossRef Link)

[6] Weizhi Nie, Anan Liu, Yuting Su, Huanbo Luan, Zhaoxuan Yang, Liujuan Cao, and Rongrong Ji, "Single/cross-camera multiple-person tracking by graph matching," *Neurocomputing*, vol. 139, pp. 220–232, 2014. Article (CrossRef Link)

[7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, 2012. Article (CrossRef Link)

[8] W. Nie, A. Liu, and Y. Su, "Multiple person tracking by spatiotemporal tracklet association," *In AVSS*, pp. 481–486, 2012. Article (CrossRef Link)

[9] J. Tu, H. Tao, and T. S. Huang. "Online updating appearance generative mixture model for meanshift tracking," *In ACCV* , vol. 1, pp. 694–703, 2006. Article (CrossRef Link)

[10] Z. Han, Q. Ye, and J. Jiao, "Online feature evaluation for object tracking using kalman filter," *In ICPR*, pp. 1–4, 2008. Article (CrossRef Link)

[11] L. Bazzani, M. Cristani, and V. Murino, "Decentralized particle filter for joint individual-group tracking," *In CVPR*, pp. 1886–1893, 2012. Article (CrossRef Link)

[12] Yaowen Guan, Xiaoou Chen, Deshun Yang, and Yuqian Wu, "Multi-person tracking-by-detection with local particle filtering and global occlusion handling," *In IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2014. Article (CrossRef Link)

[13] Maha M. Azab, Howida A. Shedeed, and Ashraf Saad Hussein, "New technique for online object tracking-by-detection in video," *IET Image Processing*, vol. 8, no. 12, pp. 794–803, 2014. Article (CrossRef Link)

[14] Arne Schumann, Martin Bauml, and Rainer Stiefelhagen, "Person tracking-by-detection with efficient selection of part-detectors," *In 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS 2013, pp. 43–50, 2013. Article (CrossRef Link)

[15] Z. Gao, H. ZHang, G.P Xu, Y.B Xue, "Multi-perspective and Multi-modality Joint Representation and Recognition Model for 3D Action Recognition," *NeuroComputing*, vol. 151, pp. 554–564, 2015. Article (CrossRef Link)

[16] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," In CVPR, pp. 1815–1821, 2012. Article (CrossRef Link)

[17] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," *In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008. Article (CrossRef Link)

[18] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," In *proc. of* ECCV 2012, *12th European Conference on Computer Vision*, pp. 702–715, 2012. Article (CrossRef Link)

[19] E. Piatkowska, A. N.Belbachir, S. Schraml, and M. Gelautz, "Spatiotemporal multiple persons tracking using dynamic vision sensor," *In CVPR Workshops*, pp. 35–40, 2012. Article (CrossRef Link)

[20] I. Zuriarrain, F. Lerasle, N. Arana-Arejolaleiba, and M. Devy, “An mcmc-based particle filter for multiple person tracking,” *In ICPR*, pp. 1–4, 2008. Article (CrossRef Link)

[21] R. Munoz-Salinas, “A bayesian plan-view map based approach for multiple-person detection and tracking,” *Pattern Recognition*, vol. 41, no. 12, pp. 3665–3676, 2008.Article (CrossRef Link)

[22] Zan Gao, Long-fei Zhang, Ming-yu Chen, Alexander Hauptmann, Hua Zhang, Anni Cai, “Enhanced and Hierarchical Structure Algorithm for Data Imbalance Problem in Semantic Extraction under Massive Video Dataset,” *Multimedia Tools and Applications*, vol. 68, no. 3, 2014. Article (CrossRef Link)

[23] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *In CVPR*, pp. 886–893, 2005. Article (CrossRef Link)

[24] M. Becha Kaaniche and F. Bremond, “Tracking hog descriptors for gesture recognition,” *In AVSS*, pp. 140–145, 2009. Article (CrossRef Link)

[25] Thomas Mauthner, Peter M. Roth, and Horst Bischof, “Learn to move: Activity specific motion models for tracking by detection,” in *Proc. of Computer Vision - ECCV 2012*, pp.183–192, 2012. Article (CrossRef Link)

