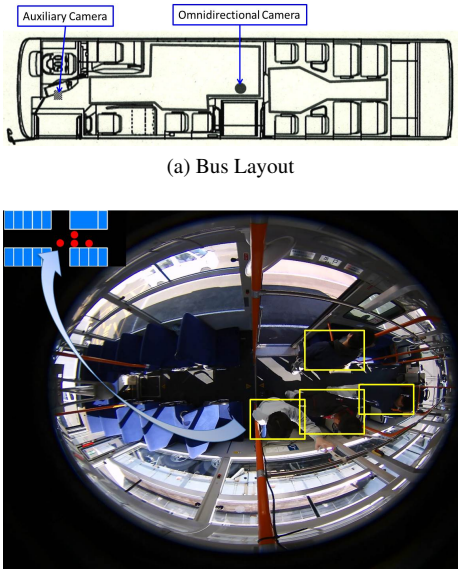
An Omnidirectional Vision System for Bus Safety Surveillance

Dao Huu Hung, Hideo Saito
Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama,
Kanagawa, 223-8522, Japan
hungdaohuu, saito@hvrl.ics.keio.ac.jp

Keiichi Yamamoto, Hiromitsu Sato
Mitsubishi Fuso Truck and Bus Corp.
10 Okura-cho, Nakahara-ku, Kawasaki-shi,
Kanagawa, 211-8522, Japan
keiichi.yamamoto, hiromitsu.h.sato@daimler.com

Abstract

Sudden brake seems to be unavoidable in driving practice to prevent from accidence. The sudden brake may often cause falling risks to on-board passengers in public transport and even threaten the lives of the senior. Thus, this paper presents an omnidirectional vision system to assist bus drivers in ensuring the safety for passengers. To this end, we configure an auxiliary camera in the boarding gate to recognize the age of passengers and an omnidirectional camera on the ceiling of the bus to detect and track the passengers. Their trajectories and positions are mapped into a bus layout, together with their age information transferred from the auxiliary camera. This bus layout is shown on a small display in front of bus drivers. Instead of looking to the mirror to guess and gather information of passengers, the driver can collect such necessary information for safe driving. We set up experiments on a real bus and demonstrate the system by preliminary results.



(a) Bus Layout

(b) Panoramic Image captured by Omnidirectional Camera and a virtual bus layout in the top-left corner

Figure 1: Our proposed omnidirectional vision-based bus surveillance system. Passengers' positions are mapped into a bus layout in the top-left corner of Fig. 1b. Red color for standing passengers indicates potential hazard situations. Sitting passengers are denoted by yellow color.

1. Introduction

Video surveillance has a wide range of applications in both outdoor and indoor environments such as crowd density estimation [13, 7], crowd behavior analysis [25], public traffic monitoring [5], Person re-identification [3], and fall detection of the elderly [14, 15], etc. Among these, this paper concerns a video surveillance system to assist bus drivers to ensure the safety of on-board passengers, in particular, the safety of the senior.

There are many existing works in the literature of computer vision, dedicated to assisted driving. They include pedestrian detection [11], on-road vehicle detection [21], traffic sign detection and recognition [12], lane detection [4], blind spots elimination [10], hazard situation detection at railway-road crossing [20], and driver inattention recognition [9], etc. These applications primarily focus on creating comfortable driving environments for drivers, helping drivers prevent from accidence by automatically detecting obstacles and hazard situations possibly occurring outside the vehicle, and making a way towards an automated driving system. However, our system concentrates on safety surveillance of on-board passengers.

A few studies devoted to bus or public transport surveillance systems. Passengers are detected and tracked for vandalism deterrence by conventional camera [6]. The monitored view is quite restricted although optimal camera placement in a bus is studied by taking a trade-off between coverage area and cost [1]. To overcome the limited views,

an omnidirectional camera is utilized in this paper.

