Ground-Moving-Platform-Based Human Tracking Using Visual SLAM and Constrained Multiple Kernels

Kuan-Hui Lee, Jenq-Neng Hwang, *Fellow, IEEE*, Greg Okopal, and James Pitton

***Abstract*—This paper proposes a robust ground-moving-platform-based human tracking system, which effectively integrates visual simultaneous localization and mapping (V-SLAM), human detection, ground plane estimation, and kernel-based tracking techniques. The proposed system systematically detects humans from recorded video frames of a moving camera and tracks the humans in the V-SLAM-inferred 3-D space via a tracking-by-detection scheme. To efficiently associate the detected human frame by frame, we propose a novel human tracking framework, combining the constrained-multiple-kernel tracking and the estimated 3-D information (depth), 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 humans in a camera equipped on a ground-moving platform such as a dash camera and an unmanned ground vehicle.**

***Index Terms*—Mobile vision, human detection, human tracking, visual SLAM.**

I. Introduction

RECENTLY, an emerging application of video analysis for Intelligent Transportation Systems (ITS) is the use of the mobile vision, since autonomous machines such as vehicles, robots, and drones, are newly developed for different applications [1]–[3]. As more and more development of the autonomous technologies, many applications of video analyses are considered to be applied to autonomous platforms. Within the applications, human tracking is one of the most critical tasks in mobile vision for the development of intelligent autonomous systems. By human tracking, humans' moving trajectories can be obtained to process high level analytics and applications, for example, human counting, human flow estimation, criminal tracking, detection and avoidance of abnormal behaviors, and so on. Therefore, human tracking on a moving platform is well developed [24], [26].

Manuscript received April 1, 2015; revised October 29, 2015 and February 1, 2016; accepted April 16, 2016. Date of publication May 18, 2016; date of current version November 23, 2016. The Associate Editor for this paper was V. Punzo.

K.-H. Lee and J.-N. Hwang are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: ykhlee@uw.edu; hwang@uw.edu).

G. Okopal and J. Pitton are with the Applied Physics Laboratory, University of Washington, Seattle, WA 98195 USA (e-mail: okopal@apl.washington.edu; pitton@apl.washington.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2557763

Human tracking is quite challenging since humans may vary greatly in appearance due to different viewing perspectives, non-rigid deformations, intra-class variability in shape and other visual properties. Especially under moving cameras, tracking of human becomes more challenging than that in static cameras, because of the combined effects of egomotion, blur, and the issues mentioned above. 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, thus makes the task more difficult. Instead of using background modeling based methods to extract the human blobs, human detectors are widely used to detect the human in the video frame. Therefore, the challenge is to successfully detect the humans in moving cameras, and then apply the tracking techniques to the detected ones, resulting in the so-called tracking-by-detection schemes. However, human detectors may effectively extract human blobs, they still have some limitations. First, human detectors may produce false alarm or miss detection; such errors can be reduced by further applying effective tracking techniques. Second, when humans are partially/fully occluded, the detections can fail and the tracking can be unreliable until the human 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.

In order to effectively track the humans and robustly handle occlusion problems in a ground-moving platform, a 3-D based tracking scheme with ability of occlusion handling is critically needed. Taking advantage of the tracking-by-detection scheme, we propose a robust ground-moving platform based human tracking system, which successfully integrates V-SLAM, human detection, ground plane estimation, and kernel-based tracking techniques. In this paper, we extend our previous work in [4], analyzing the performance of the proposed system, and experimenting on more videos of different scenarios. The proposed system starts with human detection and V-SLAM for camera calibration. Then, the ground planes are estimated based on the camera motions, so that we can infer the 3-D locations (relative to the cameras) of the humans. By taking 3-D information into account, we reformulate the tracking problem based on the Constrained Multiple-Kernel (CMK) approach [5], 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 humans effectively, but can also robustly handle occlusion during tracking, especially for ground moving platforms such as the unmanned ground vehicles.

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 human tracking, which combines the CMK tracking with the 3-D information to formulate the association of the detected targets. The experimental results are shown in Section V, followed by the conclusion in Section VI.

II. Related Work

Recently, tracking-by-detection is becoming a widely used scheme for human tracking. By applying a human detector to each frame of a video sequence to detect the existence of humans, the tracking scheme becomes a task to associate the detected human objects with each other frame-by-frame.

In general, human detection follows two basic steps [3]: foreground segmentation and object classification. Foreground segmentation first extracts blobs of interest from the image frame, avoiding as many background regions as possible. Then, object classification classifies the extracted blobs as humans or non-humans. 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 [6], [7]. The motion-based approaches utilize inter-frame motion and optical flow [8], [9], so as to minimize the inclusion of background regions, and extract candidates by considering the aspect ratio, size, and position. The approaches to object classification are mainly 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 classes and perform correlation between the image and the human body templates [10], [11]. The appearance methods [12]–[17] define a set of image features (descriptors), and a classifier is pre-trained by positive examples (human objects) and negative examples (non-human objects) with various learning algorithms, such as neural network, SVM, AdaBoost and etc.

The human detector proposed in [13] considers histogram of oriented gradient (HOG) as the features since HOG can efficiently represent the shape of human. The implicit shape model (ISM) [14] human detector 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) [15] extends the idea of [13], 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. C4 human detector [16] adopts a local-binary-pattern-like feature to efficiently represent the human template and achieve real-time performance. Recently, a human detector which extracts low-level visual features based on spatial pooling is proposed in [17]. These human detectors can be embedded independently in the proposed system, so as to functionally perform human detection.

