Research on pedestrian tracking algorithm based on deep learning framework

Chuanxu Wang¹, Xiaozhen Sun², * and Hui Li³
¹,²,³School of Qingdao University of Science & Technology, Institute of information science and technology University, Qingdao, China

*Corresponding author e-mail: sunxiaozhen0006@163.com

Abstract. Pedestrian tracking is a hot topic in the field of computer vision. Current pedestrian tracking methods may treat the reappear target as a new target to tracking, which can lead to tracking failure. For this purpose, this paper uses the deep learning framework to complete the whole algorithm of pedestrian tracking to solve this problem. The pedestrian tracking algorithm is divided into pedestrian detection and tracking. In the detection part, the Faster-RCNN framework is used to detect pedestrians; and in the tracking part, we use the Person-ReID method to track pedestrians, converted the pedestrians tracking problem to feature extraction and matching problem between different frames, and improved the tracking results effectively according to the deep learning framework. According to the experiment results on the simple standard dataset and RGB-D People dataset, the tracking mAP of our algorithm get 92.51% and 76.9%.

1. Introduction

Pedestrian tracking algorithm is an important research topic in computer vision field. Early target tracking using the template matching method[1], which put forward by Lucas and Kanade in 1981. Many scholars have made many contributions in this field before, and many tracking algorithms have been proposed.

The tracking based on regional[2]is divided the human body into multiple regions, through to tracking the small areas to track the whole, but the algorithm will be failure once the pedestrian occlusion appears. The tracking based on contour[3] achieves tracking by using a closed contour of human body and constantly updated, but it is difficult to initialize the contour. The tracking based on features[4] is to extract the feature of every image, and tracking by corresponding the features of the frame before and after, but the robustness of algorithm mainly depends on the feature extraction methods of system selection. Model-based tracking[5,6] describe the structure of the body details by using the linear, projection or 3D model, but the system requires a lot of calculation parameters and calculation, which leads to low efficiency. All the tracking method has its own advantages and disadvantages.

Since the CNN has attained the remarkable achievements by using Deep Learning to ImageNet Dataset in 2012 by Hinton team, the Deep Learning gets attention quickly and has been widely used in various research fields. Wang et al.[7] using automatic encoder initializes the network parameters and...
constantly fine-tune online of the training network classification layer to track the target. Ma et al.[8] through visual analysis, fusion multiple features to improve the effect of tracking, and achieve the end-to-end tracking all convolution network[9]. And Wang et al.[10] come up with building the recognition model based on the deep learning to assist the multiple target tracking algorithm. Considering the better performance of deep learning, this article also apply deep learning framework to pedestrian tracking algorithm.

2. Pedestrian Detection
Pedestrian detection is the premise and key of pedestrian tracking algorithm, the target of the detection is to extract and accurate positioning the pedestrians who appear on the video sequence. This article uses Faster RCNN[11] as pedestrian detection network, the process is shown in figure 1.

![Figure 1. Faster RCNN detection process.](image)

Input the entire images to the CNN for feature extraction, to obtain the feature map of the images. Then treat the extracted characteristics as a 51 * 39 of 256 channels images, to input them into the RPN network and form the Region proposal. The essence is the sliding windows on the extracted feature map on CNN, to get multi-scale and multi length-width ratio of the Region proposal. Namely, consider three scales and length-widths ratio respectively, a total of nine possible candidate anchor, of the characteristics of the images of each position.

Mark the anchors that the candidate area and ground truth overlapping proportion is the largest as the foreground, then choose the anchors with bounding box overlapping ratio greater than 0.7 to notes as foreground, which the overlapping ratio less than 0.3 as the background. Finally, output the probability of foreground and background from the 256D characteristics.

Treat the extracted ROI area as a fixed value, with the extracted proposal based on RPN network input to Fast RCNN, to detect and identify targets in the proposal.

During the test, the trained feature model and test dataset are stored in the network, then run the network to detect the pedestrians of the test dataset. In the end, save the detected pedestrians coordinates.

According to the Faster RCNN network, we define the loss function as:

\[ L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \].

While, \(i\) is an index of the mini-batch anchor, \(p_i^*\) is the prediction probability of the target anchor. If the anchor is positive, the ground-truth label \(p_i^*\) is equal to 1, and if the anchor is negative, the \(p_i^*\) is equal to 0. \(t_i\) is a vector, which means the four parametric coordinates of the bounding box, \(t_i^*\) is the coordinate vectors corresponding with the positive anchor of ground-truth box.