In public transport system, recognizing high-level events is also vital in dealing with anti-social behavior. High-level sequential observation modeling framework [16] is proposed to temporally correct uncertain sensing outputs, i.e. position and gender, before reasoning high-level events by rule-based event composition framework [17] and intelligent sensor information system [18]. One typical high-level event considered in these papers is a male and disgust passenger abusing or threatening drivers. Meanwhile, our paper focuses on understanding passengers' statuses, such as "*safe*" in case of sitting in saloon areas and "*hazard*" in case of for example, senior people standing on the bus floor.

The reason why we consider people standing on the bus floor as a hazard situation, especially the elderly, is because the sudden brake often causes falling risks to them. Sudden brake seems to be unavoidable and frequently occurs in driving practice for accidence prevention. The falling has profound implications and even threatens the lives of fallen senior passengers. This hazard situation possibly happens when passengers are sparse. But it is argued that high human density in a crowded bus makes it unlikely.

To this end, we configure a two-camera surveillance system on a real bus. One auxiliary camera is set up at the boarding gate of the bus for recognizing the age of passengers. One omnidirectional camera is mounted on the ceiling of the bus to monitor the whole interior of the bus as shown in Fig. 1. Their positions and trajectories are mapped into a bus layout, along with their age information transferred from the auxiliary camera. A small display in front of the bus driver shows the bus layout so that the bus driver keeps the weather eye on the statuses of all passengers. If all passengers are sitting, indicating "*safe*" status, a sudden brake seems to be possible. Otherwise, in particular, a senior passenger standing on the bus floor implies a potential hazard situation then the bus driver should avoid any unnecessary sudden brake. It is noted that we limit our discussion in this paper to the omnidirectional vision module.

Although the omnidirectional camera covers the full interior of the bus, it poses some challenges in people detection and tracking. Firstly, the panoramic image is highly distorted. As a result, HOG-based people detector [8] often produces both false negative and false positive results since image distortion makes HOG templates of background resemble that of human. Secondly, human shape and orientation change drastically across the panoramic image. For example as shown in Fig. 3, the human orientation varies 360 degree when people travel around the image. Hence, the exhaustive search for HOG human templates to detect people becomes intractable.

In bus surveillance, occlusion often occurs not only among passengers but also by the saloon. Lower body parts are often invisible. Unconstrained illumination, reflection and dynamic background in the window areas also pose difficulty. In addition, missing-frame phenomenon usually happens in practice of video surveillance. Consequently, it is hard to find a motion model to accommodate the passenger movement in case of missing-frame phenomenon.

State-of-the-art online visual tracking approaches [23] commonly assume known objects in the first frame. However in practice, when and where passengers starting to board the bus are unknown to build appearance models. Hence, we suggest using adaptive background subtraction [26] to segment moving foregrounds from image sequences. Foreground blobs are grouped together to detect people [14]. We track the detected person consistently in several frames, i.e. 5 frames, by simple Euclidean distance-based data association then take the tracked results (after 5 frames) to initialize an appearance model for tracking in the particle filter framework.

In particle filter tracking framework, a set of particles are re-sampled and propagated by a motion model considering the variations of 6 affine parameters [23] to find probable positions, dimensions and orientations of candidates. Designing an optimal motion model is extremely difficult to cope with unconstrained human movements, especially in consideration of missing-frame phenomenon. In top-view omnidirecitonal cameras, the movement of person far away from the camera is slow but seems to be quite faster when the person is near and under the camera (see Fig. 3) due to significant changes in human pose or shape deformation. If particles are allowed to propagate to a wide area, the tracker is likely to be drifted away to nearby similar appearance objects. Limiting propagated space forces particles not to reach where the objects are actually. Therefore, the hybrid particle filter tracking method that is a combination of color particle filter tracking and data association-based tracking is proposed to deal with missing-frame phenomenon. On the one hand, since associating data from noisy foreground-based people detection results is not robust, comparing appearance between templates and detected foreground regions may improve data association-based tracking. On the other hand, the set of particles is not only sampled from the set of particles in the previous frame but also partially sampled from foreground-based people detection results in the current frame. Because of the proposed re-sampling mechanism, the missing-frame phenomenon can be overcome.

