ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414747

MULTIPLE HUMAN TRACKING IN NON-SPECIFIC COVERAGE WITH WEARABLE CAMERAS

Sibo Wang[1,2], *Ruize Han*[1,2†], *Wei Feng*[1,2], *Song Wang*[1,3]

[1]College of Intelligence and Computing, Tianjin University, China
[2] Key Research Center for Surface Monitoring and Analysis of Cultural Relics, China
[3] Department of Computer Science and Engineering, University of South Carolina, USA
{wangsibo_123, han_ruize, wfeng}@tju.edu.cn, songwang@cec.sc.edu

ABSTRACT

Compared to fixed cameras, wearable cameras have time-varying non-specific view coverage and can be used to alternately observe people at different sites by varying the camera views. However, such view change of wearable cameras may introduce intervals of transitional frames without useful information, which brings new challenge for the important multiple object tracking (MOT) task – existing MOT methods can not handle well frequent disappearing/reappearing targets in the field of view, especially in the presence of informationless transitional sequences of frames. To address this problem, in this paper we propose a Markov Decision Process with jump state (JMDP) to model the target's lifetime in tracking, and use optical flow of the camera motion and the statistical information of the targets to model the camera state transition. We further develop a frame-level classification algorithm to locate the transitional sequence. By combining all of them, we formulate the proposed non-specific-coverage MOT problem as a joint state transition problem, which can be solved by the state transfer mechanism of the targets and the camera. We collect a new dataset for performance evaluation and the experimental results show the effectiveness of the proposed method.

Index Terms— Multi-human tracking, wearable cameras, abnormal frames

1. INTRODUCTION

Multiple object tracking (MOT), especially multiple human tracking (MHT), has wide applications in video surveillance and human-machine interaction. Most existing methods use fixed cameras for video collection, whose field of view (FOV) is unchanged and limited. In contrast, wearable cameras, e.g., GoPro and Google glass, worn by and moved with wearers, have time-varying non-specific observation coverage [1–5] and can be used to track and observe people at different sites by varying the camera views, which enables more flexible and wide-range outdoor video surveillance of crowded scenes. The goal of this paper is to study the new problem of MHT in non-specific fields using wearable cameras.

This is a highly challenging problem given the indeterminate change of the camera FOVs. We consider two typical situations in this paper: 1) camera view may suddenly move alway from any targets, leading to frames without any people, e.g., viewing the sky or ground for a break, and 2) the camera FOV is quickly changed to cover different human groups at different sites, also leading to temporal intervals without useful information. In either situation, the camera FOV exhibits a sudden and large change and we refer to the resulting intervals of transition as a 'transitional sequence'. Figure 1(a) shows a sample video of situation 1) where the transitional sequence contains no targets of interest. By taking the first and the last frames of transitional sequence as the cut-off boundary, this video can be divided into three disjoint video segments shown in Fig. 1(b). Although high inter-frame continuity is shown within each segment, the inter-frame continuity between the segments, i.e., at the boundary frames a and b in Fig. 1(b), is very poor. An example of situation 2) is shown in Fig. 1(c), where the camera alternately change the view to observe different groups of people at different sites with transitional sequences. In both situations, the resulting videos lack the throughout inter-frame continuity and appearance consistency that are required by most existing MOT methods [6–8]. While few existing MOT methods may partially address this problem by integrating a person re-identification mechanism [9–11], they substantially increase the searching space and still have difficulty in handling targets of similar appearance but in different groups.

In order to handle the above two common situations, in this paper we build a Markov Decision Process [12] with jump state (JMDP) for each target to complete the tracking decision and state transition. Meanwhile, a camera state transition mechanism is introduced to judge the FOV of camera at each time. For that, we also train a binary classifier to accurately identify the beginning and end of each transitional sequence based on the location information of the targets and the background information of the image in each frame.

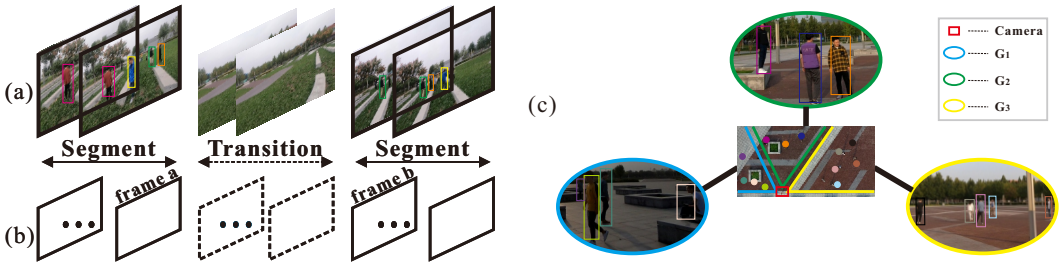

Fig. 1. An illustration of transitional sequences caused by (a, b) temporally moving the camera view away from the same group of targets and (c) changing the camera view from one site to another to observe different groups of people.