After human object detection, a tracking framework is applied to the detected objects. There have been literatures of human tracking with moving cameras, e.g., Kalman filters [11], [18] and Particle filters [19], [20] are widely used in tracking. Some literatures adopt Multi-Hypothesis Tracking (MHT) [21], [22] and Joint Probabilistic Data Association Filters (JPDAFs) [23] to optimize detected target association by considering information over several time steps. Ess *et al.* [24] perform multi-body tracking by combining an ISM detector [14] and a stereo-odometry-based tracker. Adnriluka *et al.* [25] detect people using a part-based detector [15], and then use a Gaussian process latent variable model to compute the temporal consistency of detections over time. Leibe *et al.* [26] 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. In addition to the frame-by-frame based target association techniques mentioned above, recently, more and more methods tend to find globally optimal solutions across the entire sequence. Some approaches formulate the tracking problem as a min-cost flow network problem, and others use iterative hierarchical methods to link tracklets. Zhang *et al.* [27] map the maximum-a-posteriori (MAP) data association problem into a cost-flow network with a non-overlap constraint on trajectories. In [28], the authors use a cost function with the objects' birth and death states, and show that the global solution can be obtained with a greedy algorithm. In [29], Wu *et al.* propose a coupling formulation to avoid the problem of error propagation in tracking-by-detection scheme, and further solve the partial/complete occlusions. The approach in [30] incorporates constraints of piecewise constant-velocity path smoothness based on the flow network framework. Yang *et al.* [31] create a conditional random field from the set of tracklets to remove the assumption of independence between the tracked objects. 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 can also be applied to the moving cameras, combined with V-SLAM, to calibrate and to localize the 3-D positions of the camera and static features [32], [33]. 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 [26], 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.* [34] 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 approach in [35] estimates the ground plane by using sparse features, dense inter-frame stereo and object detection based on a real-time monocular SfM framework. Unlike the

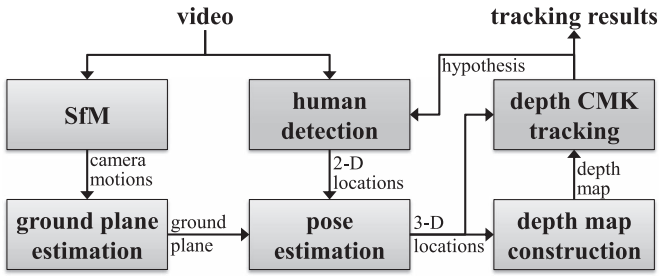


Fig. 1. Overview of the proposed system.

tracking-by-detection methods mentioned above, we take advantage of the SfM framework to locate pedestrians in 3-D space, and combine CMK tracking with the 3-D information, so as to deal with the partial/total occlusion problem during tracking.

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 within f_s video frames from a ground-moving platform, where we assume the height of the camera is known. Meanwhile, a pre-trained human detector is adopted to detect humans in the video frames. 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 humans' 2-D locations to 3-D locations. Based on the humans' 3-D locations relative to the camera motion, a depth map is constructed to represent the relative depth of the detected humans. Finally, the proposed Depth CMK tracking technique which combines the CMK tracking and the depth information is used to globally associate the detections frame-by-frame. If a detection is missing during the tracking, a hypothesized association (detection) is inserted based on the tracking results, so that the detected humans can be tracked continuously.

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 that 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

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

where X_p is the 3-D location of the pth feature point, x_{pq} is the pth observed 2-D location corresponding to X_{pq} from the qth video frame (the original formulation was referred to the qth camera), $\hat{\mathbf{P}}_q$ is the projective matrix of the qth frame, $\hat{\mathbf{P}}_q \cdot X_{pq}$ is the reprojection of X_p onto the qth 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. Human Detection

For human detection, we try to use some existing pre-trained human detectors [13]–[17], for comparative study, to detect humans in the video frames, i.e., the HOG-based human detector [13], the ISM human detector [14], the DPM human detector [15], the C4 human detector [16], and the spatial pooling based human detector [17]. These human detectors can be embedded independently in the proposed system, so as to functionally perform human 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 detected human blobs are first applied saliency detection [36], and then morphological operations (closing + opening + closing with 3×3 square structuring elements) to accurately obtain the segmentations. However, if the saliency detection is not sufficient, we thus create an ellipse mask to be the segmentation. To determine the saliency detection is sufficient, we calculate the ratio of the saliency area over the whole area of the human blob. If the ratio is larger than a threshold, which is empirically chosen to be 0.7, this saliency detection is regarded as sufficient.

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 re-estimated 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 re-estimate the ground planes in every f_g frames, $f_g \leq f_s$. Therefore, we collect f_g ground planes, each is calculated by each pair of consecutive camera motions, to form a set of ground planes $\{(\mathbf{g}_q, \psi_q)\}$, where $\mathbf{g}_q \in \mathbb{R}_3$ is the normal vector of the ground plane and $\psi_q \in \mathbb{R}$ is the offset of the plane. Finally, we can combine them into a single 4-by-f_g matrix $\mathbf{D}$

$$\mathbf{D} = \left[(\mathbf{g}_q, \psi_q)^{\mathrm{T}} \cdots \left(\mathbf{g}_{q+f_g}, \psi_{q+f_g}\right)^{\mathrm{T}} \right]. \tag{2}$$

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

To recover the uncorrupted version from the corrupted matrix $\mathbf{D}$, we employ the Robust Principle Component Analysis (RPCA) [37] to extract a low-rank 4-by-f_g matrix $\mathbf{A}$ from $\mathbf{D}$. Thanks to the characteristic of RPCA, the low rank matrix $\mathbf{A}$ consists of the uncorrupted data which represent more reliable ground planes for the application. The mean vector of the

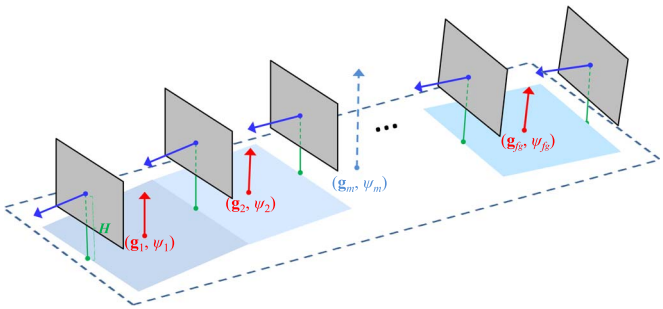


Fig. 2. Example of the ground plane estimation. Gray planes are the video frames, and H is the height of the camera. The final ground plane for f_g ground planes (dot-line plane) is obtained by a set of ground planes (solid planes).

matrix $\mathbf{A}$, annotated by $(\mathbf{g}_m, \psi_m)$, is considered to be our final ground plane for those f_g consecutive frames and is more resilient to the existence of noise in the system. Fig. 2 shows an example of using a ground plane set $\{(\mathbf{g}_q, \psi_q)|q = 1, \ldots, f_g\}$ to estimate the final ground plane $(\mathbf{g}_m, \psi_m)$.