The paper is continued with section 2 describing our proposed system. Experimental setup and our promising preliminary results are demonstrated in section 3. Surprisingly, performance of our hybrid particle filter tracking is better than that of SVD [24] and DLT [22] trackers. Finally, conclusions and future works come in the last section.

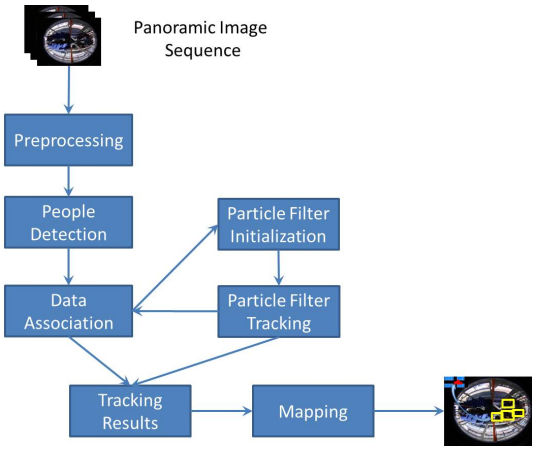


Figure 2: The flowchart of our proposed system

2. Our proposed system

This section describes our proposed system in detail, with the flowchart shown in Fig. 2.

In the preprocessing step, we set ROI for input images empirically within an enclosed ellipse to exclude dynamic background in window areas. The panoramic image sequences are segmented by adaptive background subtraction method [26]. Foreground images are enhanced by morphological operators such as open and close. Foregrounds are labeled by connected components, producing foreground blobs. Small foreground blobs, likely caused by noise, are removed. Finally, we have a pool of labeled foreground blobs for detecting people by the one presented in [14].

2.1. Data association-based tracking

It is assumed that detected people are represented by rectangular bounding boxes, as shown in Fig. 1. We perform tracking people by associating the detection results across two successive frames. Suppose that we detect M people $\{P_1^{t-1}, P_2^{t-1}, ..., P_i^{t-1}, ..., P_M^{t-1}\}$ in the frame $t-1$, where P_i^{t-1} represents the position of the person centroid i^{th} in the image coordinate. Our task is to associate these M people to N people $\{P_1^t, P_2^t, ..., P_j^t, ..., P_N^t\}$, detected in the frame t, based on Euclidean distances between the two sets of centroid positions. As a result, we have the Euclidean distance matrix $D = \{d_{ij} | 1 \leq i \leq M, 1 \leq j \leq N\}$, where $d_{ij} = |P_i^{t-1} - P_j^t|$. D is a symmetric matrix. The results of data association are determined by

$$\begin{aligned} pair(i,j) &= \underset{i,j}{\operatorname{argmin}}\ d_{ij} \\ s.t &\qquad d_{ij} < T_d \end{aligned} \tag{1}$$

where T_d a distance threshold, preventing from associating too far-away people. To track multiple people, both row i^{th} and column j^{th} are excluded from the matrix D and Eq. 1 is repeated until D becoming a 1D matrix.

If a track is consistently maintained in several frames, i.e. 5 frames, its appearance will be taken as a template for particle filter tracking initialization. Once the particle filter is initialized, the template is used to improve data association results by computing Bhattacharyya distance between its color histograms in HSV space.

$$d(H, H_T) = \sqrt{1 - \sum_i \frac{\sqrt{H(i) \cdot H_T(i)}}{\sqrt{\sum_i H(i) \cdot \sum_i H_T(i)}}} \tag{2}$$

where H and H_T color histograms of associated regions and the template, respectively. The track is updated if the Bhattacharya distance is smaller than a particular threshold, i.e. 0.55, indicating at least a half match in our experiment.