Classification loss \(L_{cls}\) is a logarithmic loss function of two categories (target vs. non-target), which defined as:

\[ L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1-p_i)(1-p_i^*)] \].

For regression loss \(L_{reg}\), we use:
Among, $R$ is the robust loss function:

$$L_{\text{reg}}(t_i, t^*_i) = R(t_i - t^*_i).$$

(3)

3. Pedestrian Tracking

The function of pedestrian tracking algorithm is to match the detected pedestrians between adjacent frames, thus obtaining the trajectory of pedestrians in the field of camera vision.

The pedestrian tracking section in this paper adopts the Person Re-identification algorithm based on CNN deep learning framework. The Person Re-ID algorithm [12,13] judges whether there is a pedestrian in the images or the video sequences by using the computer vision technology.

In the process of tracking, we input the test the video frames and corresponding Bounding Box of the pedestrian detection coordinates into the ResNet-50 as the CNN network model[14, 15] to extract the characteristics. Given an input image, will produce 1024 channels feature maps. Through the feature maps, transform the characteristic of pedestrians by using 512 * 3 * 3 convolution layer.

To finally find the target pedestrians, we compare the extracted features with target pedestrians through the identification net. Firstly, we will use the network of Rol Pooling layer to obtain the 1024 * 14 * 14 areas from feature maps. And then put them into the ResNet-50 network. At the last, integrate them into 2048-dimensional feature vector. By feature extraction and comparison to achieve the goal of the pedestrian tracking. In order to optimize the track to pedestrians, we will calculate the cosine similarity of the predicted pedestrians with target pedestrians by using the integrated characteristics to project to 256d normalization vector space. The tracking process diagram as shown in figure2:

![Figure 2. The flow chart of pedestrian tracking system.](image)

This method chooses to use online approximation for optimization. Note the feature that has been labeled in the mini-batch as $x \in \mathbb{R}^D$, which D is the feature dimension, and a lookup table (LUT) $V \in \mathbb{R}^{Q \times D}$ is retained to store all the identities that have been labeled. In the forward propagation, $V^T x$ is used to calculate the cosine similarity between the samples in the mini-batch with the identity that has already labeled. In the backward propagation, assumption the target pedestrian is classified as t, then update t in LUT as: $V_t \leftarrow V_t + (1-\gamma)x$, where $\gamma \in (0,1)$. For the features that are unlabeled, we use the method of loop queue to save them in the most recent mini-batch. The unlabeled identity is denoted as $U \in \mathbb{R}^{Q \times D}$, where Q is the queue length. Then we can also compute the cosine similarity by $U^T x$. After each iteration, the new feature vector is placed into the queue and eliminated the outdated. That, we define the probability that the eigenvectors x as the i-th class of pedestrians is:

$$p_i = \frac{\exp(v_i^T x / \tau)}{\sum_{j=1}^L \exp(v_j^T x / \tau) + \sum_{t=1}^Q \exp(u_t^T x / \tau)}.$$

(5)

Similarly, in the circular queue, we define the probability that the eigenvectors x as the i-th unlabeled pedestrians is:

$$q_i = \frac{\exp(u_i^T x / \tau)}{\sum_{j=1}^L \exp(v_j^T x / \tau) + \sum_{t=1}^Q \exp(u_t^T x / \tau)}.$$

(6)

Where $\tau$ controls the degree of probability distribution.
4. Experimental results and analysis

Our experiments uses the RGB - D People Dataset[16,17], it contains 3000 + RGB - D video frames, is made up of three Kinect device sensors to obtain from university hall. These data contains different direction and different occlusion degree of walking and standing.

Our experiments is using one of the sensors of video frames of the dataset for pedestrian detection and pedestrian tracking as the training samples, at the same time, the other two sensors of the video frames are as the test sample of this algorithm. In order to have a comparison, we also use the standard dataset of TB – 100 of Jogging image sequence to do experiment, and compared with the experimental results of the RGB - D People Dataset.

In the figure 3 and figure 4, we shows the two datasets of some tracking video frames respectively. By tracking images, Jogging image sequence of the two goals in the frame are occlusion from the 47th to the 67th frame, but the tracking accuracy is still ideal, behind the 87th frame and other few frames that displayed are not lost the target. While, in RGB - D Dataset video frame, the tracking accuracy is not ideal, the pedestrians that appear in the nearby are tracking higher accuracy, but when the pedestrians are in the strong light and appear in the distance of the camera, our tracking would be easy to track failure. At the same time, we are given in table 1 about the detection and tracking parts compared with the two datasets. From the table 1, we can also be seen that due to the Jogging are only two goals in image sequence, and the occlusion is not serious, the detection and tracking mAP of the pedestrians are better than the RGB-D People Dataset. In the end, we also calculate the tracking accuracy of the two targets of the Jogging image sequence, compared with the previous model, the results as shown in table 2.