D. Constrained Multiple-Kernel Tracking

The objective of the CMK tracking [5] is to retrieve a candidate object, which can be described as multiple kernels with pre-specified constraints among these kernels, so that the minimum mismatch (cost) between the tracked object and the candidate model can be reached. When a target is described by color histogram in the feature space, in order to differentiate the contribution of a pixel at different spatial location in the target area, a spatially-weighted color histogram, the so-called kernel, is used to represent the tracked object and the candidate model.

For an object described by 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 defined to be the Kullback-Leibler (K-L) distance between two spatially-weighted color histogram related to a specific tracked target part

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

Moreover, the constraint functions $\mathbf{C}(\mathbf{x}) = \mathbf{0}$ need to be considered to confine the kernels based on their spatial inter-relationships. Thus, the problem can be further formulated by

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

In order to gradually decrease the total cost function and maintain 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 [38] to iteratively solve the constrained optimization problem. 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., $\mathbf{C}(\mathbf{x}) = \mathbf{0}$)

$$\begin{aligned}\delta_\mathbf{x} &= \alpha\left(-\mathbf{I} + \mathbf{C}_\mathbf{x}\left(\mathbf{C}_\mathbf{x}^T\mathbf{C}_\mathbf{x}\right)^{-1}\mathbf{C}_\mathbf{x}^T\right)\mathbf{J}_\mathbf{x} + \left(-\mathbf{C}_\mathbf{x}\left(\mathbf{C}_\mathbf{x}^T\mathbf{C}_\mathbf{x}\right)^{-1}\mathbf{C}(\mathbf{x})\right) \\ &= \delta_\mathbf{x}^A + \delta_\mathbf{x}^B\end{aligned} \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state vector; $\mathbf{C}(\mathbf{x}) = [c_1(\mathbf{x}) \cdots c_m(\mathbf{x})]^T$ is the matrix including m constraint functions, and $c_j(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ is the i-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 $\mathbf{x}$, and $\alpha > 0$ is the step size.

As proved in [5], $\delta_\mathbf{x}^A$ and $\delta_\mathbf{x}^B$ are orthogonal to each other. Moving along the $\delta_\mathbf{x}^A$ decreases the total cost function $J(\mathbf{x})$ while keeping the same values of $\mathbf{C}(\mathbf{x})$. On the other hand, moving along the $\delta_\mathbf{x}^B$ can lower the absolute values of $\mathbf{C}(\mathbf{x})$. Owing to these characteristics, the optimal solution can be reached in an iterative manner. The iteration is stopped until either the cost function and the absolute values of constraint functions are both lower than some given thresholds ε_j and ε_c respectively, or the iteration count is larger than a threshold T (Algorithm 1 in [5]).

When occlusion occurs, not all the kernels can be used for matching. To solve the issue, each kernel is assigned with an adaptively adjustable weight w_κ in (3)

$$J(\mathbf{x}) = \sum_{\kappa=1}^{N_k} w_\kappa \cdot J_\kappa(\mathbf{x}). \tag{6}$$

Thus, the movement vector in (5) is modified to be

$$\delta_\mathbf{x} = \alpha\left(-\mathbf{I} + \mathbf{C}_\mathbf{x}\left(\mathbf{C}_\mathbf{x}^T\mathbf{C}_\mathbf{x}\right)^{-1}\mathbf{C}_\mathbf{x}^T\right)\mathbf{W}\mathbf{J}_\mathbf{x} + \left(-\mathbf{C}_\mathbf{x}\left(\mathbf{C}_\mathbf{x}^T\mathbf{C}_\mathbf{x}\right)^{-1}\mathbf{C}(\mathbf{x})\right) \tag{7}$$

where $\mathbf{W} = \begin{bmatrix} w_1\mathbf{I} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{N_k}\mathbf{I} \end{bmatrix}$ and $\mathbf{I}$ is an $(n/N_k) \times (n/N_k)$ identity matrix, n is the dimension of the state space.

The value w_κ corresponding to the κ-th kernel is adaptively updated according to the inverse of the distance value $\mathbf{J}_\kappa(\mathbf{x})$ and is normalized to make the sum equal to N_κ. The idea of the adaptation is that the movement vector will have more confidence on a kernel with smaller distance than that with the higher one. When a kernel is ineffective due to occlusion, the movement vector will adaptively count on the other effective kernels (i.e., kernels with lower distance) [5].

IV. Depth CMK Human Tracking

Once obtaining the 3-D locations of the humans 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 minimum color distance. 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 humans, but also well handles occlusions.

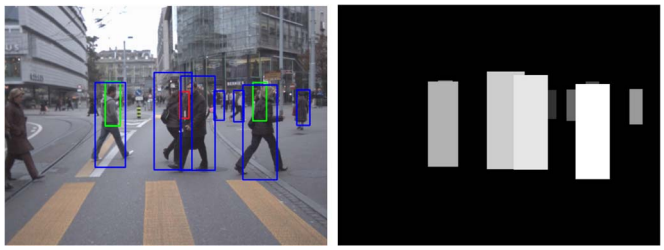


Fig. 3. Example of the depth map, showing (a) detections and (b) depth map, where higher intensity indicates detected humans are closer to the camera.

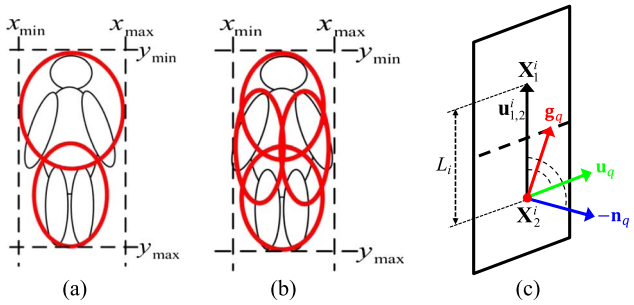


Fig. 4. Layout for (a) two kernels and (b) four kernels, both in 2-D space. (c) Illustration of the 3-D-based constraints in the case of a 2-kernel layout.