2.2. Hybrid color particle filter tracking

In the problem of visual tracking, a target at time t is represented by a state vector $X_t = \{x, y, sx, sy, \theta, \phi\}$ including position and scale changes in x, y coordinates, rotation θ, and deformation ϕ. The evolution of the state vector X_t is a Markov process, that is, X_t is modeled by Gaussian distributions around its preceding state X_{t-1}.

$$p(X_t | X_{t-1}) = \mathcal{N}(X_t; X_{t-1}, \Psi) \tag{3}$$

where Ψ a diagonal covariance matrix of state vector X_t. When a new image arrived, providing new measurements Z_t, we need to recursively estimate the state vector X_t of the target. In other words, the objective of visual tracking is to reconstruct the pdf $p(X_t | Z_{1:t})$ recursively by two states: prediction and update [2].

Firstly, we predict the target state at time t by using the motion model in Eq. 3. The prior pdf of the state at time t is attained by Chapman-Kolmogorov equation with an assumption of known pdf $p(X_{t-1} | Z_{1:t-1})$ at time $t-1$.

$$p(X_t | Z_{1:t-1}) = \int p(X_t | X_{t-1}) p(X_{t-1} | Z_{1:t-1})\, dX_{t-1} \tag{4}$$

Secondly, a new measurement Z_t at time t is used to update the prior by Bayesian rule

$$p(X_t | Z_{1:t}) = \frac{p(Z_t | X_t) p(X_t | Z_{1:t-1})}{\int p(Z_t | X_t) p(X_t | Z_{1:t-1})\, dX_t} \tag{5}$$

Analytical solutions for the recursive relations in Eqs. 4 and 5 are definitely intractable [2]. Approximate solutions by particle filters are preferable, in particular, both posterior pdf $p(X_t | Z_t)$ and the measurement pdf $p(Z_t | X_t)$ being

non-Gaussian [19]. The heart of particle filter tracking is to propagate a set of weighted particles $S = \{X_t^i, \omega_t^i | i = \overline{1, N}\}$ by using the motion model, each composing of one hypothetical state vector and corresponding weight such that $\sum_{i=1}^{N} \omega_t^i = 1$, to approximate the posterior pdf $p(X_t|Z_{1:t})$ as

$$p(X_t|Z_{1:t}) \approx \sum_{i=1}^{N} \omega_t^i \delta(X_t - X_t^i) \tag{6}$$

The weights are chosen by

$$\omega_t^i \propto \omega_{t-1}^i \frac{p(Z_t|X_t^i)p(X_t^i|X_{t-1}^i)}{q(X_t|X_{t-1}^i, Z_t)} \tag{7}$$

where $q(\cdot)$ the important density. It is noted in Eq. 7 that the weights are proportional to the state evolution and the measurement. In color particle filter tracking, the weights are evaluated based on Bhattacharyya distances in Eq. 2.

$$\omega_t^i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{d^2}{2\sigma^2}} \tag{8}$$

where d Bhattacharyya distance. Since small Bhattacharyya distances indicate high appearance similarity between the particles and the target, Eq. 8 assigns high weights to particles having small Bhattacharyya ditances.

However after a few iterations, some particles likely have negligible weights. Updating these particles is wasting time, leading to degeneracy phenomenon. One way to avoid the degeneracy phenomenon is to resample the set of particles based on its weights. During the filtering process, It had better to choose particles with high weights multiple times, rather than taking those with low weights, to propagate in the next time steps.

In visual tracking, it is very difficult to design an optimal motion model to accommodate with unconstrained movements of the target. In particular, our problem considers tracking people by a top-view omnidirectional camera. The movements of people who are far away from the camera seem to be slow. However, they tend to be faster when traveling near and under the camera, as shown in Fig. 3, due to pose variations and shape deformation. Missing-frame phenomenon often happening in practice also poses challenges for designing an optimal motion model. Choosing large covariance matrix Ψ in Eq. 3 is infeasible since the track is likely drifted away to nearby objects with similar appearance. But choosing small covariance matrix Ψ, prediction steps may not propagate particles to the regions where the target actually is. The problem becomes more challenging in case of missing-frame phenomenon.

