Computer Vision and Image Understanding 153 (2016) 29–36

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Tracking by switching state space models

Junseok Kwon[a,*], Ralf Dragon[b], Luc Van Gool[b]

[a] School of Computer Science and Engineering, Chung-Ang Univercity, Seoul, South Korea
[b] Computer Vision Lab, D-ITET, ETH Zurich, Sternwartstrasse 7, Zurich 8092, Switzerland

ARTICLE INFO

Article history:
Received 14 July 2015
Accepted 5 March 2016

Keywords:
Visual tracking
Probabilistic model
State space switch
Ground plane

ABSTRACT

We propose a novel tracking method that allows to switch between different state representations as, e.g., image coordinates in different views or image and ground plane coordinates. During the tracking process, our method adaptively switches between these representations. We demonstrate the applicability of our method for dynamic cameras tracking dynamic objects: Using the image based representation (non-smooth trajectories if the camera is shaking) together with the ground plane based one (estimation uncertainty in visual odometry or ground plane orientation), the disadvantages of both representation forms can be overcome: Non-occluded observations on the image plane provide strong appearance cues for the target. Smooth paths on the ground plane provide strong motion cues with the camera motion factored out. Following a Bayesian tracking approach, we propose a probabilistic framework that determines the most appropriate state space model (SSM)—image or ground plane or both—at each time instance. Experimental results demonstrate that our method outperforms the state-of-the-art.

1. Introduction

Tracking a target in realistic videos still is highly challenging. Recent tracking methods proposed advanced appearance and motion models [1,6,7,11,14,15,25,27,29,31,35,36,39]. However, these methods still find it difficult to handle severe occlusions. A promising solution for such problems is to track the target on the ground plane. Most ground plane methods assume the use of multiple, fixed cameras [4,22,33,37]. On the ground plane the target is no longer occluded by the other objects. In a multi-camera setup, chances are also high that the target can be seen unoccluded by at least one camera, which then allows the tracker to use up-to-date appearances for the target. On the other hand, in most real cases, we only have a single video stream as input and such luxury cannot be depended on. Moreover, with the positions and settings of our camera changing over time, the ground plane estimated can more easily get flawed or may even become unavailable, thereby even impairing tracking performance. Thus, for our setup it would seem best to keep the image plane in mind as well, where appearances can be updated and the target can be successfully tracked when there is a clear view without occlusions.

Based on these observations, our method adaptively switches between state space models (SSMs) over time and takes advantage of both the image and ground planes, as shown in Fig. 1. The advantage of referring motions to the ground plane is that the camera motion is factored out. Hence the target motion becomes smooth, often mostly linear, and easy to track. The linearity assumption can help survive occlusions which we cannot circumvent by calling on other cameras in our case. When there are no occlusions or when no good ground plane estimation is available, the image plane can be used.

The challenge is then to appropriately select the SSMs that are used for each frame. We present a Bayesian framework that performs this scheduling. The SSMs are selected on the basis of confidence scores: a likelihood score for the observation and a smoothness score for the target motion, respectively. Without occlusions, the likelihood score calculated from the image plane tends to be larger than the smoothness score calculated from the ground plane. If that is the case, our method chooses the image plane as the preferable SSM and predicts the target location on the image coordinate. The likelihood score of an occluded target tends to decrease more than its smoothness score. When the smoothness score becomes larger than the likelihood score, the ground plane gets selected as the most appropriate SSM and is used to predict the target location. Our method can select the image and ground planes at the same time when both the likelihood and smoothness scores are large. In this case, the method gets the target location by combining all predictions obtained from the image and ground planes. In particular, our method follows the framework of Koolen et al. [23] to probabilistically select the activated SSMs.

