Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080545

Robust Detection and Tracking Algorithm of Multiple Objects in Complex Scenes

Hong-Yu Hu[1,2,*], *Zhao-Wei Qu*[1], *Zhi-Hui Li*[1] *and Qing-Nian Wang*[2]

[1] College of Transportation, Jilin University, Changchun 130022, China
[2] State Key Laboratory of Automobile Dynamic Simulation, Jilin University, Changchun 130022, China

Received: 17 Sep. 2013, Revised: 15 Dec. 2013, Accepted: 16 Dec. 2013
Published online: 1 Sep. 2014

Abstract: Detection and tracking of multiple targets in complex environment with an uncalibrated CCD camera is developed in this paper. 1) A background initialization algorithm based on clustering is presented. All stable non-overlapping intervals in the temporal training sequence of each pixel are located as possible backgrounds by slip window; then the background interval is obtained from the classified data set of possible backgrounds by unsupervised clustering. 2) Moving multi-targets are tracked through integration of the motion and shape features by Kalman filter model. In order to ensure the continuity and the stabilization, occlusion processing is performed. The proposed approach is validated under real traffic scenes. Experimental results show that detection and tracking algorithms are robust and adaptive and could be well applied in real-world.

Keywords: Intelligence Transportation System, Background Initialization, Motion Tracking, Gaussian Mixture Model.

1 Introduction

Vision-based detection has become an efficient technique support for control and management of urban traffic [1]. Eliminating background, extracting moving objects efficiently and adaptively in complex environment is necessary for a robust detection and tracking system. In the video sequence, the background is more stable to detect than foreground. Background initialization is the fundamental component of background model. Background models such as Gaussian Model [2] and Mixture Gaussian Model [3] assume that the initial background could be obtained using a short temporal training sequence with foreground objects exist. However, these models could not deal with some practical situations including moving foreground objects efficiently. Temporal average method [4] considers the average of pixel values in temporal sequence is the background, but it could not get over the influence of continual moving objects in the training set; Median method has been employed for initializing the background model. But the Median could tolerate up to only 50% outliers, which cannot satisfy the requirements of some complicated environments. Furthermore, reference [5] assumes that the background pixel always be stable. All stable non-overlapping sub-intervals in the training sequence could be located for each pixel as possible background, and the most stable sub-interval in temporal is denoted as the background interval based on heuristic principle. These methods work well when the proportion of foreground objects in the training set is above 50%. While they could not overcome the impact by the slow motion of massive objects.

In the field of motion tracking, several methods have been studied. Koller[6] uses 3D model to track vehicle at intersections which depends on detailed geometric object models. Paragios [7] tracks moving objects based on active contour models, but this approach needs a suitable design for tracking initialization, which is easy to cause some significant measurement errors; Kanhere [8] presents an automatic technique for detecting and tracking vehicle with vehicle base fronts (VBFs) at a low angle, even in the presence of severe occlusion. Chachich [9] uses color signatures in quantized RGB space for tracking vehicles. However, the process of tracking could not well deal with the case of occlusion. The limitations of the method would result in a serious impact for its universal application. Zhang [10] presents a multilevel framework consisting of the intraframe, interframe, and tracking levels to handle vehicle occlusion. Tsai [11] and

* Corresponding author e-mail: dayuhoo@gmail.com

Yang [12] propose multiple pedestrians tracking with occlusion handling in dynamic scenes based on color models. These algorithms were studied from different angles, and achieved some good results on occlusion processing.

We propose a novel algorithm for detection and tracking of multiple objects. A background initialization algorithm based on clustering classifier is developed, all stable non-overlapping intervals in the temporal sequence of each pixel are located as possible backgrounds by slip window, and then the background interval is obtained from the classified data set of possible backgrounds by unsupervised clustering; multiple targets tracking algorithm is presented, which is modeled by extended Kalman filter combined with occlusion processing. In experiments, the proposed algorithm is tested under different traffic scenes. The results show that the method is robust and self-adaptive in continuous frames even in the case of temporal occlusion.