A. Depth Map Construction

Based on the 3-D locations of the humans, we can construct a depth map to describe the relative 3-D locations of all the tracked targets. Fig. 3 shows an example of the depth map, where Fig. 3(a) shows the result of human detection and Fig. 3(b) is the corresponding depth map. The depth map depicts the relative distance between the camera and the detected humans, higher (brighter) intensity means the detected humans are closer to the camera. As shown in the Fig. 3(a), two green-boxed targets are severely 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 ith target is occluded or not, in terms of the visibility $v_i \in [0, 1]$

$$v_i = \frac{\text{visible area of } i\text{th target}}{\text{total area of } i\text{th target}}. \tag{8}$$

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

B. Problem Formulation

In [5], the CMK tracking scheme tracks video objects in 2-D space (image), i.e., $\mathbf{x} \in \mathbb{R}^{2 \cdot N_k}$ in (6). To efficiently integrate the depth information into the CMK framework, we need to reformulate the problem. First we extend the (6) from 2-D to 3-D space

$$J^i(\mathbf{X}) = \sum_{\kappa=1}^{N_k} w_\kappa^i \cdot J_\kappa^i(\mathbf{X}),\ \mathbf{X} \in \mathbb{R}^{3 \cdot N_k}. \tag{9}$$

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

$$J(\mathbf{X}) = \sum_{i=1}^{N_q} v_i \cdot J^i(\mathbf{X}) = \sum_{i=1}^{N_q} v_i \cdot \left(\sum_{\kappa=1}^{N_k} w_\kappa^i \cdot J_\kappa^i(\mathbf{X}) \right) \tag{10}$$

where N_q is the number of the targets in the qth video frame, $\mathbf{X} \in \mathbb{R}^{3 \cdot N_q \cdot N_k}$, and $\mathbf{X} = [(\mathbf{X}_1^1)^T \ \cdots \ (\mathbf{X}_\kappa^i)^T]^T$ for the ith target and the κth kernel.

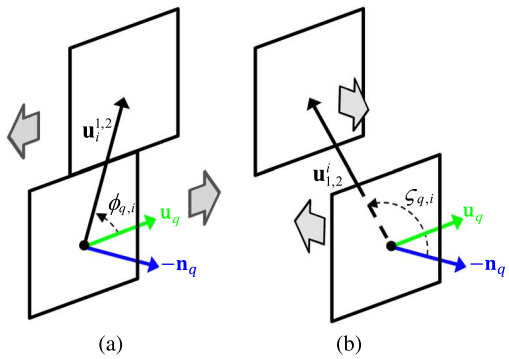


Fig. 5. Constraints for binding two kernels in 3-D space along the (a) left–right direction and the (b) forward–backward direction.

Necessarily, the constraint functions $\mathbf{C}(\mathbf{X}) = \mathbf{0}$ must be considered to maintain the relative locations of the kernels. In [5], 2-kernel and 4-kernel layouts are proposed to describe a human, as shown in Fig. 4(a) and (b). Unlike the constraints used in [5], 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. 4(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 the ith target in the qth frame; where $\mathbf{n}_q$ is the normal vector of the qth 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 $\mathbf{u}_{1,2}^i = \mathbf{X}_1^i - \mathbf{X}_2^i$. First, the distance between two kernel centers should be the same, which implies

$$\left\|\mathbf{u}_{1,2}^i\right\|^2 = (L_i)^2, \text{ for the } i\text{th target} \tag{11}$$

where L_i is the initial distance between $\mathbf{X}_1^i$ and $\mathbf{X}_2^i$. Second, both the angle between $\mathbf{u}_q$ and $\mathbf{u}_{1,2}^i$, and the angle between $-\mathbf{n}_q$ and $\mathbf{u}_{1,2}^i$, should be consistent. Therefore, we can have

$$\begin{cases} \frac{\mathbf{u}_q \cdot \mathbf{u}_{1,2}^i}{\|\mathbf{u}_q\| \left\|\mathbf{u}_{1,2}^i\right\|} = \cos(\phi_{q,i}) \\ \frac{-\mathbf{n}_q \cdot \mathbf{u}_{1,2}^i}{\|-\mathbf{n}_q\| \left\|\mathbf{u}_{1,2}^i\right\|} = \cos(\varsigma_{q,i}) \end{cases}, \text{ for the } i\text{th target in the } q\text{th frame.} \tag{12}$$

Those constraints can bind the kernels with each other in the 3-D space during tracking. The constraint with angle $\phi_{q,i}$ limits right-left movement of the kernels, as shown in Fig. 5(a); while the constraint with angle $\varsigma_{q,i}$ limits forward-backward movement of the kernels, as shown in Fig. 5(b).

(a) (b)

Fig. 6. 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.

TABLE I
CONFIGURATIONS OF THE DATASETS

Sequence	Resolution	# frames	frame per second (fps)
ETHMS Seq#1	640×480	1000	15
ETHMS Seq#2	640×480	800	15
ETHMS Seq#3	640×480	320	15
ETHMS Seq#4	640×480	800	15
Downtown#1	1920×1080	580	30
Downtown#2	1920×1080	220	30
UWcamp#1	1920×1080	1400	30