* Corresponding author. Fax: +82 2 820 5301.
E-mail addresses: kwonj@vision.ee.ethz.ch, jskwon@cau.ac.kr (J. Kwon).

http://dx.doi.org/10.1016/j.cviu.2016.03.006
1077-3142/

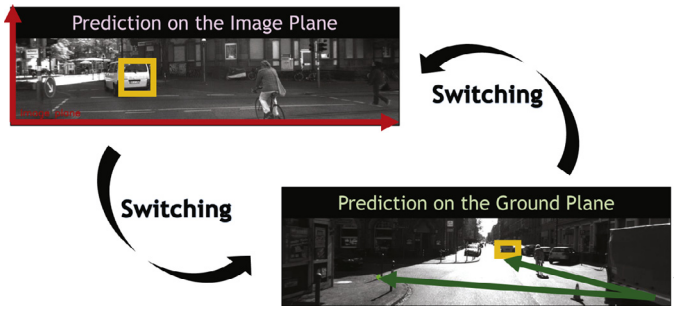


Fig. 1. Visual tracking by changing the SSMs. Our tracker adaptively selects its SSM over time to predict the target state, i.e. it uses either the image, or the ground plane, or both. Under occlusions, the ground plane is typically employed to predict the target state. While the camera motion is removed, the target motion becomes very smooth on the ground plane. Then this strong motion cue helps survive occlusions during the tracking process, although a position of the occluded target may not be accurately estimated. With our single camera setup, the ground plane is wrongly obtained at some frames. In these frames, our method uses the image plane alternatively.

2. Related work and contributions

Tracking methods that change the SSMs: The closest approach of our method is to switch target models, e.g. between multi-view face models to handle pose variations in face tracking [38], where the multi-view face models are represented by different SSMs. On the other hand, camera hand-over systems perform SSM changes as well. For instance, Zajdel et al. [42] switched between non-overlapping cameras to track multiple targets, where these cameras have different SSMs. Our approach is different in that different types of SSMs are mixed, for the same camera and for the same target representation. The methods proposed in Murphy [30] and Peursum et al. [32] switched the dynamic and observation models, while our method switches the state space model.

Tracking methods that use the ground plane: Tracking with the help of the ground plane, especially to counter occlusions, has been explored in several publications. Several of those also assumed the availability of multiple, static cameras [5,10,22,33,37] or needed fully calibrated cameras, e.g. Berclaz et al. [4]. While not many approaches use a single camera, our method uses a single and moving camera to get accurate tracking results.

Sampling based tracking methods: The particle filter was introduced by Isard and Blake [18] to deal with non-linear and non-Brownian motions. Our method also employs the particle filter for visual tracking, but does not fix one SSM and instead switches between multiple possibilities. That eases the handling of complex motion patterns even more, as one can use the SSM where they appear simplest (principle of parsimony). For instance, under shaky camera motions even simple linear trajectories look erratic in the image plane, but if ego motion and ground plane orientation is available, the trajectories are linear again in the ground plane coordinate system.

Kreucher et al. [24] used multiple kinematic models and adaptively determined which model best describes the target motion. Xu and Li [41] fused two different observation, namely color histogram and gradient. Liang-qun et al. [28] selected appropriate models for maneuvring target tracking in a cluttered environment. Hlinomaz and Hong [17] updated particle filter at multi rates to robustly track the target. Our method follows the philosophy of aforementioned methods using multiple models but proposed a more sophisticated strategy of determining appropriate models at each time.

3. Bayesian tracking with multiple SSMs

The goal of visual tracking is to find the state $\hat{\mathbf{X}}_t$ that maximizes the posterior $p(\mathbf{X}_t|\mathbf{Y}_{1:\,t})$, where $\mathbf{X}_t = (X_t^x, X_t^y, X_t^s)$ denotes the target state at time t which includes the x, y center positions and the scale s of the bounding box, and where $\mathbf{Y}_{1:\,t}$ indicates the observations up to time t. We keep the fixed aspect ratio of the bounding box over time. One can achieve this goal by using the MAP estimation:

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t} p(\mathbf{X}_t|\mathbf{Y}_{1:t}), \tag{1}$$

where

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \propto p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}). \tag{2}$$

In (2), the likelihood is designed as

$$p(\mathbf{Y}_t|\mathbf{X}_t) = e^{-\lambda_1 \mathrm{Dist}(\mathbf{Y}_t(\mathbf{X}_t), P_t)}, \tag{3}$$

where λ_1 is a weighting parameter, and the function Dist returns the appearance similarity between the candidate patch $\mathbf{Y}_t(\mathbf{X}_t)$ and the reference target patch P_t. The patch P_t is obtained by the ASLA method in Jia et al. [19]. In (3), $\mathbf{Y}_t(\mathbf{X}_t)$ indicates the observation $\mathbf{Y}_t$ inside of the bounding box described by $\mathbf{X}_t$.

