Driving Recorder Based On-Road Pedestrian Tracking Using Visual SLAM and Constrained Multiple-Kernel

Kuan-Hui Lee, Jenq-Neng Hwang
Department of Electrical Engineering,
Box 352500, University of Washington,
Seattle, WA, 98195, USA
{ykhlee, hwang}@uw.edu

Greg Okapal, James Pitton
Applied Physics Laboratory
Box 355640, University of Washington,
Seattle, WA, 98195, USA
{Okopal, pitton}@apl.washington.edu

***Abstract*— This paper proposes a robust driving recorder based on-road pedestrian tracking system, which effectively integrates Visual Simultaneous Localization And Mapping (V-SLAM), pedestrian detection, ground plane estimation, and kernel-based tracking techniques. The proposed system systematically detects the pedestrians from recorded video frames and tracks the pedestrians in the V-SLAM inferred 3-D space via a tracking-by-detection scheme. In order to efficiently associate the detected pedestrian frame-by-frame, we propose a novel tracking framework, combining the Constrained Multiple-Kernel (CMK) tracking and the estimated 3-D (depth) information, to globally optimize the data association between consecutive frames. By taking advantage of the appearance model and 3-D information, the proposed system not only achieves high effectiveness but also well handles occlusion in the tracking. Experimental results show the favorable performance of the proposed system which efficiently tracks on-road pedestrian in a moving camera equipped on a driving vehicle.**

Keywords— mobile vision; pedestrian detection; object tracking; Visual SLAM;

I. INTRODUCTION

Nowadays, an emerging application of video analysis for Intelligent Transportation Systems (ITS) is the use of the driving recorder, which is a device that records video for a vehicle to create a record of driving. A driving recorder, which gradually becomes a necessity in a vehicle, can be regarded as a new mechanism of moving video surveillance on the roads. Tracking of pedestrians, especially tracking under moving cameras, is a quite challenging task because of the combined effects of egomotion, blur, and rapidly changing lighting conditions. The introduction of a moving camera invalidates many effective moving object tracking techniques used in static camera, such as background subtraction and a constant ground plane assumption. Therefore, the challenge is to successfully detect the pedestrians in moving cameras, and then apply the tracking techniques to the detected ones, resulting in the so-called tracking-by-detection schemes. However, when pedestrians are partially/fully occluded, the detections can fail and the tracking can be unreliable until the pedestrains reappear in the frames. Hence, to handle the occlusion issues during tracking, the 3-D information obtained by using multi-view stereo techniques, such as visual odometry and Visual Simultaneous Localization And Mapping (V-SLAM), can be further adopted to infer the relative 3-D locations between the targets [1].

In this paper, taking advantage of the tracking-by-detection scheme, we propose a robust driving recorder based on-road pedestrian tracking system, which successfully integrates V-SLAM, pedestrian detection, ground plane estimation, and kernel-based tracking technique. The proposed system starts with pedestrian detection and V-SLAM estimation of the 3-D locations of the pedestrians. By taking 3-D information into account, we reformulate the tracking problem based on the Constrained Multiple-Kernel (CMK) approach [2], which can effectively resolve the occlusions during tracking, to globally optimize the data association between consecutive frames. Hence, the proposed system can not only track the pedestrians effectively, but can also robustly handle occlusion during tracking.

The rest of the paper is organized as follows. Section II gives a brief survey on the related work. In Section III, the overview of the proposed system, including adopted algorithms, is briefly described. Section IV depicts the proposed pedestrian tracking, which combines the CMK tracking with the 3-D information to formulate the association of the detected pedestrians. The experimental results are shown in Section V, followed by the conclusion in Section VI.

II. RELATED WORK

In general, pedestrian detection by a moving camera follows two basic steps [1]: foreground segmentation and object classification. Foreground segmentation first extracts blobs of interest from the image, avoiding as many background regions as possible. Then, object classification classifies the extracted blobs as pedestrians or non-pedestrians. The approaches of the foreground segmentation can be classified into 1) image-based and 2) motion-based. The image-based approaches mainly rely on the color, intensity, edges, and gradient orientation of pixels [3],[4]. The motion-based approaches utilize inter-frame motion and optical flow [5],[6], so as to avoid as many background regions as possible, and extract candidates by considering the aspect ratio, size, and position. The approaches to object classification are purely based on 2-D information, and can also be broadly divided into 1) template-based and 2) appearance-based. The template-based approaches use predefined patterns of the human/vehicle

classes and perform correlation between the image and the template. The template is a human body in case of the pedestrian detection [7]–[9]. The appearance methods [10]–[14] define a set of image features (descriptors), and a classifier is pre-trained by positive examples (human and/or vehicle) and negative examples (nonhuman and/or nonvehicle) with various learning algorithms, such as neural network, SVM, AdaBoost and etc.