Our main contributions are: (1) To the best of our knowledge, we are the first to study the multiple human tracking (MHT) in non-specific coverage using the wearable camera, by considering both the cross-transitional-sequence MHT for each group and switch-group MHT for multiple groups; (2) We establish a Markov Decision Process with jump state (JMDP) and a camera state transition mechanism to handle the non-specific-coverage tracking problem; (3) We collect a new dataset of videos with various kinds of transitional sequences for performance evaluation. We have released this dataset to public[1].

2. PROPOSED METHOD

In this section, we first introduce Markov Decision Processes (MDP) with jump state to model the tracking shown in Fig. 1. We further propose the camera state transition (CST) model to identify the jump state caused by the camera movement and a Transitional Sequence Identification (TSI) model to locate the transitional sequence.

2.1. MDP with Jump State (JMDP)

Given a new input video frame, a JMDP is first initialized for each detected target, and the state is initialized to active. Next, we apply a single object tracking (SOT [13, 14]) approach, e.g., ECO [15], to keep tracking each target. The target state is set as tracked when the target maintains the active state on more than α frames. When the tracking process gets unreliable, e.g., the tracking score is low or the tracking result is inconsistent with the detection result, we suspend the tracker and set the target to the lost state. We then perform data association in DMAN [10] to compute the similarity between the tracklet and detections that are not covered by any tracked target. After that, the similarity scores are used in the Hungarian algorithm [16] to obtain the assignment between the detections and the lost targets. According to the assignment, lost targets that are linked to object detections are transferred to tracked state. Otherwise, they stay as lost. In particular, whenever entering a transitional sequence, the states of all targets are transferred to jump state immediately. The specific state transitions are described in Fig. 2.

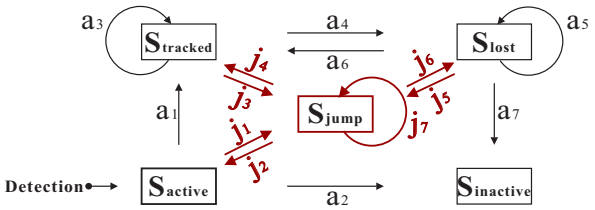

Fig. 2. The target JMDP in our framework.

Active: It is the initial state of any target. Whenever an object is detected by the object detector, it will enter s_{active}.
Tracked: A tracked target can keep as s_{tracked}, or transition to s_{lost} if the target is lost due to some reason, such as occlusion or disappearance from the field of view of the camera.
Lost: A lost target can stay as s_{lost}, or go back to s_{tracked} if it appears again, or transition to s_{inactive} if it has been lost for a sufficiently long time.
Jump: At the beginning of a transition frame sequence, all the targets with any of the above states will enter s_{jump} immediately. First, we save the current state of each target as s_{p}, and then transfer it to s_{jump}, e.g. for the actions j_1, j_3, j_5 in Fig. 2. When the camera moves back to these targets again, we associate the current detection results with the targets with state s_{jump}. We need to determine whether to go back to the previously saved state s_{p} according to the associated score d calculated by DMAN [10], as shown by actions j_2, j_4, j_6 and j_7, in Fig. 2. In this case, the MDP transforms is

$$s_{i,t} = \begin{cases} s_{\text{p}}, & \text{if } s = s_{\text{jump}} \text{ and } d > \tau \\ s_{\text{jump}}, & \text{otherwise,} \end{cases} \quad (1)$$

Inactive: It is the terminal state for any target and an inactive target stays as inactive forever.

The above is our overall tracking framework. However, when tracking among multiple groups at different sites as shown in Fig. 1(c), there are multiple transition sequences $\mathcal{J} = \{J_1 \cup J_2 \cup ... \cup J_i\}$ in a video. It is essential to know which group the camera is looking at before and after each transitional sequence J_i.