In case the tracking is run through multiple state space models, one can specify the particular SSM, m, for the state prediction $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1})$ from (2), which yields an SSM prediction $p(m|\mathbf{Y}_{1:t-1})$ and a state prediction given the SSM, $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m)$. One can then collect information from a finite number of SSMs as:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}) = \sum_{i=1}^{M} p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m=i)p(m=i|\mathbf{Y}_{1:t-1}), \tag{4}$$

with M is the total number of SSMs.

In practice the predictions of some SSMs may be unavailable or very noisy during the tracking process. Following the recent work in Koolen et al. [23], we make the set of SSMs, $\mathcal{W}_t$, whose predictions are available at time t (*called awake set*), and also the set of SSMs, $\mathcal{S}_t$, whose predictions are not available at time t (*called asleep set*). If a SSM belongs to the awake set, our method predicts the state $\mathbf{X}_t$ with $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m)$. Otherwise, our method does not predict the state. Using the awake set $\mathcal{W}_t$, the state prediction therefore is:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}) = \frac{\sum_{i\in\mathcal{W}_t} p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m=i)p(m=i|\mathbf{Y}_{1:t-1})}{p(\mathcal{W}_t|\mathbf{Y}_{1:t-1})}, \tag{5}$$

with $p(\mathcal{W}_t|\mathbf{Y}_{1:t-1}) = \sum_{i\in\mathcal{W}_t} p(m=i|\mathbf{Y}_{1:t-1})$, i.e. the average prediction by awake SSMs.

The best state $\hat{\mathbf{X}}_t$ that maximizes the posterior is then found by using the Particle Filter (PF) Isard and Blake [18]. The PF represents the posterior $p(\mathbf{X}_t|\mathbf{Y}_{1:\,t})$ by using N particles: $p(\mathbf{X}_t^{(n)}|\mathbf{Y}_{1:t})$ for $n = 1, \ldots, N$. The MAP formulation in (1) is then

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t^n} p(\mathbf{X}_t^{(n)}|\mathbf{Y}_{1:t}) \text{ for } n = 1, \ldots, N, \tag{6}$$

where $\mathbf{X}_t^{(n)}$ denotes the nth particle. In (6), our method chooses the best particle that maximizes the posterior.

Because the PF is a well known sampling method, we briefly introduce it in terms of our visual tracking problem. The PF mainly consists of three steps, namely transition, weighting, and re-sampling steps. At each frame, our method first determines the awake set using (11). In the transition step, the method moves particles with the transition probability in (16). In the weighting step, the method calculates weights of particles using the likelihood in

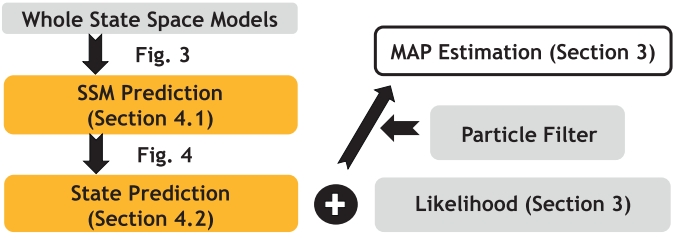


Fig. 2. Procedure of the proposed method.

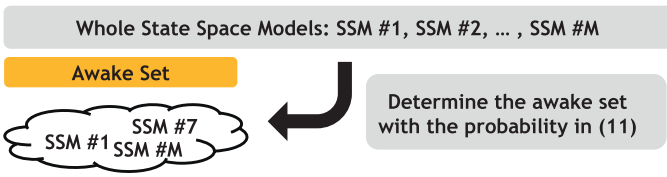


Fig. 3. State space model prediction. In Section 4.1, our method predicts SSMs by determining whether each SSM is awake or not.