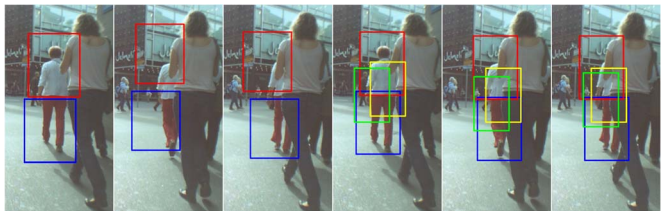
Hence, for each target i, the movement vector $\delta_{\mathbf{X}}$ can be iteratively solved by using the projected gradient method

$$\delta_{\mathbf{X}} = \alpha\left(-\mathbf{I} + \mathbf{C}_{\mathbf{X}}\left(\mathbf{C}_{\mathbf{X}}^T\mathbf{C}_{\mathbf{X}}\right)^{-1}\mathbf{C}_{\mathbf{X}}^T\right)\mathbf{V}\mathbf{W}\mathbf{J}_{\mathbf{X}} + \left(-\mathbf{C}_{\mathbf{X}}\left(\mathbf{C}_{\mathbf{X}}^T\mathbf{C}_{\mathbf{X}}\right)^{-1}\mathbf{C}(\mathbf{X})\right) \quad (13)$$

where $\mathbf{C}(\mathbf{X}) = [c_1(\mathbf{X})\cdots c_m(\mathbf{X})]^T$ is the matrix including m constraint functions, and $c_j(\mathbf{X}) : \mathbb{R}^{3\cdot N_q\cdot N_k} \rightarrow \mathbb{R}$ is the jth constraint function, $\mathbf{V} = \begin{bmatrix} v_1\mathbf{I}_v & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_{N_q}\mathbf{I}_v \end{bmatrix}$, $\mathbf{I}_v$ is an $3N_k \times 3N_k$ identity matrix, $\mathbf{W} = \begin{bmatrix} w_1\mathbf{I}_w & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{N_k}\mathbf{I}_w \end{bmatrix}$, $\mathbf{I}_w$ is an $3N_q \times 3N_q$ identity matrix, and $\mathbf{J}_{\mathbf{X}}$ is the gradient vector of the total cost function with respect to $\mathbf{X}$.

C. Hypothesized Association

However, due to unreliable detection or occlusion, humans 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. Inserting hypothesized association not only improves detection rate but also helps to continuously track the targets. Fig. 6 shows an example of the hypothesized association in case of occlusion. The person wearing gray clothes is detected in frame 230, but is then fully 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. 6(b) (see Table I).



(a) (b)

Fig. 7. Example of the CMK tracking when a tracked pedestrian is partially occluded, in the (a) 2-kernel case and the (b) 4-kernel case.

On the other hand, if a tracked target cannot be successfully associated to detections for several frames (empirically set as 5 in this work), this target is treated as a missed target.

V. EXPERIMENTAL RESULTS

In this section, we show experimental results of the proposed system on the ETH Mobile Scene (ETHMS) dataset [24], which is very challenging and difficult because of the heavy occlusions and busy crowd in the camera views. We test only four sequences (#Seq1 $\sim$ #Seq4) because these sequences are widely tested in other methods [24], [28]. 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 [24], [27], [28]. In order to have a fair comparison of our proposed scheme with previous works, we use the same video sequences as used in their simulations. The configurations of the tested videos are shown in Table I. All the experiments are processed on a personal computer with a P4 2.67 GHz 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. Here we assume the height of the camera is known. In the human detection, the pre-trained detectors [13]–[17] are functionally applied to the proposed system to detect humans within the 30 meters. In the CMK tracking, K-L distance is used for all distance measures, and the 8-bins histogram of the object is constructed based on the HSV color space with a roof kernel. The proposed approach is compared with the following three approaches. The approach in [24] is a stereo algorithm based on graphical model. The approach in [28] 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. 2-Kernels vs 4-Kernels

As shown in Fig. 4, The C2K uses 2 kernels to describe a human body, while the C4K uses 4 kernels. These kernels are generated systematically and automatically once the multiple-kernel tracking is activated. When occlusions occur, the C4K has better performance than the C2K since more kernels can allow more partial occlusion scenarios. Fig. 7 shows a typical example of a partially occluded pedestrian. The 4-kernel

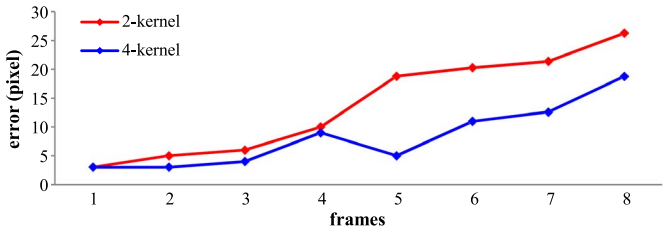


Fig. 8. Error (pixel) of the tracking results during the partial occlusion.

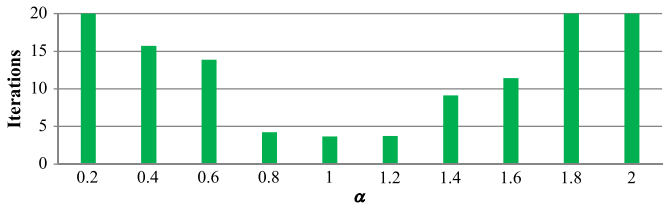


Fig. 9. Example of the converging iterations with different step sizes (α), where $\varepsilon_{\rm ms} = 10$ mm, and $T = 20$.

case is shown in Fig. 7(a). When a kernel is occluded, other non-occluded kernels can still shift to desired positions, so as to achieve a good tracking result for the next frame. This is because the occlusions often happen to right side or left side, and the 2 additional side-kernels can provide higher robustness. As for the 2-kernel case shown in Fig. 7(b), the kernels may possibly shift toward wrong directions because of mismatch between the target and the occlusion. Fig. 8 shows the error (pixel) of the tracking results during the partial occlusion, where the error is defined as the distance between the tracked target's centroid and the ground truth. The results start from the same detection, with the same error in frame 1. When occlusion occurs (frames 4 $\sim$ 8), the error of the 2-kernel case increases more rapidly than that of 4-kernel case. This indicates that C4K performs better during partial occlusions.

However, in general, the 2-kernel case performs better than the 4-kernel case. As described in Section IV-B, two kernels require three constraints to limit right-left movement and forward-backward movement of the kernels. On the other hand, four kernels require twelve constraints, in which three constraints for each pair of kernels. Such constraints sometimes deteriorate the convergence of the optimization, often resulting in worse tracking performance. Although the C4K is better than the C2K in case of partial occlusions, it still cannot compensate the disadvantages in general cases. Therefore, the overall performance of the C2K is better than that of the C4K for whole videos in different testing datasets, as shown in Table III.