[26] Xiaoyan Jiang, Erik Rodner, and Joachim Denzler, “Multi-person trackingby-detection based on calibrated multi-camera systems,” in *Proc. of Computer Vision and Graphics - International Conference*, pp. 743–751, 2012. Article (CrossRef Link)

[27] Michael Bredereck, Xiaoyan Jiang, Marco Korner, and Joachim Denzler, “Data association for multi-object tracking-by-detection in multi-camera networks,” in *Proc. of Sixth International Conference on Distributed Smart Cameras*, pp. 1–6, 2012. Article (CrossRef Link)

[28]Z. Gao, H. Zhang, G.P Xu,Y.B Xue and A. G. Hauptmannc, “Multi-View Discriminative and Structured Dictionary Learning with Group Sparsity for Human Action Recognition,” *Signal Processing*, 2015. Article (CrossRef Link)

[29] B. Yang and R. Nevatia, “Multi-target tracking by online learning of nonlinear motion patterns and robust appearance models,” *In CVPR*, pp. 1918–1925, 2012. Article (CrossRef Link)

[30] M. Patzold, R. Heras Evangelio, and T. Sikora, “Boosting multi-hypothesis tracking by means of instance-specific models,” *AVSS*, pp. 416–421, 2012. Article (CrossRef Link)

[31] T. Gehrig and J. W.McDonough, “Tracking multiple speakers with probabilistic data association filters,” *CLEAR*, pp. 137–150, 2006. Article (CrossRef Link)

[32] H. Jiang, S. Fels, and J. J. Little, “A linear programming approach for multiple object tracking,” *In CVPR*, 2007. Article (CrossRef Link)

[33] J. Berclaz, F. Fleuret, and P. Fua, “Robust people tracking with global trajectory optimization,” *In CVPR*, pp. 744–750, 2006. Article (CrossRef Link)

[34] Liu Feng, Liu Xiaoyu, and Chen Yi, “An efficient detection method for rare colored capsule based on RGB and HSV color space,” in *Proc. of 2014 IEEE International Conference on Granular Computing*, pp. 175–178, 2014. Article (CrossRef Link)

[35] Mohammad Nazmul Alam Khan, Guoliang Fan, Douglas R. Heisterkamp, and Liangjiang Yu, “Automatic target recognition in infrared imagery using dense HOG features and relevance grouping of vocabulary,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 293–298, 2014. Article (CrossRef Link)

[36] S. Pellegrini, A. Ess, and L. J. Van Gool, “Improving data association by joint modeling of pedestrian trajectories and groupings,” *In ECCV*, pp. 452–465, 2010. Article (CrossRef Link)

[37] P. Gorur and B. Amrutur, “Speeded up gaussian mixture model algorithm for background subtraction,” *AVSS*, pp. 386–391, 2011. Article (CrossRef Link)

[38] R. Kasturi, D. B. Goldgof, P. Soundararajan, and et al, “Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol,” *IEEE Trans. Pattern Anal*, vol. 31, no. 2, pp. 319–336, 2009. Article (CrossRef Link)

[39] B. Benfold and I. Reid, “Stable multi-target tracking in real-time surveillance video,” *In CVPR*, pp. 3457–3464, 2011. Article (CrossRef Link)

[40] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” *In CVPR*, 2008. Article (CrossRef Link)

[41] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" *In CVPR*, pp. 1345–1352, 2011. Article (CrossRef Link)

[42] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," *In ICCV Workshops*, pp.120–127, 2011. Article (CrossRef Link)

Yu-ting Su received the B.S., M.S. and Ph.D. degrees in electronic engineering from Tianjin University of China, in 1995, 1998 and 2001, respectively. He is currently a Professor at the School of Electronic Information Engineering in Tianjin University. His research interests include digital video coding, digital watermarking and data hiding multimedia forensics, and multimedia retrieval.

Xiao-rong Zhu received the B.S. degrees in Xidian University of China, and received M.S. degrees in the school of electronic engineering from Tianjin University of China. Her research interests include computer vision, digital watermarking and data hiding multimedia forensics.

Wei-zhi Nie received the B.S. and M.S. degrees in electronic engineering from Tianjin University of China. He is currently pursuing the Ph.D. degree from Tianjin University. His research interests include multiple object tracking, computer vision and location-based social network.