[1]https://github.com/github19970909/NSMHT

2.2. Camera State Transition (CST)

To handle the above two situations, e.g., those shown in Fig. 1(a) and (c), we propose a model of camera state transition, for recording the current view direction of the camera. First, a camera state c is initialized to $(0, 0)$ for each target appearing in its current FOV, and $(0, 0)$ is used to represent the target is re-observed (after disappearing for a while) by the camera. When the tracking enters the start (e.g., frame a in Fig. 1(b)) or the end (e.g., frame b in Fig. 1(b)) of the transitional sequence J_i, the camera state transition is performed for each target according to its $[N, C]$ values, where we use $N \in \{1, -1\}$ to show the trend of the number of people at a frame, where $N = 1$ indicates the number of people is increasing, which means that the camera is turning to new people group as $\mathrm{f}_{12}{}^{+}$,$\mathrm{f}_{2n}{}^{+}$,$\mathrm{f}_{n2}{}^{+}$ shown in Fig. 3. $N = -1$ indicates that the number of people is decreasing, which means that the camera is leaving the current group as $\mathrm{f}_{12}{}^{-}$,$\mathrm{f}_{2n}{}^{-}$,$\mathrm{f}_{n2}{}^{-}$ shown in Fig. 3. Considering all possible directions of the camera, we divide the camera motion into eight directions, and get the moving direction of the camera according to the optical flow direction. C is used to represent the moving direction of the camera, $C =< C_{\mathrm{h}}, C_{\mathrm{v}} >$, $C_{\mathrm{h}}/C_{\mathrm{v}} \in \{1, -1, 0\}$ is the horizontal/vertical direction, where 1 represents the moving direction is right/up, and -1 represents the moving direction is left/down, 0 means there is no movement in the horizontal or vertical direction.

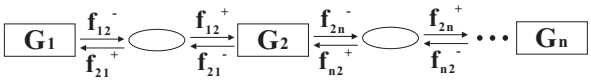

Fig. 3. Camera state transition. $\mathrm{f}_{jj'}{}^{-}/\mathrm{f}_{jj'}{}^{+}$ indicates the number of people is reducing/increasing from group j to j'.

We update the camera state c at time t as

$$c = c + C_t \times \left[(N_t \odot -1) + (N_t \odot 1) \left(C_t \oplus C_{t-1} \right) \right], \quad (2)$$

where N_t and C_t are defined as above with a frame index t, and t denotes the frame number of the start or end of the identified transitional sequence. Here $\odot$ and $\oplus$ denote the xnor and xor operations, respectively. For example,

$$N_t \odot -1 = \begin{cases} 1, & \text{if } N_t = -1 \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

$$C_t \oplus C_{t-1} = \begin{cases} 1, & \text{if } C_t \neq C_{t-1} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

2.3. Transitional Sequence Identification (TSI)

In order to find the start (e.g., frame a in Fig. 1(b)) and end (e.g., frame b in Fig. 1(b)) of transitional sequence mentioned before, we need to recognize whether the frame is in the transitional sequence or not, which we refer to as frame-based transitional sequence identification.

There are several clues to help transitional sequence identification. First, when the camera view is significantly changed in the considered two situations, it usually starts and ends with a sudden and dramatic camera movement. Therefore, a frame with a very large optical flow is more likely to be the start of transitional sequence. Second, the change in the number of people in the FOV can also be used to help recognize transitional sequence. Following these clues, for each frame t, we first calculate the optical flow between frame $t-1$ and t and take the average value O_t. Second, we calculate P_t, the number of people detected on frame t. Finally, we construct a 12-dimensional feature vector and train a binary support vector machine (SVM) classifier with RBF kernel $[O_{t-6}, \cdots, O_t, P_{t-6}, \cdots, P_t]$ for identifying the transitional sequence. Influence of different features in this method will be discussed in Section 3.4.

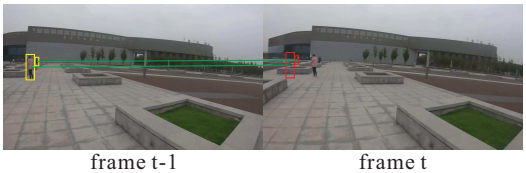

Fig. 4. The red boxes in frame t are the predictions of the position of yellow boxes in frame t-1 by directly using the constant velocity motion model.