B. Convergence

The convergence of (10) relies on the step size α. To speed up the convergence of the optimization, one possible way is to increase the step size α. However, increasing α also increases the probability of oscillation. If the mean-shift vector is smaller than a threshold ε_{ms}, or the iterations are up to T, the optimization is regarded as converged. Fig. 9 shows an example of the converging iterations with different step sizes. From the figure, when α is smaller, the optimization takes more iterations to converge. On the contrary, when α is larger, the optimization still needs many iterations due to the oscillation of gradient descent. For example, with $\alpha \leq 0.6$ or $\alpha \geq 1.6$, the convergence requires more than 10 iterations. As for $0.8 \leq \alpha \leq 1.2$, the optimization can converge in 5 iterations. Thus, to achieve better performance, we empirically choose $\alpha = 1.0$ for the step size, and $\varepsilon_{ms} = 10$ mm, $T = 20$ for the stopping criterion.

TABLE II
COMPARISON OF DETECTION RATE AND FPPI

Method	Detector	Detection rate (%)	False Positive Per Image (FPPI)
[24]	ISM	47	1.5
[24]	HOG	67.5	1.0
[28]	DPM	49.53	0.93
[28]	SP	51.86	0.92
[28] + NMS	DPM	49.94	0.93
[28] + NMS	SP	52.05	0.92
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
C2K	SP	77.82	0.83

C. Detection Performance

The competing approaches are evaluated with different human detectors, in terms of the detection rate and false positive per image (FPPI). This shows the performance of inserting hypothesized association during tracking. The results of the ETHMS dataset are shown in Table II. As shown in the table, the approaches with SP achieve slightly better performance than that with other detectors, indicating that SP provides more robust detections to facilitate better tracking. Nonetheless, the tracking performance of the DPM is comparable with that of the SP, which implies the DPM can provide sufficient detections during our CMK tracking. The results show that both the proposed approach (C2K, C4K) and the approach in [24] are superior to the approach in [28]. Since both approaches further utilize the 3-D information, instead of 2-D information only in [28], they can 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. This implies that missing detection can be improved by the tracking techniques, and thus better detection results benefit the tracking performance.

D. Tracking Performance

To evaluate the tracking performance, we consider the following metrics which are widely used in the previous work [24], [27]:

- 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%.

TABLE III
TRACKING RESULTS COMPARISON

Dataset	Method / Detector	GT	MT	PT	ML	FM	IDS
ETHMS Seq#1	[24] / ISM	73	**66**	**5**	2	**8**	1
	[28] / DPM	73	54	13	6	19	8
	[28] + NMS / DPM	73	55	12	6	19	8
	C4K / DPM	73	58	10	5	7	2
	C2K / DPM	73	**64**	**7**	2	**3**	3
	C2K / SP	73	**66**	**5**	2	**3**	3
ETHMS Seq#2	[24] / ISM	55	**46**	**5**	4	**9**	0
	[28] / DPM	55	36	10	9	11	0
	[28] + NMS / DPM	55	36	10	9	11	0
	C4K / DPM	55	43	5	8	7	0
	C2K / DPM	55	**43**	**8**	5	**7**	0
	C2K / SP	55	**44**	**7**	5	**7**	0
ETHMS Seq#3	[28] / DPM	24	**17**	5	2	**1**	2
	[28] + NMS / DPM	24	17	5	2	1	2
	C4K / DPM	24	17	5	2	0	1
	C2K / DPM	24	17	5	2	0	1
	C2K / SP	24	**18**	4	2	**0**	1
ETHMS Seq#4	[24] / ISM	88	**74**	**8**	6	**20**	3
	[28] / DPM	88	52	16	20	29	14
	[28] + NMS / DPM	88	55	14	19	29	14
	C4K / DPM	88	64	14	10	18	6
	C2K / DPM	88	**71**	**9**	8	**11**	6
	C2K / SP	88	**72**	**8**	8	**11**	6
Downtown#1	[28] / DPM	5	**4**	**1**	0	**1**	0
	[28] + NMS / DPM	5	4	1	0	1	0
	C4K / DPM	5	5	0	0	0	0
	C2K / DPM	5	**5**	0	0	**0**	0
	C2K / SP	5	**5**	0	0	**0**	0
Downtown#2	[28] / DPM	38	10	10	15	2	1
	[28] + NMS / DPM	38	**10**	10	15	**2**	1
	C4K / DPM	38	12	9	16	1	0
	C2K / DPM	38	**13**	8	16	**1**	0
	C2K / SP	38	**15**	8	14	**1**	0
UWcamp#1	[28] / DPM	13	9	3	0	1	0
	[28] + NMS / DPM	13	**9**	3	0	**1**	0
	C4K / DPM	13	12	1	0	0	0
	C2K / DPM	13	**12**	1	0	**0**	0
	C2K / SP	13	**13**	0	0	**0**	0

- FragMentation (FM): the number of times a trajectory is interrupted.
- ID Switches (IDS): the number of times two trajectories switch their IDs.

The results of several sequences in dataset are shown in Table III, where GT denotes the number of trajectories in the ground truth. Higher MT, lower FM, and lower IDS imply the better performance. From the MT results, the approach in [28] is not as good as other approaches. Due to the limitation of 2-D information, the approach in [28] cannot well solve occlusion issues, thus results in lower MT and higher FM. On the other hand, the 3-D based approaches are able to significantly improve the performance by efficiently handling occlusions. The C2K performs better than the C4K, because the detected humans are too small to be clearly described by 4 kernels, which easily include some additional background regions so that the tracking is easily impacted by the background. Besides, there are more constraints for binding 4 kernels, making the optimization of the cost function more divergent. When comparing the C2K with the approach in [24], although the MT/PT/ML results are comparable, the FM results of the CMK are much better than that of [24]. This implies that the proposed 3-D based CMK framework can effectively associate the targets, so as to perform well on successively tracking the targets. Moreover, the consistent improvement of the proposed approach indicates that the 3-D based tracking scenario is much better than the 2-D based ones. Several visual tracking results are shown in Fig. 10, where (a), (b), (c), (d) are the ETHMS dataset, and (e), (f), (g) are the datasets recorded by ourselves. The results show favorable performance of the proposed system, not only successively tracking humans but also well handling occlusion in the tracking. Moreover, since relative 3-D locations of the humans are obtained, with GPS information, we can also construct 3-D visualization of the dynamic scenes. Fig. 11 shows the 3-D visualization of the dataset in Fig. 10(f) and (g), showing the dynamic scenes in different aspects of views. All the videos associated with the simulations reported in this paper can be viewed our website.[1]