After object detection, a tracking framework is applied to the detected objects. There have been literatures of pedestrian and vehicle tracking with moving cameras, e.g., Kalman filters [9], [15] and Particle filters [16]–[18] are widely used in tracking. Ess *et al*. [19] perform multi-body tracking by combining an ISM detector [12] and a stereo-odometry-based tracker. Adnriluka *et al*. [20] detect people using a part-based detector [13], and then use a Gaussian process latent variable model to compute the temporal consistency of detections over time. Leibe *et al*. [21],[22] propose the use of a color model and the event cone, i.e., the time-space volume in which the trajectory of a tracked object is sought in 3-D space. Recently, many approaches [23]–[26] formulate the tracking problem as a min-cost flow network problem. These approaches globally optimize the trajectories of all objects, instead of locally optimizing for each object. However, the performance highly relies on the reliable detection. If the detection misses or long-time occlusion happens, the performance deteriorates significantly.

Alternatively, several approaches based on the structure-from-motion (SfM) framework have been developed. The SfM framework is originally used for static 3-D scene reconstruction in multi-view stereo applications. It is also applied to the moving cameras, combined with V-SLAM, to calibrate and to localize the 3-D positions of the camera and static features in the world coordinate system [27],[28]. Based on the SfM, many researchers develop approaches to detect, track, and reconstruct moving objects within the static background, the so-called dynamic scene reconstruction. In [21] and [22], the authors reconstruct not only the static background but also humans and vehicles, and track them by the tracking-by-detection scheme. Kundu *et al*. [29] presents a real-time and incremental V-SLAM system that allows choosing between full 3-D reconstruction or simply tracking of the moving objects. The advantage of the SfM based approaches is to locate the objects in 3-D space, so as to deal with the occlusion problem during tracking. Owing to this reason, we also choose to reconstruct the 3-D locations of the objects based upon the SfM framework.

III. Overview of the Proposed System

Fig. 1 shows the overview of the proposed system. First, the proposed system calibrates the camera motions by the classic SfM pipeline. Meanwhile, a pre-trained human detector is adopted to detect pedestrians in the video frames from a driving recorder. Then, according to the calibrated camera motions, we can thus estimate the ground plane, which is used by the pose estimation step to back-project the pedestrians' 2-D locations to 3-D locations. Then, a depth map is constructed to represent the detected pedestrians. Finally, the proposed tracking technique combining the CMK tracking [2] and the depth information is applied to track the detected pedestrians.

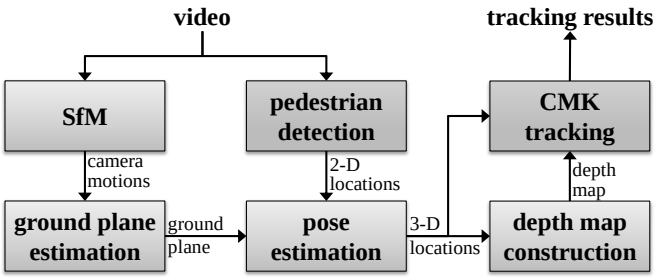


Fig. 1. Overview of the proposed system.

A. Structure-from-Motion (SfM)

The monocular V-SLAM framework follows a standard bundle adjustment formulation, where an interest point operator is first applied to extract feature points which are matched between consecutive frames. Then, according to the matched feature points, a 5-point based RANSAC algorithm is used to estimate the initial epipolar geometry. Finally, the camera motions are determined by camera resection. The set of 3-D points and the corresponding feature points are used in the bundle adjustment process to iteratively minimize the reprojection error:

$$X_p = \arg\min_{X,\hat{\mathbf{P}}} \sum_{\forall p,q} d(\hat{\mathbf{P}}_q \cdot X_{pq}, x_{pq}), \quad (1)$$

where X_p is the 3-D location of the p^{th} feature point, x_{pq} is the observed 2-D location corresponding to X_p from the q^{th} video frame (the original formulation was referred to the q^{th} camera), $\hat{\mathbf{P}}_q$ is the projective matrix of the q^{th} frame, $\hat{\mathbf{P}}_q \cdot X_{pq}$ is the reprojection of X_p onto the q^{th} frame, and $d(\bullet)$ is the distance measurement between the reprojected locations and the observed locations in the image. Such nonlinear least-square problem is solved by the Levenberg-Marquardt algorithm.