Table 1. The comparison of experimental results between two datasets.

| Dataset       | Detection mAP | Tracking mAP |
|---------------|---------------|--------------|
| Jogging       | 99.8%         | 92.51%       |
| RGB-D         | 90.7%         | 76.9%        |

Table 2. The comparison of the tracking accuracy of Jogging Dataset.

| Methods       | target1       | target2       |
|---------------|---------------|---------------|
| Wang et al[10]| 88.6%         | 87.62%        |
| Ours          | 93.48%        | 91.53%        |

Figure 3. The tracking frames of the Jogging Dataset.

Figure 4. The tracking frames of RGB-D People Dataset.

5. Conclusion

This paper puts forward the implementation of the pedestrian tracking algorithm based on deep learning framework. The experimental results show that the algorithm of pedestrian tracking accuracy has obviously improved than before algorithm. Under the condition of the pedestrian occlusion and target disappear again, the method achieves the tracking accuracy effectively by feature extraction and matching. Compared with the traditional pedestrian tracking method, this algorithm can achieve the continuous tracking. Especially, the experimental results of the standard database and RGB - D People database can prove the effectiveness of the proposed method. In the future, we will use the depth information further, to make up for the RGB image since the pedestrians are occlusion seriously, and improve the tracking accuracy.
Acknowledgments
This work was financially supported by China Nature Science Fund Project 61472196, 61672305, 61702295 and Shandong Nature Science Fund Project ZR2015FM012.

References
[1] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision [C]. Proceedings of the 7th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: ACM, 2: 674-679(1981).
[2] V Hadjioannou, CX Mavromoustakis, G Mastorakis, EK Markakis, D Valavani. Context awareness location-based Android application for tracking purposes in assisted living [C]. IEEE International Conference on Telecommunications and Multimedia, 1-7(2016).
[3] M Demi. Contour Tracking with a Spatio-Temporal Intensity Moment [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38 (6) :1141-1154(2016).
[4] P Men , C Wang, H Guo, J An. Multi-feature Fusion for the Edge Tracing of PolSAR Oil Slick [C]. IEEE International Conference on Information Science and Control Engineering, 478-482(2016).
[5] J Stuhmer, S Nowozin, A Fitzgibbon, R Szeliski, T Perry. Model-Based Tracking at 300Hz Using Raw Time-of-Flight Observations [C]. IEEE International Conference on Computer Vision, 3577-3585(2015).
[6] B Liu, O Jesorsky, R Kompe, Robust real-time multiple object tracking in traffic scenes using an optical matrix range sensor[C]. IEEE International Conference on Intelligent Transportation Systems, 742-747(2007).
[7] Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking [C]. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 809-817(2013).
[8] Ma C, Huang J B, Yang X K, et al. Hierarchical convolutional features for visual tracking [C]. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 3074-3082(2015).
[9] Wang L J, Ouyang W L, Wang X G, et al. Visual Tracking with Fully Convolutional Networks[C]. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 3119-3127(2015).
[10] Wang Huiyan, Yang Yutao, Zhang Zheng, Yan Guoli, Wang Jingqi, Li Xiaolan, Chen Weigang, Hua Huang, Wang Xun. Multi-pedestrian tracking algorithm assisted by in-depth learning [J]. Chinese Journal of Image and Graphics. 22 (3): 349-357 (2017).
[11] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]. IEEE International Conference on Neural Information Processing Systems, 39(6):91-99(2015).
[12] Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. Image and Vision Computing, (2014).
[13] L. Zheng, Y. Yang, and A. G. Hauptmann. Person reidentification: Past, present and future. arXiv preprint arXiv:1610.02984, (2016).
[14] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, Xiaogang Wang. Joint Detection and Identification Feature Learning for Person Search[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 3376-3385(2017).
[15] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850, (2016).
[16] L. Spinello, K. O. Arras. People Detection in RGB-D Data[C]. IEEE International Conference on Intelligent Robots and Systems (IROS), 38(2):3838-3843(2011).
[17] M. Luber and L. Spinello and K. O. Arras People Tracking in RGB-D Data With On-line Boosted Target Models[C]. IEEE International Conference on Intelligent Robots and Systems (IROS), 30(1):3844-3849(2011).