Visual Persons Behavior Diary Generation Model based on Trajectories and Pose Estimation

Chen Gang[1 2*], Chen Bin[2,3], Liu Yuming[4], Li Hui[4]

[1]Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, 610041
[2]University of Chinese Academy of Sciences, Beijing, 100049
[3]Guangzhou Institute of Electric Technology, Chinese Academy of Science, Guangzhou, 510070)
[4]Electronic dispatch control center of Yunnan power grid group co. LTD

*Corresponding author e-mail:chengangfzu@foxmail.com

Abstract The behavior pattern of persons was the important output of the surveillance analysis. This paper focus on the generation model of visual person behavior diary. The pipeline includes the person detection, tracking, and the person behavior classify. This paper adopts the deep convolutional neural model YOLO (You Only Look Once)V2 for person detection module. Multi person tracking was based on the detection framework. The Hungarian assignment algorithm was used to the matching. The person appearance model was integrated by HSV color model and Hash code model. The person object motion was estimated by the Kalman Filter. The multi objects were matching with exist tracklets through the appearance and motion location distance by the Hungarian assignment method. A long continuous trajectory for one person was get by the spatial-temporal continual linking algorithm. And the face recognition information was used to identify the trajectory. The trajectories with identification information can be used to generate the visual diary of person behavior based on the scene context information and person action estimation. The relevant modules are tested in public data sets and our own capture video sets. The test results show that the method can be used to generate the visual person behavior pattern diary with certain accuracy.

1. Introduction

The trajectory of multiple persons was an important research significance and application value in the visual surveillance. Due to the constraints of the surveillance environment such as the scene light conditions, occlusion, scale changing, object accessing and out of scenarios, it is hard work for tracking in the long-term for multiple persons. Based on the HOG (Histogram of Gradient), color histogram, SIFT (Scale Invariant Feature Transform) features and trackers such as Kalman Filter, nonlinear particle filter and correlation filter, many method has been used on the persons tracking, as in[1] [2] [3]. Due to the deep learning has been widely studied in object detection and its performance is very good, multi-persons tracking based on detection framework becomes a hotspot research.

Literature[4] combines with the YOLO[5] model on multiple persons detection results to get trajectories. Literature[6] combines with YOLO object detection framework, and use the Kalman filter to get the estimated object location in the next video frame. Through the judgment on location coincidence between track estimate and detection, the dynamic allocation model was used to achieve multi-target tracking. Although the feature and model parameter are relatively simple, the framework has achieved a good performance in the MOT challenge 2016 dataset. The algorithm was improved when combined the person deep descriptor[7].According to the relevant research shows that the object detection and tracking based on the deep learning has important research foreground and application value. The purpose of this paper is to obtain statistical analysis of the relevant person object behavior patterns during the long-term tracking in surveillance video. The challenges and difficulties of tracking for long-term include: (1) Occlusion. Long-term tracking is blocked by other object more possibility. (2) Scale factors. (3) Object appearance changed. In the special application (object distance is less than 5 meters in the surveillance), face detection and recognition can reduce the difficulty of long time tracking. This paper uses the Geographic Information System (GIS) method by adding attribute information to the trajectories. The attribute information include color feature, object bounding box area, object moving speed, face feature etc. The persons behavior pattern in the surveillance were interpretation through integrating persons pose estimation, trajectories and scene context. The persons pose estimation was achieved the real-time in the literature[8].The persons key points combing with the object detection could be used for atomic action analysis. The scene context information combing with the trajectories were used for estimate person behavior in this paper. For example, a person trajectories location was near the equipment so we can reason that the person was operating equipment. But the person may had a rest. So the pose estimation can make more accurate judgments on the person behavior.

2. Method

The deep learning detection framework YOLO was used to detect the person. And the Kalman Filter was selected as the tracker to generate the short trajectory segments. Based on the spatial-temporal consistency of trajectory attributes, the iterative algorithm was used to connect the segment trajectories. The space coordinate mapping relationship between the image coordinate and the surveillance scene is established through camera calibration technology. The algorithm pipeline includes person detection, the trajectory segments generation, the trajectories segments connection, the relationship between scene context and trajectories, the person pose estimation. The algorithm framework as shown in algorithm I.