Unlike the classic multi-view stereo algorithm, where the bundle adjustment is used to globally optimize camera motions based on all the images, V-SLAM is used to locally optimize the camera motions based on f_s consecutive frames. In other words, a windowed bundle adjustment is applied to obtain more robust and more accurate camera motions along the time.

B. Pedestrian Detection

For pedestrian detection, we use a pre-trained human detector [11]–[14] to detect pedestrians in the video frames. The human detector proposed in [11] considers *Histogram Of Gradient* (HOG) as the features, which can efficiently represent the shape of human. The *Implicit Shape Model* (ISM) [12] uses a voting scheme based on multi-scale interest points to generate a large number of detections hypotheses, and a codebook is used to preserve the trained features. The *Deformable Part Model* (DPM) [13] extends the idea of [11], using a root model and several part models to describe different partitions of an object. The part models are spatially connected with the root model according to the predefined geometry, so as to precisely depict the object. Recently, C^4 [14] adopts a local-binary-pattern-like feature to efficiently

represent the human template and achieve real-time performance. These human detectors can be independently embedded in the proposed system, so as to functionally perform pedestrian detection. To efficiently lockdown the targets, we start to track a target which has to be detected in three consecutive frames; otherwise, the detections are regarded as false alarm. Furthermore, the detections are refined by morphological operations to accurately determine their locations.

C. Ground Plane Estimation

Due to unpredictability of road conditions, a ground plane estimated in the beginning may not be applicable for the entire video sequence. Therefore, the ground plane needs to be continuously reestimated based on the updated camera motions from the SfM. Since the noises produced by camera calibration usually have adverse impact on ground plane estimation, we collect f_s ground planes, each is calculated by a pair of consecutive camera motions, to form a set of ground planes $\{(\mathbf{g}_q, \psi_q)\}$, where $\mathbf{g}_q$ is the normal vector of the ground plane and ψ_q is the offset of the plane. Then, we can combine them into a single 4-by-f_s matrix $\mathbf{D}$,

$$\mathbf{D} = [(\mathbf{g}_q, \psi_q) \quad \cdots \quad (\mathbf{g}_{q+f_s}, \psi_{q+f_s})]. \tag{2}$$

Note that some $\{(\mathbf{g}_q, \psi_q)\}$ are unreliable due to varying road conditions and noisy camera calibration.

To recover the uncorrupted version from the corrupted matrix $\mathbf{D}$, we employ the Robust Principle Component Analysis (RPCA) [30] to extract a low-rank 4-by-f_s matrix $\mathbf{A}$ from $\mathbf{D}$. Thanks to the characteristic of RPCA, the low rank matrix $\mathbf{A}$ consists of the uncorrupted data which represents the ground planes for the application. The mean vector of the matrix $\mathbf{A}$ is considered to be our final ground plane for those f_s consecutive frames and is more resilient to the existence of noise in the system.

D. Constrained Multiple-Kernel (CMK) Tracking

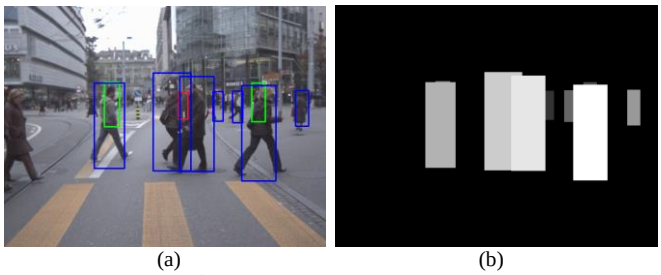
The objective of the CMK tracking [2] is to retrieve a candidate object, which can be described as multiple kernels with pre-specified constraints among these kernels, so that the maximum similarity between the tracked object and the candidate model can be reached. For N_k kernels, the total cost function $J(\mathbf{x})$ is defined to be the sum of the N_k individual cost functions $J_\kappa(\mathbf{x})$, which is designed to be inversely proportional to the similarity. Moreover, each individual cost is also assigned by a weight w_κ which is proportional to the similarity:

$$J(\mathbf{x}) = \sum_{\kappa=1}^{N_k} w_\kappa \cdot J_\kappa(\mathbf{x}), \tag{3}$$

where $simi_\kappa(\mathbf{x})$ is the similarity function at the location $\mathbf{x}$ in the state space domain.