(3). The re-sampling step is same with the conventional PF. Fig. 2 and Algorithm 1 illustrate the whole process of our method.

Algorithm 1 Tracking by changing the state space models.

Input: $\mathbf{X}_{t-1} = (\mathbf{X}^x_{t-1}, \mathbf{X}^y_{t-1}, \mathbf{X}^s_{t-1})$
Output: $\hat{\mathbf{X}}_t = (\hat{\mathbf{X}}^x_t, \hat{\mathbf{X}}^y_t, \hat{\mathbf{X}}^s_t)$
1: Determine the awake set $\mathcal{W}_t$ using (11).
2: **for** $n = 1$ to N **do**
3: Move the nth particle $\mathbf{X}^{(n)}_t$ using (16).
4: Calculate a weight of the particle using (3).
5: **end for**
6: Re-sampling step.
7: Estimate the best state $\hat{\mathbf{X}}_t$ using (6).

4. Modeling

$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1})$ in (2) is determined through (5). That equation in turn consists of the SSM prediction $p(m|\mathbf{Y}_{1:t-1})$ and the state prediction given the SSM $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m)$, which are derived in Sections 4.1 and 4.2, respectively.

4.1. SSM prediction: $p(m|\mathbf{Y}_{1:t-1})$

In order to predict the SSM, our method adopts the method of Koolen et al. [23], as shown in Fig. 3. If a new observation $\mathbf{Y}_t$ is available, $p(m|\mathbf{Y}_{1:t})$ is obtained from $p(m|\mathbf{Y}_{1:t-1})$ in the step-wise manner:

$$p(m|\mathbf{Y}_{1:t}) = \frac{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)}{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1})} p(m|\mathbf{Y}_{1:t-1}). \tag{7}$$

We update $p(m|\mathbf{Y}_{1:t-1})$ differently depending on the sets $\mathcal{W}_t$ and $\mathcal{S}_t$, where weights of asleep SSMs are left unaffected, whereas weights of awake SSMs are updated by using (7), as follows:

$$p(m|\mathbf{Y}_{1:t}) = \begin{cases} \alpha_t p(m|\mathbf{Y}_{1:t-1}) & \text{if } m \in \mathcal{W}_t \\ p(m|\mathbf{Y}_{1:t-1}) & \text{if } m \in \mathcal{S}_t \end{cases}, \tag{8}$$

where

$$\alpha_t = \frac{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)}{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1})}. \tag{9}$$

Determining the awake set $\mathcal{W}_t$: To efficiently update $p(m|\mathbf{Y}_{1:t-1})$ in (8), we define a Markov chain on the labels $l = \{w, s\}$ with the transition probabilities $\theta(l|w)$ and $\theta(l|s)$, that is

$$p(m, l|\mathbf{Y}_{1:t}) \leftarrow \alpha_t \theta(l|w) p(m, l = w|\mathbf{Y}_{1:t-1}) + \theta(l|s) p(m, l = s|\mathbf{Y}_{1:t-1}), \tag{10}$$

where w and s denote awake and asleep labels, respectively. In (10), $\theta(l|w)$ is the transition probability of preserving the label w or changing to the label s. $\theta(l|s)$ is the transition probability of preserving the label s or changing to the label w.

The basic idea behind the aforementioned updating strategy is to make an awake (asleep) SSM to be awake (asleep) at the next iteration with a high probability. The probability increases by multiplying α_t in (10) when a awake model is good to track the target at the current frame. Notably α_t measures the confidence of an SSM as addressed in following section. Then the mth SSM is determined as awake, like this:

$$m \in \mathcal{W}_t \text{ with the probability } \frac{p(m, w|\mathbf{Y}_{1:t-1})}{p(m, w|\mathbf{Y}_{1:t-1}) + p(m, s|\mathbf{Y}_{1:t-1})}, \tag{11}$$

while multiple SSMs can be determined as awake. If all SSMs are determined as asleep, our method awakes the mth SSM with the probability $\frac{p(m|\mathbf{Y}_{1:t-1})}{\sum_{j\in\mathcal{W}_t} p(m=j|\mathbf{Y}_{1:t-1})}$.