Therefore to deal with missing-frame phenomenon, we propose hybrid color particle filter tracking. The hybrid method differs from the conventional particle filter tracking, first introduced in [19], only in the most important step, the resampling. The conventional way of resampling is to draw samples directly from the pool of particles in the previous time step. It has been shown in our experiment a poor performance in case of missing-frame phenomenon. Hence, our resampling strategy is to draw samples not only from the pool of particles in the previous time step but also partially from the data association tracking results. As shown in the flowchart in Fig. 2, both data association-based tracking and particle filter tracking work collaboratively. When a new frame arrives, data association-based tracking is firstly performed. The target template helps improve the robustness of data association. Subsequently, the results of data association will be used in resampling step of particle filter tracking to partially draw a new set of particles. Hence, the robustness against missing-frame phenomenon and sudden quick movements can be ameliorated. This tracking mechanism is able to improve overall tracking robustness.

2.3. Mapping to bus layout

The position mapping is very straightforward due to using top-view and fixed camera. The saloon regions are manually labeled. Let denote $S = \{S_1, S_2, ..., S_i, ..., S_n\}$ as a set of saloon regions where $S_i = \{X_i, Y_i, W_i, H_i\}$ the i^{th} saloon region at position $\{X_i, Y_i\}$, enclosed by a rectangle $\{W_i, H_i\}$. Suppose that a detected person is also enclosed by a rectangle $R_P = \{W, H\}$. The person is mapped into S_i region if R_P and S_i significantly overlap each other.

$$\frac{R_P \cap S_i}{R_P} \geq 0.75 \tag{9}$$

The detected person is mapped into i^{th} saloon region in the bus layout if Eq. 9 is satisfied. Their status is denoted by yellow color, indicating a safe status. Otherwise, it is denoted by red color, indicating a potential hazard situation of a person standing on the bus floor.

3. Experiments and preliminary results

3.1. Experimental setup

We set up an experimental environment in a real bus as shown in Fig. 1. Panoramic sequences are processed in a laptop PC, powered by chip set core i7 2.7GHz, Ram 16GB, and NVIDIA GPU with CUDA. Various people act as passengers getting on and taking off the bus. During the journey of the bus, these passengers perform a variety of actions such as standing, sitting, and changing seats, etc. We recorded 35 video samples in total for testing which are divided into Simple and Complex sets. The former comprises of 15 samples. Each contains one or several people getting on the bus and subsequently taking off the bus. The bus is empty at the beginning and moving after taking people. By contrast, the latter is more challenging, due to initially

Table 1: Preliminary results on the Simple set of our method are quantitatively compared with those of SVD [24] and DLT [22].

	GT	Our Method	SVD	DLT
1-1	1	0	0	0
1-2	1	1	0	0
1-2m	1	1	0	0
1-2s	1	1	0	1
1-3	1	1	0	0
1-3_	1	1	0	0
1-3m	1	1	0	0
1-3s	1	1	0	0
1-5	1	1	0	0
1-5m	1	1	0	0
1-7	1	1	0	0
1-7m	1	1	0	0
2-2	2	2	0	0
2-6	2	2	0	0
3-3_3-6	3	0	0	0

having several people and being recorded under strong sunshine, although the bus is stationary, One or several people get on and take off the bus afterwards, meanwhile the others stay relatively non-moving.

3.2. Preliminary results

In this section, we present the preliminary results of testing our system on the in-house video collections. We also perform a quantitative comparison between our hybrid particle filter tracking method and some state-of-the-art methods, i.e. SVD [24] and DLT [22]. We measure the number of successful tracks, which are defined as successful continuous tracks from the boarding to exit gates. Table 1 shows the ground truth, our results, along with SVD and DLT results[1], tested on the Simple set. While performance of our method is quite good, both SVD and DLT surprisingly produce very poor performance. Fig. 3 visually illustrates the tracking results of our method, in comparison with those of SVD and DLT. Both SVD and DLT commonly fail to track people when their bodies are bended to sit down or when people travel under the omnidirectional camera, causing their appearance to change significantly and drastically. In addition, missing-frame phenomenon is overcome by our method, as shown in Fig. 3, meanwhile it forces SVD and DLT to all failure.