In addition, the constraint functions $\mathbf{C}(\mathbf{x}) = \mathbf{0}$ need to be considered to maintain the relative locations of the kernels. The constraint functions confine the kernels based on their spatial inter-relationships. Thus, the problem could be further formulated as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} J(\mathbf{x}), \text{ subject to } \mathbf{C}(\mathbf{x}) = \mathbf{0}. \tag{4}$$



(a) (b)

Fig. 2. An example of the depth map, showing (a) detections and (b) depth map, where higher intensity indicates detected pedestrians are closer to the camera.

By well defining the similarity function and the constraint functions, all kernels can not only be directed to the most similar region but also maintain the constrained conditions.

IV. CMK Based Pedestrian Tracking

Once obtaining the 3-D locations of the pedestrians from the pose estimation stage (see Fig. 1), we apply the CMK tracking technique to track them. In other words, we associate the targets in the current frame with the detections in the next frame. First, for each target, the 3-D location of its candidate is predicted by Kalman filtering. Then, the CMK tracking is used to effectively relocate the candidate's 3-D location by achieving best color similarity. On the other hand, the depth information helps to understand the relative 3-D locations between the targets, so that we are able to handle occlusion issues in the tracking. By efficiently combining depth information into the CMK tracking, the proposed system not only effectively tracks the pedestrians, but also well handles occlusions.

A. Depth Map Construction

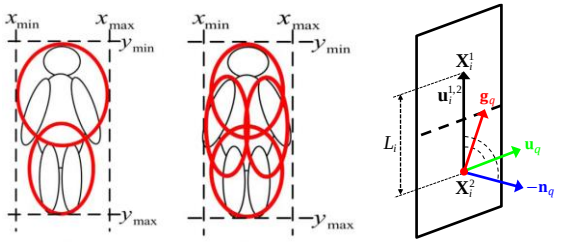
Based on the 3-D locations of the pedestrians, we can construct a depth map to describe the relative 3-D locations of all the tracked targets. Fig. 2 shows an example of the depth map, where Fig. 2 (a) shows the result of pedestrian detection and Fig. 2 (b) is the corresponding depth map. The depth map depicts the relative distance between the camera and the detected pedestrians, higher intensity means the detected pedestrians are closer to the camera. As shown in the Fig. 2 (a), two green-boxed targets are almost occluded by the other targets in front of them, while a red-boxed target is totally occluded by other targets. Thanks to the depth map, we can approximately evaluate if the i^{th} target is occluded or not, in terms of the visibility $v_i \in [0,1]$:

$$v_i = \frac{\text{visible area of } i^{th} \text{ target}}{\text{total area of } i^{th} \text{ target}}. \tag{5}$$

If $v_i = 1$, it implies the i^{th} target is totally visible without being occluded by other targets; if $0 < v_i < 1$, it means the i^{th} target is partially occluded; otherwise, it is totally occluded.

B. Problem Formulation

In [2], the CMK tracking scheme tracks video objects in 2-D space (image), i.e., $\mathbf{x} \in \mathbb{R}^2$ in Eq. (3). To efficiently integrate the depth information into the CMK framework, we need to



(a) (b) (c)

Fig. 3. Layout for (a) 2-kernel and (b) 4-kernel, both in 2-D space. (c) Illustration of the 3-D based constraints in case of 2-kernel layout.

(a) (b)

Fig. 4. An example of the hypothesized association in (a) frame 230 and (b) frame 233; the blue boxes represent the detections, and the red box is the inserted hypothesized association.

reformulate the problem. First we extend the Eq. (3) from 2-D to 3-D space,

$$J^i(\mathbf{X}) = \sum_{\kappa=1}^{N_k} w_\kappa \cdot J_\kappa^i(\mathbf{X}), \quad \mathbf{X} \in \mathbb{R}^3. \tag{6}$$

This equation is regarded as the local optimization for each individual target *i* with multiple kernels. Second, considering the depth information, we assign the visibility of each target as a weight to deal with the global optimization. In other words, the total cost function becomes:

$$J(\mathbf{X}) = \sum_{i=1}^{N_q} v_i \cdot J^i(\mathbf{X}) = \sum_{i=1}^{N_q} v_i \cdot \left(\sum_{\kappa=1}^{N_k} w_\kappa \cdot J_\kappa^i(\mathbf{X}) \right), \tag{7}$$

where N_q is the number of the targets in the q^{th} video frame, and $\mathbf{X} \in \mathbb{R}^{3\times N_q}$.