Algorithm I the statistical human behavior pattern algorithm based on object long tracking trajectory in surveillance

```
Input video frames {$I_t$: t=1,…,∞}，Candidate face templates for person objects recognition
Output The object long term tracking trajectory to a specific person object, target behavior type statistics
For Each  $I_t$ with t = 1,…,∞ Do
          DetectPerson()(YOLO detect object bounding box)
          DetectFace()(Face object bounding box)
Tracking Initialization
For Each {$FaceROI_r$ t:r=0,…,R} and {$PersonROI_r$ t:r=0,…,R} Do
  If $PersonROI_r$ and $FaceROI_r$ ( belong to nothing existing tracking
    trajectories)
Then K=K+1 (increase the trajectories number)
Track
(To existing trajectories in the video frame $I_t$, search for object detection for the each track in the $I_{t+1}$ frame. The matching feature is divided into motion matching and object appearance matching, and auxiliary face information. Motion matching is performed by Kalman Filter. Color
```

histogram and perceptual Hashing code are used in the appearance model. According to the tracking process, the trajectories position and the attribute parameters including position, speed, start frame, end frame, color feature were recorded.)

Trajectories Connection

(Some trajectory fragments belong to the same ID object to generate long-term trajectories.)

Tracks Recognition

(According to the sampling interval discrete value, the face feature was computed to compare with the candidate face libraries. The ID number was given for the corresponding trajectory. Since the face capture and recognition in video need the certain conditions, it also provide manual person object recognition method.)

Behavior Analysis

(According to the specific application scenarios include behavior inference rule, spatial relationship between persons and other contextual and some simple pose estimation to judge the person object simple behavior type. The temporal and spatial distribution of trajectories were used to the special person behavior type.)

2.1. Object detection based on YOLO

In recent years the object detection method based on the deep convolutional neural network has been developed rapidly. It has made great progress in the detection accuracy, and the detection speed has been improved very quickly. The emergence of RCNN, Fast R-CNN, Faster R-CNN, SSD, YOLO, YOLO-V9000 and other detection model were established. This paper uses the YOLO model.

2.2. Person object tracking

The object appearance model were modeled by hash code and color histogram feature descriptor. The object motion pattern was tracking using the Kalman filter. According to the threshold value, the distance measurement of motion and appearance model was used to allocate the objects to the existing trajectories. The Hungarian algorithm was used in the allocation strategy. This paper used the discrete sampling algorithm to speed up the algorithm. Each face recognition was performed once every M frames. The effect of face recognition is to identify the tracking trajectory ID and detection the objects re-enter the surveillance scene. (1) According to the scale division between head and body the head candidate region was get. And the harr-like feature, SVM (Support Vector Machine) classifier were used to detect the face in the proposal region. (2) Because of the person posture, the lighting variation, the scale and other reasons, the face cannot been detected in the each proposal region. Continue to detect the face until it is been detected.

2.3. The projection between image plane and real surveillance scene coordinate system

According to the geometric relationship of the imaging process, the mapping relationship between the 3D coordinate system of the surveillance scene and the imaging plane of the camera could be established. The commonly used concise model was the pinhole model. The surveillance scene was indoor environment and the ground plane is horizontal.

2.4. Person behavior analysis

The trajectories spatial-temporal pattern were calculated by using the trajectory spatial location and attribute data and the context information.

(1) Person behavior type was inferenced based on rules. The rules were generated by experience. According to the spatial distance between the object location and the context of the surveillance scene, the residence time, the moving frequency and other indicators the behavior pattern was inferenced. The method is suitable for that person objects were less. For example the large scale

production workshop with higher automation level was suitable environment. Through the long term trajectories, context equipment and pose estimation the behavior statistical model of worker was learned.

(2) Behavior types analysis based on pose estimation. The key points of person body in the video image are directly detected to estimate the pose of body and infer the type of human behavior. Recently person pose estimation based on video image has been get great progress, the literature [8] presents a method called PAF to detect human skeleton key point for real time.

3. Experiment

The experiment used two sets of test video. One test video sets was used for multi target tracking and long term trajectory generated performance. Another sets was used for evaluation the algorithm of person behavior visual dairy based on long time tracking trajectory, pose estimation and scene context information.