Estimating α_t: (9) consists of $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)$ and $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)$ includes the confidence score of the mth SSM, as follows:

$$p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m) \propto \int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m) d\mathbf{X}_t, \tag{12}$$

where $p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m)$ is the confidence of the mth SSM, which is designed in Section 5.2.

By using (5), (10), and (12), α_t in (9) is estimated by

$$\alpha_t \propto \frac{\int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m) d\mathbf{X}_t}{\sum_{i\in\mathcal{W}_t} \beta_t \int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m = i) d\mathbf{X}_t}, \tag{13}$$

where $\beta_t = \frac{p(m=i|\mathbf{Y}_{1:t-1})}{\sum_{j\in\mathcal{W}_t} p(m=j|\mathbf{Y}_{1:t-1})}$. In (13), the confidence integration over the state, $\int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m)\, d\mathbf{X}_t$, is approximated by the sum of all confidences obtained from particles in the PF. (12) and (13) are derived in Sections 7.1 and 7.2, respectively.

4.2. State prediction: $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m)$

Given the mth SSM, the state prediction described in Fig. 4 is

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, m) = \int p(\mathbf{X}_t|\mathbf{X}_{t-1}, m) p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1}) d\mathbf{X}_{t-1}, \tag{14}$$

where $p(\mathbf{X}_t|\mathbf{X}_{t-1}, m)$ is the state transition probability and $p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})$ is the posterior at time $t-1$. In (14), the equality holds because $\mathbf{X}_{t-1}$ is independent from m. To predict a new state, our method projects the previous state $\mathbf{X}_{t-1}$ into the mth SSM, by using the projection function $F_m(\cdot)$ and proposes a new state with the probability $p(\mathbf{X}_t|\mathbf{X}_{t-1}, m)$ as follows:

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}, m) = \mathcal{N}(F_m(\mathbf{X}_{t-1}), \sigma_m), \tag{15}$$

where $\mathcal{N}(F_m(\mathbf{X}_{t-1}), \sigma_m)$ denotes the Gaussian Normal distribution defined on the mth SSM, (mean $F_m(\mathbf{X}_{t-1})$, standard deviation σ_m).

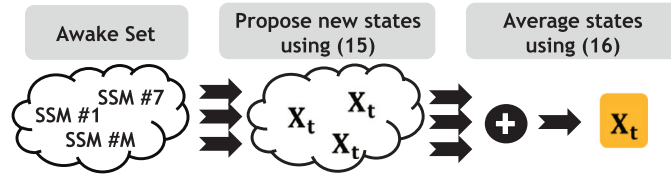


Fig. 4. State prediction. In Section 4.2, our method propose a new state by averaging the predicts obtained from all SSMs in the awake set.

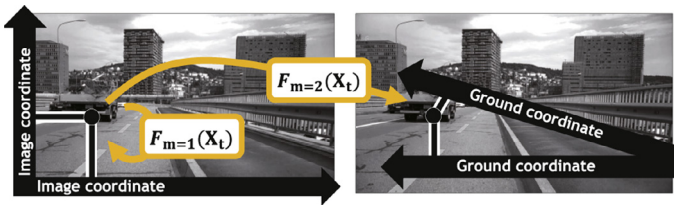


Fig. 5. The two SSMs Image Plane (left) and Ground Plane (Right).

If multiple SSMs are awake, our method averages state predictions made by all SSMs in the awake set:

$$\mathbf{X}_t \sim \sum_{i \in \mathcal{W}_t} \beta_t \, p(\mathbf{X}_t | \mathbf{X}_{t-1}, m = i). \tag{16}$$

where each SSM in the awake set is weighted by $\beta_t = \frac{p(m=i|\mathbf{Y}_{1:t-1})}{\sum_{j \in \mathcal{W}_t} p(m=j|\mathbf{Y}_{1:t-1})}$. For averaging in (16) we backproject all state predictions into the image plane and average them into a single state.