Necessarily, the constraint functions $\mathbf{C}(\mathbf{X}) = \mathbf{0}$ must be considered to maintain the relative locations of the kernels. In [2], 2-kernel and 4-kernel layouts are proposed to describe a human, as shown in Fig. 3 (a) (b). Unlike the constraints used in [2], which are mainly based on 2-D geometry, we set the constraints based on 3-D geometry. Without loss of generality, here we only discuss the 2-kernel case as shown in Fig. 3 (c), but the idea can be easily extended to the 4-kernel case. To represent a target in 3-D space, we define a *target plane* $(-\mathbf{n}_q, \pi_q^i)$ for each i^{th} target in the q^{th} frame; where $\mathbf{n}_q$ is the normal vector of the q^{th} frame, and π_q^i is the offset of the target plane. To properly set the constraints, we start to calculate two auxiliary vectors, $\mathbf{u}_q = -\mathbf{n}_q \times \mathbf{g}_q$ and $\mathbf{u}_i^{1,2} = \mathbf{X}_i^1 - \mathbf{X}_i^2$. First, the distance between two kernel centers should be the same, which implies

$$\left\| \mathbf{u}_i^{1,2} \right\|^2 = (L_i)^2, \text{ for } i^{\text{th}} \text{ target}, \tag{8}$$

where L_i is the initial distance between $\mathbf{X}_i^1$ and $\mathbf{X}_i^2$. Second, both the angle between $\mathbf{u}_q$ and $\mathbf{u}_i^{1,2}$, and the angle between $-\mathbf{n}_q$ and $\mathbf{u}_i^{1,2}$, should be consistent. Therefore, we can have

$$\begin{cases} \dfrac{\mathbf{u}_q \cdot \mathbf{u}_i^{1,2}}{\| \mathbf{u}_q \| \| \mathbf{u}_i^{1,2} \|} = \cos(\phi_{q,i}) \\ \dfrac{-\mathbf{n}_q \cdot \mathbf{u}_i^{1,2}}{\| -\mathbf{n}_q \| \| \mathbf{u}_i^{1,2} \|} = \cos(\varsigma_{q,i}) \end{cases}, \text{ for the } i^{\text{th}} \text{ target in the } q^{\text{th}} \text{ frame.} \tag{9}$$

In order to gradually decrease the total cost function and ensure the constraints satisfied during the state search, the movement vector $\delta_{\mathbf{X}}$, i.e., the gradient vector of the $J(\mathbf{X})$, is needed for using the projected gradient method to iteratively solve the constrained optimization problem [31]. The basic idea is to project the gradient vector onto two orthogonal spaces, one is related to decreasing the cost function and the other corresponds to satisfying the constraints, i.e., $\mathbf{C}(\mathbf{X}) = \mathbf{0}$. Hence, for each target *i*, we can have the movement vector:

$$\delta_{\mathbf{X}}^i = v_i \left(\alpha(-\mathbf{I} + \mathbf{C}_\mathbf{X}(\mathbf{C}_\mathbf{X}^T\mathbf{C}_\mathbf{X})^{-1}\mathbf{C}_\mathbf{X}^T)\mathbf{W}\mathbf{J}_\mathbf{X} + (-\mathbf{C}_\mathbf{X}(\mathbf{C}_\mathbf{X}^T\mathbf{C}_\mathbf{X})^{-1}\mathbf{C}(\mathbf{X})) \right), \tag{10}$$

where $\mathbf{X} \in \mathbb{R}^3$ is the state vector; $\mathbf{C}(\mathbf{X}) = \begin{bmatrix} c_1(\mathbf{X}) & \cdots & c_m(\mathbf{X}) \end{bmatrix}^T$ is the matrix including *m* constraint functions, and $c_j(\mathbf{X}) : \mathbb{R}^3 \to \mathbb{R}$ is the j^{th} constraint function; $\mathbf{C}_\mathbf{X} \in \mathbb{R}^{n\times m}$ is the gradient matrix of constraint functions with respect to $\mathbf{X}$; $\mathbf{J}_\mathbf{X}$ is the gradient vector of the total cost function with respect to $\mathbf{X}$; $\alpha > 0$ is the step size; $\mathbf{W} = \begin{bmatrix} w_1\mathbf{I} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{N_k}\mathbf{I} \end{bmatrix}$, $\mathbf{I}$ is an $\frac{n}{N_k} \times \frac{n}{N_k}$ identity matrix, and *n* is the dimension of the state space.