E. Discussions

The proposed system tracks the pedestrians in 3-D space, which is facilitated by the ground plane estimation and V-SLAM framework. Such 3-D based tracking-by-detection scheme can not only track the humans effectively, but can also robustly handle occlusion during tracking. Nevertheless, there are several limitations exist. First, the proposed approach adopts the tracking-by-detection scheme to detect and then track humans; this implies that the approach highly relies on the detection results. However, if the quality of video sequences is not sufficient for the human detector(s), the proposed approach is not able to perform well on the poor detection results. More specifically, the tracking of a specific human is always triggered by the positive detection of a target(s). In other words, the proposed approach may not work well at night or some cases of insufficient lighting. Second, the proposed approach effectively estimates ground planes based on certain video frames when since the platform moves on flat roads/fields, but will produce less reliable estimation if the roads/fields are severely bumpy, resulting in larger error of the object back-projection and impacting accuracy of the reprojected 3-D information. Hence, the proposed approach is not reliable for the unmanned aerial vehicle, because its height dynamically changes and then infers unreliable 3-D information of humans.

VI. CONCLUSION

We proposed a robust human tracking system in a moving camera. The proposed system effectively integrates the human detectors and V-SLAM framework to relocate the humans in 3-D space, followed by an innovative 3-D based CMK tracking, which not only locally associates the targets but also globally optimizes the associations according to the 3-D information. Such system can be regarded as a key component for high level applications, such as video analysis in a large scale of mobile network [39]–[41]. Besides, the proposed framework can also be further applied to other class of on-road objects if the corresponding detectors are available.

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

Fig. 10. Visual tracking results, from the top to the bottom: (a) ETHMS Seq#1, (b) ETHMS Seq#2, (c) ETHMS Seq#3, (d) ETHMS Seq#4, (e) Downtown Seq#1, (f) Downtown Seq#2, and (g) UWcamp Seq#1.

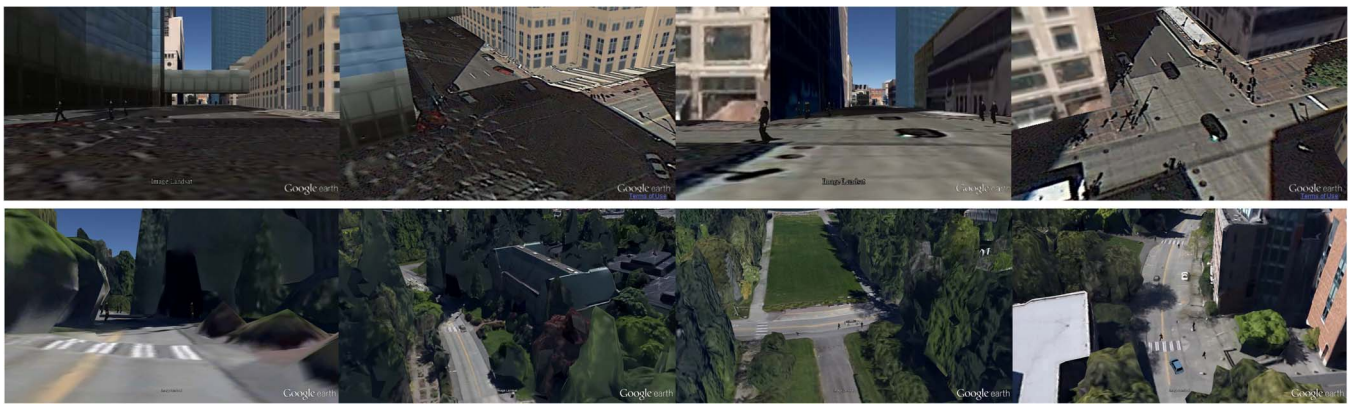


Fig. 11. 3-D visualization reconstructed from the video sequences, showing different view aspects; (top) Downtown Seq#2, (bottom) UWcamp Seq#1.

REFERENCES

[1] Google driveless car. [Online]. Available: http://en.wikipedia.org/wiki/Google_driverless_car

[2] Amazon Prime Air. [Online]. Available: http://en.wikipedia.org/wiki/Amazon_Prime_Air

[3] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[4] K.-H. Lee, J.-N. Hwang, G. Okopal, and J. Pitton, "Driving recorder based on-road pedestrian tracking using visual SLAM and constrained multiple-kernel," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst.*, Oct. 2014, pp. 2629–2635.

[5] 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. 15, no. 7, pp. 1602–1615, Nov. 2013.

[6] T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka, "Development of night-vision system," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 3, pp. 203–209, Sep. 2002.

[7] M. Bertozzi *et al.*, "Shape-based pedestrian detection and localization," in *Proc. IEEE Intell. Transp. Syst.*, 2003, vol. 1, pp. 328–333.

[8] H. Elzein, S. Lakshmanan, and P. Watta, "A motion and shape-based pedestrian detection algorithm," in *Proc. IEEE Intell. Veh. Symp.*, 2003, pp. 500–504.

[9] U. Franke and S. Heinrich, "Fast Obstacle Detection for Urban Traffic Situations," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 3, pp. 173–181, Sep. 2002.

[10] D. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: The protector system," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 13–18.

[11] D. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jun. 2007.

[12] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, vol. 2, pp. 734–741.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[14] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 259–289, May 2008.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[16] J. Wu, N. Liu, C. Geyer, and J. M. Rehg, "C4: A real-time object detection framework," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4096–4106, Oct. 2013.