2.4. Transitional Sequence Alignment (TSA)

In the transitional sequence, the camera was accompanied by the sudden and dramatic movement. At this time, the position of an object usually changes considerably, as shown in Fig. 4, where it is unreliable to use a constant velocity motion model [17, 18] to predict motion consistency as in most previous MOT tracking methods. Thus, we proposed to perceive the camera motion using the Camera State Transition (CST) model. When the camera moves quickly during the transitional sequence, we apply the transitional sequence alignment (TSA), e.g., Enhanced Correlation Coefficient (ECC) [19], to align the images. Then we apply the constant velocity motion model on the aligned images for more accurate tracking.

3. EXPERIMENTS

3.1. Dataset and Metrics

We do not find publicly available dataset for non-specific coverage switch-group tracking. Therefore, we use head-mounted GoPro to collect a new video dataset, which consists of 34 videos of length from 400 to 1,500 frames, in total 28,630 frames, taken at 5 different outdoor sites. Specifically, on 16 videos, it switches to tracking between crowd and non-crowd scenes; on 12 videos, it switches tracking between two groups of people; and on 6 videos, it switches tracking between three groups of people. We apply the standard MOT metrics for evaluating the tracking performance [20–22],

including multi-object tracking precision (MOTP) and multi-object tracking accuracy (MOTA). The main motivation and aim in this paper is to continuously track the multiple targets against the quick motion and switch of the wearable cameras. Therefore, we care more about the ID-related metrics, i.e., IDP, IDR and IDF_1 in our experimental evaluation.

3.2. Setup

We use the general YOLOv3 [23] detector for human detection. For single object tracking, we use the same features as in ECO-HC (i.e., HOG and Color Names) [15]. We use D-MAN [10] as the baseline method. And choose two MOT methods, i.e., MDP [12], and Tracktor++ [11] as the compared methods. Among them, Tracktor++ converts a detector into a Tracktor with a re-identification to improve identity preservation across frames. MDP and DMAN are single object tracker based methods, where DMAN learns deep appearance features for data association. For fair comparison, we use the same object detector for all the methods, including the proposed method and these comparison methods.

Table 1. Comparative results of different methods.

Method	IDF_1	IDP	IDR	MOTA	MOTP
MDP	50.3	51.2	49.5	73.6	82.3
DMAN	55.1	55.0	55.2	79.6	81.4
Tracktor++	47.5	51.4	44.1	77.2	83.3
Ours	72.9	72.1	73.7	78.4	81.2

3.3. Results

We evaluate the proposed method on all the videos in our dataset against the state-of-the-art methods and the results are shown in Table 1. We find that although using the same object detector as DMAN in human detection, our method outperforms DMAN by a wide margin in the ID-related metrics. As shown the last row in Table 1, on the overall dataset, our method achieves a comparable MOTA score at 78.4% and performs favorably against the state-of-the-art methods in terms of identity-preserving metrics. We improve by 17.8% in IDF_1, 17.1% in IDP, 18.5% in IDR compared with the second best performed MOT method listed in this table, which demonstrates the merits of the proposed method in maintaining the target ID. Note that, our method does not track the targets in the transitional sequence, which results in some missed detection, i.e., the false positive (FP) detections during tracking, and decreases the MOTA score to some extent. This problem can be alleviated by activating the re-tracking when the targets return to the camera's FOV. Similarly, at the beginning of the transition frames, we can ceaselessly track the targets until all of them disappear in the camera's FOV.

Moreover, we divide the dataset into one-group, two-group and three-group videos and evaluate the MOT performance, respectively. As shown in Table 2, we can first find that the tracking results in one-group case shows better performance than the two-group and three-group cases. This is due to the identification of transitional sequences in multiple groups is more complicated than single group. Meanwhile, the result reflects the importance of the accuracy of transitional sequence identification to the tracking result.

Table 2. Results on the one-, and multiple-group data.

Method	one group			multiple groups		
	IDF_1	IDP	IDR	IDF_1	IDP	IDR
MDP	50.2	50.0	50.4	50.5	52.8	48.3
DMAN	53.7	53.2	54.1	56.9	57.3	56.5
Tracktor++	44.4	47.0	42.1	51.5	57.6	46.6
Ours	74.7	73.6	75.8	70.6	70.1	71

3.4. Ablation Study

The effectiveness of each proposed module is shown in Table 3. Clearly, the method with JMDP and CST can better maintain the ID of the tracked targets. In the last row in Table 3, we demonstrate the contribution of TSA in our algorithm just like the analysis given in Sec. 2.4. The whole framework we proposed achieves 72.9% in IDF_1.