C. Hypothesized Association

However, due to unreliable detection or occlusion, pedestrians may not be detected for several frames. Hence, some tracked targets cannot be successfully associated with the detections in the subsequent frame. To consistently track a non-associated target, we insert a *hypothesized association* which has been located by the CMK tracking with the best color similarity. Fig. 4 shows an example of the hypothesized association in case of occlusion. The person wearing gray clothes is detected in frame 230, but is then occluded by another person 3 frames later. By 3-D information, we can predict his 3-D location and understand that he is occluded. Hence, a hypothesized association is used here to pretend a possible detection, as shown by the red box in Fig. 4 (b).

V. Experimental Results

In this section, we show experimental results of the proposed system on the ETH Mobile Scene (ETHMS) dataset [19], which is very challenging and difficult because of the heavy occlusions and busy crowd in the camera views. Since the proposed system is developed for the application of monocular camera, we only use the left view sequences of the dataset. In our experiments, the proposed system is evaluated by its detection performance and tracking performance. Furthermore, we compare our results with three state-of-the-art methods [19],[23],[24]. In order to have a fair comparison of our proposed scheme with previous work, we use the same

TABLE I
COMPARISON OF DETECTION RATE AND FPPI

Method	Detector	Detection rate (%)	False Positive Per Image (FPPI)
[19]	ISM	47	1.5
[19]	HOG	67.5	1.0
[24]	DPM	49.53	0.93
[24] + NMS	DPM	49.94	0.93
C4K	HOG	53.28	1.32
C4K	DPM	59.74	1.24
C2K	HOG	70.61	0.97
C2K	C^4	62.36	1.14
C2K	DPM	75.58	0.89

TABLE II
CONVERGENCE COMPARISON

Dataset	Method / Detector	GT	MT	PT	ML	FM	IDS
Seq#1	[19] / ISM	73	**66**	**5**	2	**8**	1
	[24] / DPM	73	54	13	6	19	8
	[24] + NMS / DPM	73	55	12	6	19	8
	C4K / DPM	73	58	10	5	7	2
	C2K / DPM	73	**64**	**7**	2	**3**	3
Seq#4	[19] / ISM	88	**74**	**8**	6	**20**	3
	[24] / DPM	88	52	16	20	29	14
	[24] + NMS / DPM	88	55	14	19	29	14
	C4K / DPM	88	64	14	10	18	6
	C2K / DPM	88	**71**	**9**	8	**11**	6

video sequences as used in their simulations. All the experiments are processed on a personal computer with a P4 2.67GHz CPU and 2G DDR. The implementation is constructed by C/C++, and the experimental settings are described as follows. In the SfM framework, the proposed system adopts Harris corners as the features, which are tracked by a KLT tracker. In the pedestrian detection, the pre-trained detectors [11]–[14] are functionally applied to the proposed system to detect pedestrians within the 30 meters. In the CMK tracking, K-L distance is used for all similarity measures, and the 8-bins histogram of the object is constructed based on the HSV color space with a roof kernel. To evaluate the performance, we compare the proposed approach with the other approaches. The approach in [19] is a stereo algorithm based on graphical model. The approach in [24] is a dynamic programming algorithm based on flow network framework, with and without the Non-Maxima Suppression (NMS) in it. The C2K is the proposed approach with 2 kernels and C4K is for 4 kernels.

A. *Detection Performation*

The approaches are evaluated with different human detectors, in terms of the detection rate and false positive per image (FPPI), as shown in Table I. As we can see, the approaches with DPM have better performance than that with other detectors. The results show that both the proposed approach (C2K, C4K) and the approach in [19] are superior to the approach in [24]. Since both approaches further utilize the 3-D information, instead of 2-D information only in [24], they can have effectively handle occlusion issues. When compared the C2K with the C4K, the C2K performs much better than the C4K because the C2K has better performance in the tracking, which results in increasing the detection rate and decreasing the FPPI. The detection rate of the proposed approach can achieve about 75%, owing to proper insertion of hypothesized associations and successive tracking.

B. *Tracking Performation*

To evaluate the tracking performance, we consider the following metrics which are widely used in the previous work [19],[23]:

- most tracked trajectories (MT): the number of trajectories that successfully tracked more than 80% frames in a video sequence..
- partially tracked trajectories (PT): the number of trajectories that successfully tracked between 20% and 80%.
- most lost trajectories (ML): the number of trajectories that successfully tracked less than 20%.
- Fragmentation (FM): the number of times a trajec-tory is interrupted.
- ID switches (IDS): the number of times two trajectories switch their IDs.