2 Background Initialization

Generally, if a pixel is the background point in the image, its value should keep long-term stable relatively, and it would not be changed obviously until foreground objects passing. Therefore, for each pixel, a set of values would be observed in the temporal sequence, and some stable non-overlapping intervals in the sequence could be located as possible background candidates.

The initialization background algorithm consists of two steps: the first is to obtain all stable non-overlapping intervals in the temporal training sequence for each pixel as possible background; the second is to get the background sub-set from the classified data set by unsupervised clustering to realize the background initialization.

Let $\{x_i|\ i = 1 \cdots \mathrm{N}\}$ denote N observed values of the same pixel. In order to obtain all stable non-overlapping intervals, a slip window with the fixed initial length is defined in the temporal sequence. Continuous observed values are located at first, the number of which is equal to the initial window length. If the variation of any two observed values is in a permissible range, the next observed value is pulled into the window, and the length of the window increase by 1; if the variation of observed values is out of the permissible range and the length of the window is larger than the initial length, observed values in the window are marked as a stable interval, and then renew the slip window beginning with the value after the last interval marked. If the variation of observed values is out of the permissible range and the length of the window is not larger than the initial length, the length does not alter, and the window moves one value. A set of stable intervals can be obtained when the window slips all observed values of the pixel. Let $L =< l_1, \cdots, l_k >$ represent the stable interval set, and the interval $l = \{x_i, \cdots, x_j\}$ in it should be satisfied:

$$\omega \leqslant j - i,\ \ \forall(s,t) \text{—} x_s - x_t| \leqslant \delta_{\max} \tag{1}$$

where ω is the initial length, $\delta_{\max}$ is the largest variation permitted. We set $\omega = 6, \delta_{\max} = 5$ in this paper.

Generally, the median of interval is robust, a classify set $< s_1, \cdots, s_k >$ is made up of medians of all stable intervals, and the median is expressed as:

$$s_j = median(l_j),\ 1 \leqslant j \leqslant k \tag{2}$$

The distance between medians expresses the similarity between intervals. The classify set could be classified into some sub-class according to the distance. Because the background keeps long-term stable, the distance between interval elements should be less than the largest variation $\delta_{\max}$. A circular regions is constructed with s_j as the center, $\delta_{\max}$ as the radius, the number of samples located in the region is denoted as "density". According to the third assumption, the interval which has the highest density is the background interval. And the sample which is nearest to the center of the background interval is the initial background point.

The algorithm is valid unless the density is less than 2 or the distance between samples in the interval is larger than $\delta_{\max}$. If the density is 0 (no stable interval in the temporal sequence), this may be caused by the continuous change of the background (trees rocking). The median method is used under this condition. If the density equal 1, there may be two reasons: firstly, the temporal training sequence is completely covered by the background; Secondly, it's covered completely by still objects. The median method is used for the first case to initialize the background. If the temporal sequence is completely covered by still objects, all of methods are invalid. It is not taken into account in the paper. If the distance between any two elements in the interval of $< s_1, \cdots, s_k >$ is more than $\delta_{\max}$, the interval which has the most stable degree is chosen as the background interval, and the median of it is chosen as the background point, the stable degree is expressed as:

$$SD_j = \frac{l_{j,length}}{l_{j,\delta}},\ 1 \leqslant j \leqslant k \tag{3}$$

In (3), j is the number of the interval, $l_{j,length}$ is the length of interval j , $l_{j,\delta}$ is the variance of interval j . In order to overcome the influence of the variety of illumination, weather and other environment facts, the background should be updated reasonably and timely. Based on the initialization background algorithm, background update model depending on Stauffer's method [2] combined with object's spatial property was utilized to detect moving objects.