3.1. Evaluation multi persons tracking for long-term

The video set was captured in a manufacturing factory. The video time length is 2 hours. The image resolution is 1920*1080. The ground truth was made by the video annotation tool. The evaluation indicators are as follows.

Table 1 Evaluation indicators

MOTA	the multi object tracking accuracy
IDSW	different object switching times of the same ID tracking trajectory
FRAG	number of trajectory segments

Table 2 Tracking results

Algorithm	MOTA(%)	IDSW	FRAG
C++ SORT	80.5	13	22
Ours	92%	2	3

We want to compare ours method with others. But the most of MOT algorithms were not open source. So we compared with C++ SORT algorithm that was open source. Because the face information and trajectory spatial temporal continuous processing algorithm, the IDSW and FRAG indicator of our result were less than the C++ SORT result. The result show that the method has certain trajectory correction ability and is suitable for long term object tracking.

3.2. Persons visual behavior dairy generated test

The test video was captured in Zhejiang Tobacco Factory of China. The classification and judgment of simple human behavior pattern are based one below condition.

(1) The judgment basis for relevant staff is track position and duration time. The position coincidence threshold was set with S (30cm) and the duration threshold time was set with T (20s) in the machine area called A. The openpose algorithm detected the person body key point which could enhance the person location precision.

(2) The position coincidence threshold was set with S (30cm) and the duration threshold time was set with T (20s) in the machine area called B.

(3) Rest and other activities was judged by the condition that the track person position and the track person move frequency(the track distance of unit time) is less than W(3m/30s)

Table 3 Analysis result of the test video

	The workspace A A(S)	The workspace B B(S)	Rest and other activities(S)
Ground Truth	36	192	2122
The Algorithm	16	104	1720
Accuracy	44%	54%	81%

（If the location coincidence between the track object and the workspace area and duration time satisfy the threshold condition, then the track person behavior was classified with relevant work behavior pattern.）

4. The research conclusion

This paper use the deep learning framework for person object detection, and the tracking method was under the detection framework. The face feature can been used for object ID recognition and object re-identification. For the purpose of object long term tracking we use the face feature and spatial temporal continuum track fragments processing. According to the test results, the proposed algorithm has the following characteristics. (1)The algorithm has the ability to track multi persons for long term. (2)The algorithm framework has the ability to transform between the 2D image coordinate system and the 3D real surveillance scene. The spatial-temporal pattern can been get by using the trajectory space coordinate and attribute data based on surveillance scene context information. (3)The proposed algorithm framework was the pipeline characteristic. The submodule can been replaced by the new algorithm module which makes the performance and efficiency of the framework can be improved with the computer vision technique advances quickly.

References

[1] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detectionusing a detector confidence particle filter[C].Japan:ICCV,2009.1515-1522.

[2] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M.Lui. Visual object tracking using adaptive correlation filters[C].USA:Computer Vision and Pattern Recognition, 2010.2544–2550.

[3] V. Belagiannis, F. Schubert, N. Navab, and S. Ilic.Segmentation based particle filtering for real-time 2d object tracking[C].Italy: ECCV,2012. 842–855.

[4] Konstantine Buhler,John Lambert,Matthew Vilim.Yolo Flow Real-time Object Tracking in Video[J].arXiv, 2016.1604.07468

[5] Joseph Redmon et al. You Only Look Once Unified Real-Time Object Detection[J].CoRRabs,2015.1506.02640.

[6] Bewley, Alex and Ge, Zongyuan and Ott, Lionel and Ramos, Fabio and Upcroft, Ben.Simple online and realtime tracking[C].USA:IEEE International Conference on Image Processing (ICIP),2016.3464-3468.

[7] Nicolai Wojke, Alex Bewley,Dietrich Paulus.Simple Online and Realtime Tracking with a Deep Association Metric[J].arXiv,2017.1703.07402v1.

[8] Zhe Gao and Tomas Simon and Shih-En Wei and Yaser Sheikh.Real Time Multi-Person 2D Pose Estimation using Part Affinity Fields[**C**].USA:Computer Vison and Pattern Recognition,2016.