The results of two sequences in dataset are shown in Table II, where GT denotes the number of trajectories in the ground truth. From MT results, the approach in [24] is not as good as other approaches due to the limitation of 2-D information. The 3-D based approaches are able to significantly improve the performance by efficiently handling occlusions, thus result in lower MT and higher FM. The C2K performs better than the C4K, because the detected pedestrians are too small to be clearly described by 4 kernels. This may additionally include some background regions, so that the tracking is easily impacted by the background. When compared the C2K with the approach in [19], although the MT/PT/ML results are comparable, the FM results of the CMK are much better than that of [19]. This implies that the proposed 3-D based CMK framework efficiently associate the targets, so as to perform well on successively tracking the targets. Several visual tracking results are shown in Fig. 5, where (a), (b) are the ETHMS dataset, and (c), (d) are the datasets recorded by ourselves. The results show favorable performance of the proposed system, not only successively tracking pedestrians but also well handling occlusion in the tracking. Since relative 3-D locations of the pedestrians are obtained, with GPS information, we can also construct 3-D visualization. Fig. 5 (e) shows the 3-D visualization of the dataset in Fig. 5 (d), showing the dynamic scenes in different aspects of view. All the videos associated with the simulations reported in this paper can be viewed our website[1].

VI. CONCLUSION

We proposed a robust on-road pedestrian tracking system in a moving camera. The proposed system effectively integrates the human detectors and V-SLAM framework to relocate the pedestrians in 3-D space, followed by an innovative 3-D based CMK tracking, which not only locally

[1] website: http://allison.ee.washington.edu/kuanhuilee/mcpt

associates the targets but also globally optimizes the associations according to the 3-D information. This proposed framework can also be further applied to other class of on-road objects if the corresponding detectors are available.

REFERENCES

[1] D. Gerónimo, A. M. López, A. D. Sappa and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[2] C.-T. Chu, J.-N. Hwang, H.-I Pai, and K.-M. Lan, "Tracking Human under Occlusion Based on Adaptive Multiple Kernels with Projected Gradients," *IEEE Trans. Multimedia*, vol.5, no.7, pp. 1602–1615, Nov. 2013.

[3] T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka, "Development of Night-Vision System," *IEEE Trans. Intelligent Transportation Systems*, vol. 3, no. 3, pp. 203–209, Sept. 2002.

[4] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli, and A. Tibaldi, "Shape-Based Pedestrian Detection and Localization," *Proc. IEEE Int'l Conf. Intelligent Transportation Systems*, pp. 328–333, 2003.

[5] H. Elzein, S. Lakshmanan, and P. Watta, "A Motion and Shape-Based Pedestrian Detection Algorithm," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 500–504, 2003.

[6] U. Franke and S. Heinrich, "Fast Obstacle Detection for Urban Traffic Situations," *IEEE Trans. Intelligent Transportation Systems*, vol. 3, no. 3, pp. 173–181, Sept. 2002.

[7] H. Nanda and L. Davis, "Probabilistic Template Based Pedestrian Detection in Infrared Videos," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 15–20, 2002.

[8] D. Gavrila, J. Giebel, and S. Munder, "Vision-Based Pedestrian Detection: The PROTECTOR System," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 13–18, 2004.

[9] D. Gavrila and S. Munder, "Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle," *Int'l J. Computer Vision*, vol. 73, no. 1, pp. 41–59, 2007.

[10] P. Viola, M. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *Proc. Int'l Conf. Computer Vision*, vol. 2, pp. 734–741, 2003.

[11] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[12] B. Leibe, A. Leonardis, and B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *Int'l J. Computer Vision*, vol. 77, no. 1–3, pp. 259–289, May 2008.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no.9, pp. 1627–1645, Sep. 2010.

[14] J. Wu, N. Liu, C. Geyer, and J. M. Rehg, "C4: A Real-Time Object Detection Framework," *IEEE Trans. Image Processing*, vol. 22, no.10 , pp. 4096–4106, Oct. 2013

[15] M. Bertozzi, A. Broggi, A. Fascioli, A. Tibaldi, R. Chapuis, and F. Chausse, "Pedestrian Localization and Tracking System with Kalman Filtering," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 584–589, 2004.

[16] V. Philomin, R. Duraiswami, and L. Davis, "Pedestrian Tracking from a Moving Vehicle," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 350–355, 2000.

[17] J. Giebel, D. Gavrila, and C. Schnör, "A Bayesian Framework for Multi-Cue 3D Object Tracking," *Proc. European Conf. Computer Vision*, pp. 241-252, 2004.

[18] R. Arndt, R. Schweiger, W. Ritter, D. Paulus, and O. Lóhlein, "Detection and Tracking of Multiple Pedestrians in Automotive Applications," *Proc. IEEE Intelligent Vehicles Symp.*, pp. 13–18, 2007.