3 Motion Tracking

When moving objects are extracted by background subtraction, Kalman filter is used to predict the location of moving objects and the shape feature of a target in frame sequences is utilized to reduce the cost of matching operation. The state vector of the moving target in the k th frame can be defined as follow:

$$S_k = [C_{x,k}, C_{y,k}, W_k, H_k, v_{x,k}, v_{y,k}, \Delta W_k, \Delta H_k]^T \quad (4)$$

And the observation vector could be defined as: $O_k = [\hat{C}_{x,k}, \hat{C}_{y,k}, \hat{W}_k, \hat{H}_k]^T$, So Kalman filter model could be expressed as:

$$S_k = AS_{k-1} \quad (5)$$

$$O_k = BS_k + \delta_K \quad (6)$$

In (4), $(C_{x,k}, C_{y,k})$ is the geometric center of the moving target in the k th frame. $(v_{x,k}, v_{y,k})$ is the velocity which is calculated by the location displacement of interval frames. W_k is the width and H_k is the length of the boundary box of the target and $\Delta W_k, \Delta H_k$ are varieties of them respectively. The shape feature is reflected by these parameters. A is the state transform matrix and B is the observation matrix. δ_K is the noise of the system. In the process of calculation, we set:

$$A = \begin{bmatrix} 1&0&0&0&T&0&0&0 \\ 0&1&0&0&0&T&0&0 \\ 0&0&1&0&0&0&T&0 \\ 0&0&0&1&0&0&0&T \\ 0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&1 \end{bmatrix} \quad (7)$$

$$B = \begin{bmatrix} 1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \end{bmatrix} \quad (8)$$

T is the temporal interval. In order to track a moving object in consecutive frames, the previous state of the tracked target should be predicted to estimate parameters as: $S_{\langle k+1|k\rangle} = AS_k$, $S_{k+1} = S_{\langle k+1|k\rangle} + \theta_{k+1}(O_{k+1} - BS_{\langle k+1|k\rangle})$. θ_{k+1} is the Kalman gain matrix, which is the modifiability of the state prediction. Since more than one vehicle is detected inside the search region, every detected foreground object should be compared with tracked targets to select the best candidate. We realize the matching between observed state values of the target in the present frame and predicted state values of the previous frame by minimum-distance method. If matching cost between one of objects in the present frame and the predicted target which is tracked in previous frames is least of all, and less than a threshold, it means the target is tracked successfully; Else, it maybe occur occlusion during the target moving process.

Generally, the n th moving object would not appear or disappear suddenly inside the detection region, and it would move with constant velocities by the limitation of acceleration. If the tracked target doesn't match with any observed objects at k th frame, it may be occluded by background or other objects close-by. These are cases where objects are hidden by background (trees, telegraph pole, etc.) in the process of moving or where objects move very close to each other. Thus the contours of the object will interfere and the state estimation process would be confused. Therefore, for the temporary occlusion of the motion tracking process, we use Gray Model (GM) to achieve the target state prediction. The method could be made quantitative prediction of the future state of the system depending upon historical data. GM (1, 1) is the most classic gray model. The model is the use of some new data generated from the raw data accumulated. To a certain extent, the randomness of the original data is weakened and the varieties of data could be well reflected. So it could well solve the nonlinear state prediction. Suppose there is an original data sequence U_0 of n raw data observations: $U_0 = [u_0(1), u_0(2), u_0(3) \cdots, u_0(n)]$. A new sequence could be generated by accumulating the original data: $U_1 = [u_1(1), u_1(2), u_1(3) \cdots, u_1(n)]$, where $u_1(k) = \sum_{i=1}^{k} u_0(i)$, $(k = 1, 2, 3, \cdots, n)$. The corresponding differential equations of GM (1,1):

$$\frac{dU_1}{dt} + aU_1 = b \quad (9)$$

where $\hat{a} = (a, b)^T$ is the parameters to be estimated, which could be calculated by least squares:

$$\hat{a} = (B^T B)^{-1} B^T U \quad (10)$$

where $B = \begin{bmatrix} -\frac{1}{2}[u_1(1) + u_1(2)] & 1 \\ -\frac{1}{2}[u_1(2) + u_1(3)] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}[u_1(n-1) + u_1(n)] & 1 \end{bmatrix}$,

$U = \begin{pmatrix} u_0(2) \\ u_0(3) \\ \vdots \\ u_0(n) \end{pmatrix}$

The prediction model could be constructed: $\hat{u}_0(k+1) = [\beta - \alpha u_0(1)]e^{-a(k-1)}$, where $\alpha = \frac{a}{1+0.5a}$, $\beta = \frac{b}{1+0.5a}$.