However, both SVD, DLT and our method are failure when tested on the Complex set. Since the bus is quite crowded at the beginning with relatively non-moving passengers, detecting people based on foregrounds is intractable. Even though the human appearance is manually initialized, the tracking performance is still poor. It is because under extreme lighting conditions, human appearances are very similar. It causes all trackers to drift away.

[1]We reuse the source codes, provided publicly by the authors

4. Conclusions and future works

We have presented an automated omnidirectional vision surveillance system for ensuring the safety of on-board passengers, in particular, the elderly passengers. The positions are mapped into a bus layout for interpreting their safety status that will be shown on a display, together with their age information, recognized by an auxiliary camera. Hence, bus drivers will easily and comfortably take necessary information of on-board passengers, in order to ensure their safety. Our experimental results show that state-of-the-art tracking algorithms are not robust enough to our in-the-wild samples, suggesting many rooms of improvement. An introduction of infrared sensors may alleviate low light challenges. Retrained HOG people detector by samples collected from omnidirectional cameras along with polar raster scan strategy is able to ameliorate performance.

5. Acknowledgment

This work was supported in part by MEXT/JSPS Grant-in-Aid for Scientific Research(S) 24220004.

References

[1] K. Amriki and P. Atrey. Towards optimal placement of surveillance cameras in a bus. In *ICME*, pages 1–6, 2011.

[2] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. SP*, 50(2):174–188, Feb 2002.

[3] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, pages 435–440, Aug 2010.

[4] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz. Recent progress in road and lane detection: a survey. *MVA*, 25(3):727–745, 2014.

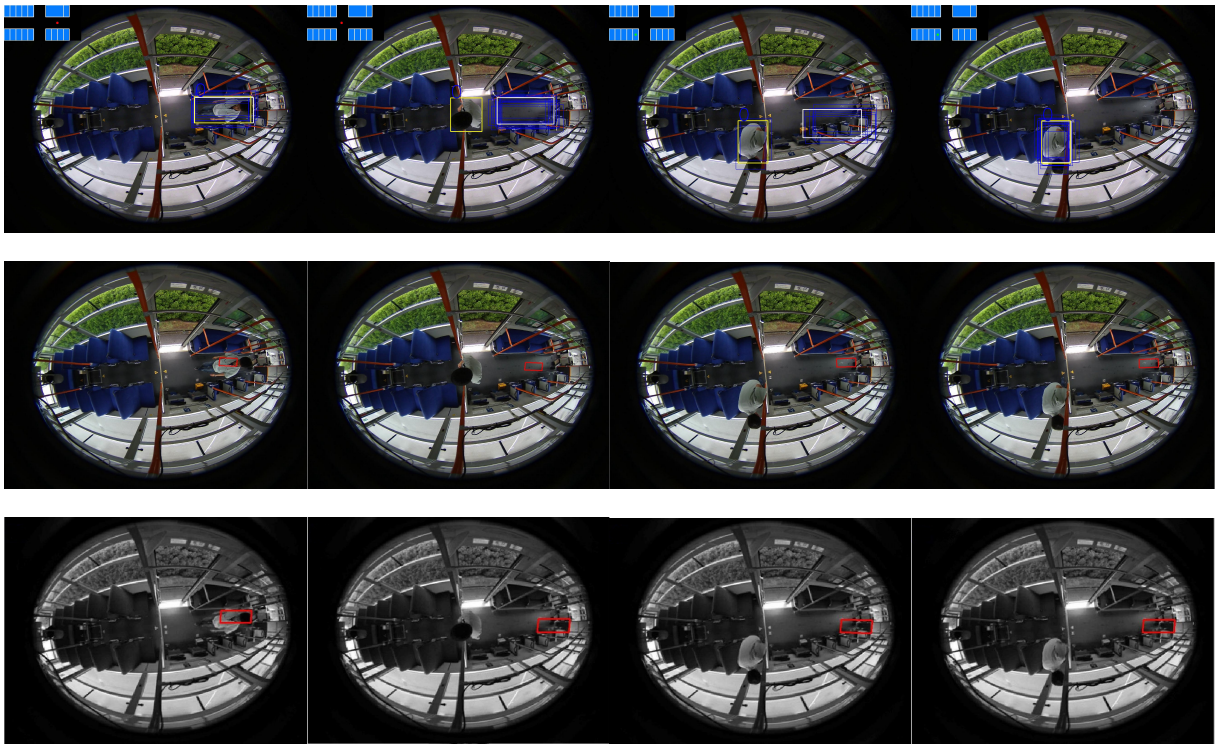
[5] N. Buch, S. Velastin, and J. Orwell. A review of computer vision techniques for the analysis of urban traffic. *IEEE Trans. on ITS*, 12(3):920–939, Sept 2011.