5. State space models for image and ground plane

In this section, we define two SSMs, corresponding to the image plane and the ground plane, as illustrated in Fig. 5. We then show how to get the confidence score of two SSMs and sample a new state based on the two SSMs.

5.1. SSM definition

For $m = 1$, our method uses the 2D image coordinate with the mapping function $F_{m=1}$:

$$\mathbf{X}_t \sim F_{m=1}(\mathbf{X}_t), \tag{17}$$

where $F_{m=1}(\mathbf{X}_t)$ is the identity function. For $m = 2$, our method uses 2D ground plane coordinates in a static ground plane coordinate system:

$$\mathbf{X}_t^{\text{ground}} \sim F_{m=2}(\mathbf{X}_t). \tag{18}$$

Because $\mathbf{X}_t^{\text{ground}}$ is located on the ground plane, only the width of the bounding box can be measured. To recover the height, we assume a fixed aspect ratio.

The mapping function $F_{m=2}$ needs the location of the camera $\mathbf{M}_t$ and the ground plane orientation $\mathbf{N}_t$ as input. Both unknowns are estimated from the camera motion using the monocular visual odometry and ground plane estimation approach of Dragon and Van Gool [9]. This approach actually constructs the 3D space. For presentation, however, only two of three bases of the 3D space are illustrated in Fig. 5. The approach is robust and fairly generic, in that it only makes the mild assumption that the camera's ego motion $\mathbf{M}_t$ and the ground plane normal $\mathbf{N}_t$ are orthogonal over longer time spans. Its accuracy regarding $\mathbf{N}_t$ and $\mathbf{M}_t$ is approximately in the order of a few degrees, and one percent, respectively. Thus, our switching approach has to be robust wrt. perturbations of $\mathbf{X}_t^{\text{ground}}$ which we analyze further in Section 6.2. Notably any ground plane estimation method could be integrated into our framework (as long as monocular, as this is an important assumption we make in order to maximize usability).

5.2. Estimating confidence of SSM

We design the confidence $p(\mathbf{Y}_{1:\,t}|\mathbf{X}_t, m)$ in (12) according to the SSMs, as shown in Fig. 6. With $F_{m=1}(\cdot)$ in (17), the confidence for the image plane is then designed based on *appearance* as

$$p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m = 1) \equiv p(\mathbf{Y}_t | F_{m=1}(\mathbf{X}_t)), \tag{19}$$

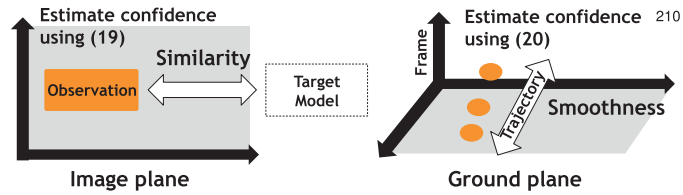


Fig. 6. Confidence estimation. If the SSM is the image plane, the confidence is degree of similarity between observation and reference appearance of the target. If the model is the ground plane, the confidence is degree of smoothness of the target trajectory over time.

where $p(\mathbf{Y}_t|\mathbf{X}_t)$ is the likelihood in (3). On the other hand, the confidence for the ground plane is designed based on *geometry*. Using $F_{m=2}(\cdot)$ from (18),

$$p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m = 2) = e^{-\lambda_2 \text{Smooth}(F_{m=2}(\mathbf{X}_t))}, \tag{20}$$

where λ_2 is a weighting parameter and Smooth returns the smoothness of the trajectory from the initial state to the current state. We measure the smoothness of the trajectory by calculating the standard deviation of the states in a trajectory.

5.3. Implementation details

When the image plane is selected as the working set, we propose both the center position (X_t^x, X_t^y) and the scale X_t^s on the image plane by using $\mathcal{N}(F_{m=1}(\mathbf{X}_{t-1}), \sigma_{m=1})$ in (15). When the ground plane is selected as the working set, however, we propose only the center position $(F_{m=2}(X_t^x), F_{m=2}(X_t^y))$ on the ground plane by using $\mathcal{N}(F_{m=2}(\mathbf{X}_{t-1}), \sigma_{m=2})$ in (15). For the scale, we use the image plane because we cannot describe the scale on the ground plane. When both planes are selected as the working set, the state proposed by the ground plane is backprojected to the image plane and is averaged with the state proposed by the image plane by using (16).