If the tracked target could not be tracking at k th frame, we keep the unmatched object with labels firstly,

(a) Source frame (b) Initial background extracted by our method proposed (c) Foreground objects detected

Fig. 1: Multiple objects detection results based on background subtraction(From the top row to the bottom row are test sequence 1/ test sequence 2/ test sequence 3.

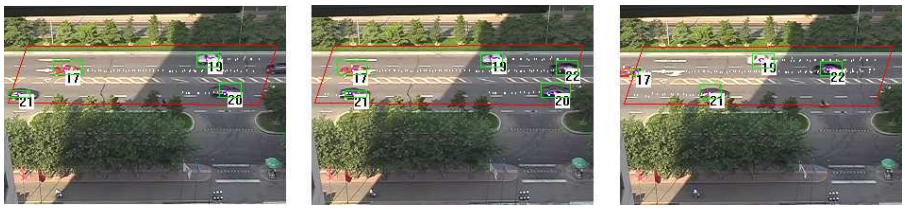

Fig. 2: Multiple vehicles tracking in the bidirectional roadway. From left to right are frame 1320, frame 1324, frame 1332 in the test sequence 4.

and construct a temporal occlusion list for the object, which save state information from $k-1$ th frame when the occlusion occurs. The GM (1, 1) model is constructed by the latest m data before occlusion. Then we update the motion and the shape parameters and predict the state with the past known data continuously in the process of the occlusion. When the new data generated, the oldest one of the m data in the sequence would be replaced. If in an occlusion time threshold T fames the target matches again, it means there has been a real occlusion, and we move the matched object of the current frame into the tracking list; if the target doesn't match with any objects in T frames, it means the target leaves the detection region.

4 Experimental Results

Several video sequences with complex environment were obtained from realistic scenarios by an uncalibrated CCD camera to test the validity of the proposed approach. The algorithm was implemented with VC.NET in a computer with Pentium 2.4GHz processor and 256M DDR.

For validating background initialization algorithm, three different traffic scenarios are tested in Fig. 1. In each video sequence, every two frame in the video sequence is chosen as a training sample. We extracted a training temporal sequence of 100 frames with existing moving objects for initial background. Sequence 1 was captured from an evening peak of an expressway. In the process of background initialization, the proportion of existing foreground objects in the training sequence is almost above 50%. Sequence 2 is an unsignalized intersection video captured at a location about 10 meters

Fig. 3: Pedestrian tracking under background occlusion of trees. From left to right are frame 1170, frame 1184, frame 1204 in the test sequence 5.

Fig. 4: Two person tracking with motion occlusion. From left to right are frame 1000, frame 1020, frame 1040 in the test sequence 6.

high. In this scenario, multiple pedestrians move slowly with together. Sequence 3 was urban mixed traffic captured from a location about 20 meters high. From the processing results, good detection performances of background and foreground extracted are showed.

Fig. 2- Fig. 4 are tracking results under three different traffic scenes indoor and outdoor. From the video processing, targets were segmented with fixed bounding boxes first, then targets were labeled to show motion tracking evidently. Fig. 2 shows that multiple vehicles are tracking in bidirectional roadways of 4 lanes. Fig. 3 is the pedestrian tracking with background occlusion with trees. Fig. 4 shows the result of two persons moving together with motion occlusion. Results show that our algorithm is robust and adaptive, could achieve consecutive, stable multiple targets motion tracking even in the case of background occlusion and motion occlusion. The algorithm processed every 2 frames (12 frame per second), which is a satisfactory real time speed.