Table 4 reveals the effects of different features in TSI: 'w/o O' and 'w/o P' denote the proposed features without the optical flow and the number of people in identifying the transitional sequence, respectively. We can see that using only one of them cannot achieve performance as good as the proposed method that combines both of them.

Table 3. Ablation study on the use of TSA.

Method	IDF_1	IDP	IDR	MOTA	MOTP
Baseline	55.1	55.0	55.2	79.6	81.4
Baseline+JMDP+CST	64.8	64.7	64.9	79.6	81.3
Baseline+JMDP+CST+TSA	72.9	72.1	73.7	78.4	81.2

Table 4. Ablation study on the use of different features.

Method	IDF_1	IDP	IDR	MOTA	MOTP
w/o P	56.1	56.0	56.2	79.7	81.4
w/o O	72.6	71.7	73.6	77.9	81.2
ours	72.9	72.1	73.7	78.4	81.2

4. CONCLUSION

In this paper, we proposed a Markov Decision Process with jump state (JMDP) and a camera state transition (CST) mechanism to handle the non-specific-coverage multiple human racking problem. To evaluate our method, we collect a new dataset with multiple types of transitional sequence. Experimental results on our collected datasets verified the effectiveness of the proposed method. Through the above efforts, we just hope to provide the fundamental resources to extend the MOT problem to wearable-camera videos, which can facilitate the video analysis to more application scenarios.

Acknowledgement: This work was supported, in part, by NSFC U1803264, 61672376, 62072334, and 61671325.

5. REFERENCES

[1] A. Cartas, P. Radeva, and M. Dimiccoli, "Activities of daily living monitoring via a wearable camera: Towards real-world applications," *IEEE Access*, no. 99, pp. 1–1, 2020.

[2] R. Han, Y. Zhang, W. Feng, C. Gong, X. Zhang, J. Zhao, L. Wan, and S. Wang, "Multiple human association between top and horizontal views by matching subjects' spatial distributions," in *arXiv*, 2019.

[3] R. Han, W. Feng, J. Zhao, Z. Niu, Y. Zhang, L. Wan, and S. Wang, "Complementary-view multiple human tracking," in *AAAI Conference on Artificial Intelligence*, 2020.

[4] R. Han, J. Zhao, W. Feng, Y. Gan, L. Wan, and S. Wang, "Complementary-view co-interest person detection," in *ACM Multimedia*, 2020.

[5] J. Zhao, R. Han, Y. Gan, L. Wan, W. Feng, and S. Wang, "Human identification and interaction detection in cross-view multi-person videos with wearable cameras," in *ACM Multimedia*, 2020.

[6] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[7] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *IEEE International Conference on Computer Vision*, 2017.

[8] A. Attanasi, A. Cavagna, L. D. Castello, I. Giardina, A. Jeli, S. Melillo, L. Parisi, F. Pellacini, E. Shen, E. Silvestri, and M. Viale, "GReTA-A novel global and recursive tracking algorithm in three dimensions," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2451–2463, 2015.

[9] Y. Young Chul, Y.-M. Song, K. Yoon, and M. Jeon, "Online multi-object tracking using selective deep appearance matching," in *IEEE International Conference on Consumer Electronics - Asia*, 2018.

[10] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M. Yang, "Online multi-object tracking with dual matching attention networks," in *European Conference on Computer Vision*, 2018.

[11] P. Bergmann, T. Meinhardt, and L. Leal-Taix, "Tracking without bells and whistles," in *IEEE International Conference on Computer Vision*, 2020.

[12] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *IEEE International Conference on Computer Vision*, 2015.

[13] R. Han, Q. Guo, and W. Feng, "Content-related spatial regularization for visual object tracking," in *IEEE International Conference on Multimedia and Expo*, 2018.

[14] R. Han, W. Feng, and S. Wang, "Fast learning of spatially regularized and content aware correlation filter for visual tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 7128–7140, 2020.

[15] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Internaltional Conference on Computer Vision and Pattern Recogintion*, 2017.

[16] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 1, pp. 32–38, 1962.

[17] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs," in *European Conference on Computer Vision*, 2012.

[18] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.

[20] L. Lealtaix, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," in *arXiv*, 2015.

[21] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Eurasip Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[22] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *European Conference on Computer Vision*, pp. 17–35, 2016.

[23] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Computer Vision and Pattern Recognition*, 2016.