6. Experiments

The proposed method (TCS) was compared with 12 state-of-the-art tracking methods: MS [8], MC [21], IVT [34], OAB [12], SemiT [13], MIL [3], MTT [43], VTS [26], STRUCK [15], TLD [20], ASLA [19], and KCF [16]. In all experiments, λ_1 in (3) and λ_2 in (20) were set to 0.05. $\sigma_{m=1}$ in (15) was set to $(\sqrt{10}, \sqrt{5}, \sqrt{0.2})$ for the x, y center positions and the scale, respectively, while $\sigma_{m=2}$ was set to $(\sqrt{0.5}, \sqrt{0.01}, \sqrt{0.01})$. In (10), we set $\theta(w|w) = \theta(s|s)$ to $\frac{2}{3}$ and $\theta(s|w) = \theta(w|s)$ to $\frac{1}{3}$. We initialize $p(1, w|\mathbf{Y}_{1:t-1}) = p(1, s|\mathbf{Y}_{1:t-1}) = p(2, w|\mathbf{Y}_{1:t-1}) = p(2, s|\mathbf{Y}_{1:t-1}) = \frac{1}{4}$. *Notably, our method always use the same parameters throughout all experiments*. For the sampling-based methods, the same number of samples, $N = 500$, was used to track the target. To obtain the tracking results of IVT, OAB, SemiT, MIL, MTT, VTS, TLD, STRUCK and ASLA, we used the software provided by the authors. Our method takes 0.1 second per frame. *The result video is provided at* http://youtu.be/OK8WE1ez_u8.

6.1. Dataset

To evaluate the tracking performance, we used three datasets[1] with moving camera and moving objects.

Our main dataset is the KITTI dataset which is captured by driving around in rural areas and on highways. Usually, the road is visible and the ground plane can be estimated. For the ground

[1] Nine sequences from http://www.cvlibs.net/datasets/kitti/, four sequences from Dragon and Van Gool [9], and 29 sequences from Wu et al. [40].

Table 1
Tracking error when our method uses different planes for the state prediction. The numbers indicate average center location errors in pixels.

Seq.	1	2	3	4	5	6	7	8	9	10	11	12	13
Image plane	207	20	75	201	50	10	16	10	**4**	114	57	138	11
Ground plane	225	23	73	145	37	**9**	10	10	75	58	49	88	11
Image+Ground planes	98	21	35	87	15	**9**	**9**	9	21	61	49	63	12
Switched planes (ours)	**6**	**17**	**20**	**11**	**4**	26	16	**6**	7	**22**	**15**	**21**	**10**

plane estimation, however, our method requires no additional information and thus makes better uses of already-available data in the scene. The KITTI dataset is adopted by many other researchers for the stereo, optical flow, detection, and tracking evaluation and provides real-world benchmarks with following difficulties to the community. (1) Background clutter: The scenes contain many similar cars and pedestrians. In the scenes, trajectories of conventional tracking methods are easily hijacked by these similar objects (e.g. ASLA in *Seqs. 1–4*), (2) Appearance change: at the beginning of tracking a target is very small and becomes larger as time goes on. This is very challenging in visual tracking because a tracker starts to track the target without enough information on the target appearance.

Our sequences of the KITTI dataset is a little shorter than normal sequences used in other papers, because both a camera and a target in the KITTI dataset move fast and thus the target appears in the scene for a short time. Nevertheless the amount of failures of the compared approaches demonstrates that the dataset are challenging enough.

Eight sequences from Jia et al. [19] include the general scenes where no ground plane is visible.