5 Conclusions

In this paper, in order to obtain initialization background from the temporal sequence with existing foreground objects, a classify data set is constructed by the median values of each stable sub-interval in the training sequence; then a background sub-set is obtained from the classify data set by unsupervised clustering. The method proposed could overcome the influence of massive objects moving slowly. Multi-targets were tracked through integration of the motion and shape features by Kalman filter modeling, and occlusion processing based on GM(1,1) is considered to improve the robustness of tracking. At last, we validate the proposed approach under real traffic scenes. And the results show the algorithm is robust and has better self-adaptability.

Acknowledgements

This work is partly supported by the National Science Foundation of China (No.51108208, No.51278220), the Postdoctoral Science Foundation Funded Project of China (No.20110491307, No.2013T60330), the Fundamental Research Fund for the Central Universities of China (No.201103146) and the Science and Technology Development Project of Jilin Province of China (No.20130522121JH).

References

[1] Buch, N., Velastin, S.A. and Orwell, J. A review of computer vision techniques for the analysis of urban traffic. IEEE Transactions on Intelligent Transportation Systems, **12**, 920-939 (2011).

[2] Wren Christopher Richard, Azarbayejani Ali, Darrell Trevor,etc. Pfinder: real-time tracking of the human body[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, **19**, 780-785 (1997).

[3] Stauffer Chris, Grimson W. E. L. Adaptive background mixture models for real-time tracking[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 246-252 (1999).

[4] Haritaoglu, I., D. Harwood, and L.S. Davis, W4: Real-Time Surveillance of People and Their Activities. IEEE Trans. Pattern Analysis and Machine Intelligence, **22**, 809-830 (2000).

[5] Wang H. Z., Suter D. A novel robust statistical method for background initialization and visual surveillance. In: Computer Vision - Accv 2006, Pt I. Berlin: Springer-Verlag Berlin, 328-337 (2006).
[6] Koller D., K. Dandilis, H. H. Nagel. Model Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes. International Journal of Computer Vision, **10**, 257-281 (1993).
[7] Paragios N., Deriche R.. Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision. J. Visual Commun. Image Represent, 249-268 (2002).
[8] Neeraj K. Kanhere, Stanley T. Birchfield, Wayne A. Sarasua. Vehicle Segmentation and Tracking in the Presence of Occlusions. Transportation Research Record: Transportation Research Board of the National Academies, 89-97 (2006).
[9] Chachich, A. A. Pau, A. Barber, K. Kennedy, E. Oleiniczak, J. Hackney, Q. Sun, E. Mireles. Traffic sensor using a color vision method. Proceedings of the International Society for Optical Engineering, **2902**, 156-164 (1997).
[10] Zhang W., Wu Q. M. J., Yang X., et al. Multilevel framework to detect and handle vehicle occlusion[J]. IEEE Transactions on Intelligent Transportation Systems, **9**, 161-174 (2008).
[11] Tsai V. J. D.. A comparative study on shadow compensation of color aerial images in invariant color models[J]. IEEE Transactions on Geoscience and Remote Sensing, **44**, 1661-1671 (2006).
[12] Yang T., Li S. Z., Pan Q., et al. Real-time multiple objects tracking with occlusion handling in dynamic scenes[C]. in Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition., **I**, 970-975 (2005).

Hong-Yu Hu, assistant professor of department of traffic information engineering and control, College of Transportation, Jilin University, China. His research interests include intelligent transportation system and traffic behavior analysis.

Zhao-Wei Qu, professor and Ph.D supervisor of department of traffic information engineering and control, College of Transportation, Jilin University,China. His research interests include traffic information collection and traffic control.

Zhi-Hui Li, associate professor of department of traffic information engineering and control, College of Transportation, Jilin University, China. His research interests include image processing and computer-aided application technology.

Qing-Nian Wang, professor and Ph.D supervisor of State Key Laboratory of Automobile Dynamic Simulation, Jilin University,China. His main research interests are: theory and control technology of hybrid vehicle, vehicle power parameter matching and optimization.