[19] A. Ess, B. Leibe, K. Schindler, and L. VanGool, "Robust Multiperson Tracking from a Mobile Platform," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1831–1846, Oct. 2009.

[20] M. Adnriluka, S. Roth, and B. Schiele, "People-Tracking-by-Detection and People-Detection-by-Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2008.

[21] B. Leibe, N. Cornelis, K. Cornelis, and L. VanGool, "Dynamic 3D Scene Analysis from a Moving Vehicle," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2007.

[22] B. Leibe, K. Schindler, N. Cornelis, and L. VanGool, "Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1683–1698, Oct. 2008.

[23] L. Zhang, Y. Li, and R. Nevatia, "Global Data Association for Multi-Object Tracking Using Network Flows," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2008.

[24] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011.

[25] Z. Wu, A. Thangali, S. Sclaroff and M. Betke, "Coupling Detection and Data Association for Multiple Object Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.1948–1955, Jun. 2012.

[26] A. A. Butt and R. T. Collins, "Multi-target Tracking by Lagrangian Relaxation to Min-Cost Network Flow," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2013.

[27] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real Time Localization and 3D Reconstruction" *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 363–370, 2006.

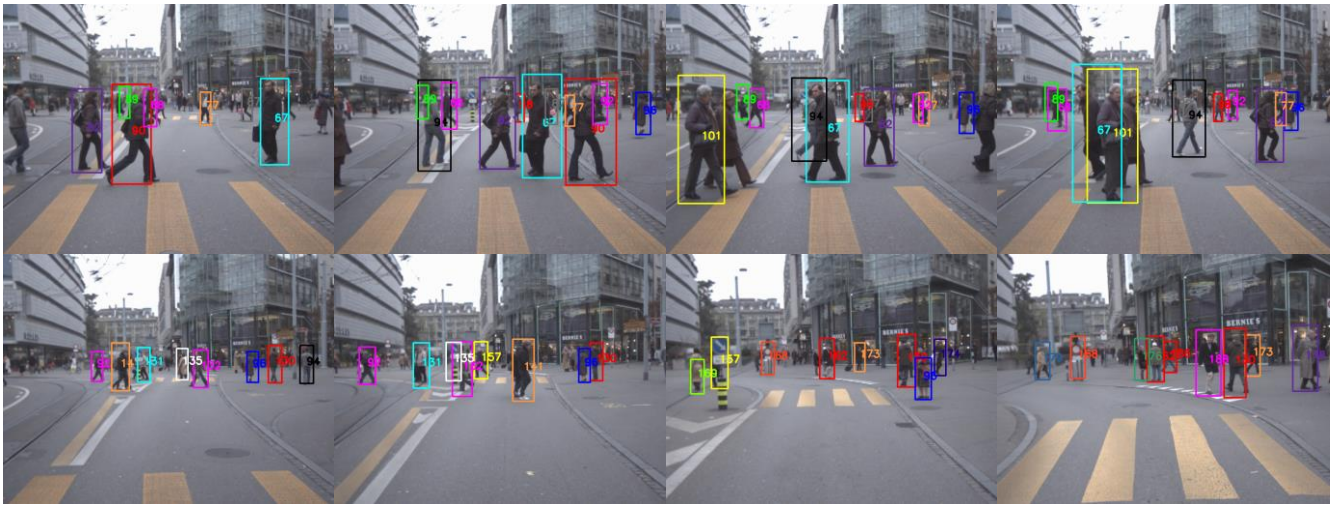
[28] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G.Welch and H. Towles, "Detailed Real-Time Urban 3D Reconstruction from Video," *Int'l J. Computer Vision*, vol. 78, no. 2–3, pp. 143–167, Jul. 2008.

[29] A. Kundu, K. M. Krishna and C. V. Jawahar, "Realtime Multibody Visual SLAM with a Smoothly Moving Monocular Camera," *Proc. Int'l Conf. Computer Vision*, pp. 2080–2087, Nov. 2011.

[30] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao, "Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization," *Proc. Neural Information Processing Systems*, Dec. 2009.

[31] J. A. Snyman, *Practical Mathematical Optimization*. NewYork, NY, USA: Springer Science + Business Media, 2005, ch. 3.

(a)



(b)

(c)

(d)

(e)

Fig. 5. Visual tracking results. (a) ETHMS Seq#1, (b) ETHMS Seq#4, (c) Downtown Seq#1, (d) Downtown Seq#2, and (e) 3-D visualization of the Downtown Seq#2, in different aspects of views.