[17] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.* [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7229360

[18] M. Bertozzi *et al.*, "Pedestrian localization and tracking system with Kalman filtering," in *Proc. IEEE Intell. Veh. Symp.*, 2004, pp. 584–589.

[19] J. Giebel, D. Gavrila, and C. Schnör, "A Bayesian framework for multi-cue 3D object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 241–252.

[20] R. Arndt, R. Schweiger, W. Ritter, D. Paulus, and O. Lóhlein, "Detection and tracking of multiple pedestrians in automotive applications," in *Proc. IEEE Intell. Veh. Symp.*, 2007, pp. 13–18.

[21] I. J. Cox, "A review of statistical data association techniques for motion correspondence," *Int. J. Comput. Vis.*, vol. 10, no. 1, pp. 53–66, Feb. 1993.

[22] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 843–854, Dec. 1979.

[23] T. E. Fortmann, Y. Bar Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, Jul. 1983.

[24] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1831–1846, Oct. 2009.

[25] M. Adnriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[26] B. Leibe, N. Cornelis, K. Cornelis, and L. van Gool, "Dynamic 3D scene analysis from a moving vehicle," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[27] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[28] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1201–1208.

[29] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, "Coupling detection and data association for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1948–1955.

[30] A. A. Butt and R. T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1846–1853.

[31] B. Yang and R. Nevatia, "Multi-target tracking by online learning a CRF model of appearance and motion patterns," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 203–217, Apr. 2013.

[32] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real time localization and 3D reconstruction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 363–370.

[33] M. Pollefeys *et al.*, "Detailed real-time urban 3D reconstruction from video," *Int. J. Comput. Vis.*, vol. 78, no. 2, pp. 143–167, Jul. 2008.

[34] A. Kundu, K. M. Krishna, and C. V. Jawahar, "Realtime multibody visual SLAM with a smoothly moving monocular camera," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2080–2087.

[35] S. Song and M. Chandraker, "Robust scale estimation in real-time monocular SfM for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1566–1573.

[36] C. Yang, L. Zhang, H. Lu, M.-H. Yang, and X. Ruan, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[37] 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," in *Proc. Neural Inf. Process. Syst.*, Dec. 2009, pp. 1–9.
[38] J. A. Snyman, *Practical Mathematical Optimization*. NewYork, NY, USA: Springer Sci. Bus. Media, 2005, ch. 3.
[39] K.-H. Lee and J.-N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Trans. Multimedia*, Special Issue Multimedia: The Biggest Big Data, vol. 17, no. 9, pp. 1429–1438, Sep. 2015.
[40] X. Chen, J.-N. Hwang, K.-H. Lee, and R. L. de Querioz, "Quality-of-content (QoC)-driven rate allocation for video analysis in mobile surveillance networks," in *Proc. IEEE 17th Int. Workshop. Multimedia Signal Process.*, Oct. 2015, pp. 1–6.
[41] X. Chen, J.-N. Hwang, D. Meng, K.-H. Lee, R. L. de Queiroz, and F.-M. Yeh, "A Quality-of-Content (QoC)-based joint source and channel coding for human detections in a mobile surveillance cloud," *IEEE Trans. Circuits Syst. Video Technol.* [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7428864&tag=1

Kuan-Hui Lee received the B.S. degree from National Taiwan Ocean University, Keelung, Taiwan, in 2003; the M.S. degree from National Cheng Kung University, Tainan, Taiwan, in 2005; and the Ph.D. degree from University of Washington, Seattle, WA, USA, in 2015.

From 2007 to 2009, he was with HTC Corporation, developing multimedia applications on smart phones. His research interests include computer vision, image processing, and machine learning.

Jenq-Neng Hwang (F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from University of Southern California, Los Angeles, CA, USA.

In the summer of 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, where he has been a Full Professor since 1999 and is currently the Associate Chair for Research in the Department of Electrical Engineering. He has written more than 300 journal and conference papers, and also book chapters in the areas of multimedia signal processing and multimedia system integration and networking, including an authored textbook entitled *Multimedia Networking: From Theory to Practice* (Cambridge University Press). He has close working relationship with the industry on multimedia signal processing and multimedia networking.

Dr. Hwang received the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He is a founding member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society and was the Society's representative to the IEEE Neural Network Council from 1996 to 2000. He is currently a member of the Multimedia Technical Committee of the IEEE Communication Society and of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He served as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE *Signal Processing Magazine*. He is currently on the Editorial Board of *ETRI Journal*, *International Journal of Data Mining and Bioinformatics*, and *Journal of Signal Processing Systems*. He was the Program Cochair of the 1998 International Conference on Acoustics, Speech, and Signal Processing and the 2009 International Symposium on Circuits and Systems.

Greg Okopal received the M.S. and Ph.D. degrees in electrical engineering from University of Pittsburgh, Pittsburgh, PA, USA.

In 2009, after having earned his degrees, he joined the Applied Physics Laboratory, University of Washington, Seattle, WA, USA, where he is currently a Senior Engineer. In 2008 and 2014, he was twice a Visiting Scientist with the NATO Center for Maritime Research and Experimentation. His research focuses on algorithms for signal processing, sensor fusion, and autonomy.

James Pitton received the Ph.D. degree in electrical engineering from University of Washington, Seattle, WA, USA, in 1994.

He is a Senior Principal Engineer with the Applied Physics Laboratory, University of Washington (APL-UW) and an Affiliate Associate Professor of electrical engineering with University of Washington. In 1999, he joined APL-UW, where he was the Head of the Environmental and Information Systems Department from 2002 to 2007. From 2007 to 2010, he was the Associate Director for Ocean and Undersea Science with the U.S. Office of Naval Research Global (ONR Global), London, U.K. He also held research positions with AT&T Bell Laboratories, Murray Hill, NJ, USA, and with the Statistical Sciences Division, MathSoft, Seattle. He has served on the organizing committees of numerous workshops and conferences, including the Workshop on Machine Intelligence for Autonomous Operations organized jointly between ONR Global, UK DSTL, and NURC. His ongoing research interests are focused on algorithms for information processing and autonomous systems, with an emphasis on sonar, automatic classification, nonstationary signal processing, and array processing.