[6] B. C. Chee, M. Lazarescu, and T. Tan. Detection and monitoring of passengers on a bus by video surveillance. In *ICIAP*, pages 143–148, Sept 2007.

[7] K. Chen, S. Gong, T. Xiang, and C. Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, pages 2467–2474, 2013.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[9] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Trans. on ITS*, 12(2):596–614, June 2011.



(i) Frame 46 (j) Frame 47 (k) Frame 57 (l) Frame 58

Figure 3: Visual results of our hybrid particle filter tracking method (first row), SVD (second row) and DLT (last row). Missing-frame phenomenon happens between frames 46 and 47. Both SVD and DLT are failture to track the person. Howver, our hybrid particle filter tracking can track the person due to the proposed resampling mechanism.

[10] T. Ehlgen, T. Pajdla, and D. Ammon. Eliminating blind spots for assisted driving. *IEEE Trans. on ITS*, 9(4):657–665, Dec 2008.

[11] D. Geronimo, A. Lopez, A. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on PAMI*, 32(7):1239–1258, July 2010.

[12] J. Greenhalgh and M. Mirmehdi. Recognizing text-based traffic signs. *IEEE Trans. on ITS*, PP(99):1–10, 2015.

[13] D. H. Hung, S. L. Chung, and G. S. Hsu. Local empirical templates and density ratios for people counting. In *ACCV*, pages 90 – 101, 2010.

[14] D. H. Hung and H. Saito. The estimation of heights and occupied areas of humans from two orthogonal views for fall detection. *IEEJ Transactions on EIS*, 133(1):117 – 127, 2013.

[15] D. H. Hung, H. Saito, and G. S. Hsu. Detecting fall incidents of the elderly based on human-ground contact areas. In *ACPR*, pages 516–521, 2013.

[16] J. Ma, W. Liu, and P. Miller. Handling sequential observations in intelligent surveillance. In *Scalable Uncertainty Management*, pages 547–560. Springer, 2011.

[17] J. Ma, W. Liu, P. Miller, and W. Yan. Event composition with imperfect information for bus surveillance. In *AVSS*, pages 382–387, Sept 2009.

[18] P. Miller, W. Liu, C. Fowler, H. Zhou, J. Shen, J. Ma, J. Zhang, W. Yan, K. McLaughlin, and S. Sezer. Intelligent sensor information system for public transport–to safely go. In *AVSS*, pages 533–538. IEEE, 2010.

[19] K. Nummiaro, E. Koller-Meier, and L. Van Gool. An adaptive color-based particle filter. *IVC*, 21(1):99–110, 2003.

[20] H. Salmane, L. Khoudour, and Y. Ruichek. A video-analysis-based railway-road safety system for detecting hazard situations at level crossings. *IEEE Trans. on ITS*, PP(99):1–14, 2015.

[21] Z. Sun, G. Bebis, and R. Miller. On-road vehicle detection: a review. *IEEE Trans. on PAMI*, 28(5):694–711, May 2006.

[22] N. Wang and D.-Y. Yeung. Learning a deep compact image representation for visual tracking. In *NIPS*, pages 809–817, 2013.

[23] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418. IEEE, 2013.

[24] F. Yang, H. Lu, and M.-H. Yang. Learning structured visual dictionary for object tracking. *IVC*, 31(12):992–999, 2013.

[25] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *MVA*, 19(5-6):345–357, 2008.

[26] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, pages 28–31, 2004.