6.2. Analysis of the proposed method

Advantage of our SSM switching process: To investigate the advantage of our SSM switching process, we made three trackers that track the target using the prediction from the image plane, the prediction from the ground plane, and the averaged prediction from image and ground planes, respectively, while other parts of the trackers stay the same. As shown in Table 1, our proposed SSM switching method produces the best tracking results: The SSM switching process helps a tracker to greatly improve the tracking performance. The tracker that only uses the image plane for prediction easily fails to track the target when there are severe occlusions. The tracker that only uses the ground plane for prediction robustly tracks the target in spite of occlusions but it misses the target when the ground plane is wrongly estimated by a single camera. The tracker that averages two planes for prediction is less sensitive to the noise, which is caused by one erroneous plane. Although a state is wrongly predicted by a plane, for example, the tracker can correct the state by averaging it with a more accurate prediction from the other plane. If the predicted states are very different from each other, however, averaging them cannot make a good single state. Hence our SSM switching method either chooses a accurate one between two states when two states are very different, or uses both states when they agree each other.

There are some failure cases. In Seqs. 6 and 7, ground plane estimates are severely inaccurate. Then the proposed method generates large center location errors in compared to other three baselines.

Fig. 7 exemplarily illustrates which plane our method employs at each frame. Our method typically chooses the image plane when the target is not occluded by the others and a clear observation is available. On the other hand, our method uses the ground plane to track the target when the target is severely occluded. This experiment demonstrates that our method is doing well in finding a

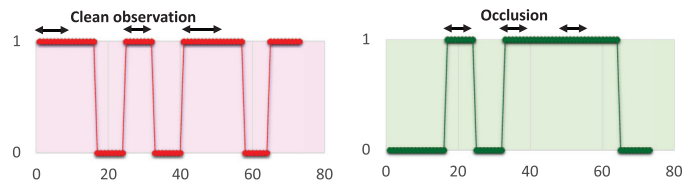
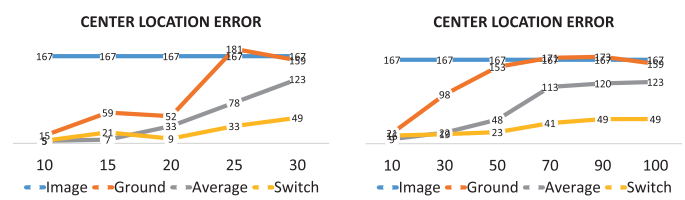


Fig. 7. Selected planes (image plane: left, ground plane: right) by our method at each frame in the *Seq. 4*. The X axis is a frame index. The Y axis has a 1 value when the plane is used at the frame.



(a) CLE vs. Degree of noise (b) CLE vs. Percentage of noise frames

Fig. 8. Center location error (CLE) when using an erroneous ground plane estimate in the *Seq. 8*. (a) 10%–30% uniformly-distributed noise on all frames. (b) 30% uniformly-distributed noise on 10%–100% of the frames.

good plane for visual tracking and appropriately switches between state planes. At some frames, our method uses both the image and ground planes and improves the tracking performance by combining tracking results from both.

Robustness toward imprecise ground plane estimation $\mathbf{N}_t$: In this experiment, we analyze the effect of imprecise estimation of $\mathbf{N}_t$. We make use of ground truth visual odometry and ground plane orientation provided in the KITTI dataset. The components of $\mathbf{N}_t$ are perturbed by uniformly-distributed noise in a varying range. We analyze the influence in terms of center location error if all frames are perturbed, and if only a fraction is perturbed, respectively (Fig. 8).

As it can be seen, our method outperforms the pure image-based approach by far and is still robust when the more perturbation is added and the more frames are perturbed. As a model violation (e.g. a person is not walking on the ground plane, but going up stairs) would result in a similar error, we may also state that our state switching raises the robustness toward ground plane motion violation.

Note that our method works in all datasets and for diverse camera motion. In the dataset used in Wu et al. [40], the ground plane may be not visible, temporarily occluded or imprecisely estimated. In this case, our method automatically chooses the image plane for visual tracking and performs as good as our baseline tracker, ASLA. For example, as visible in Table 1, row “image plane”, our method shows similar tracking performance as ASLA (Table 2, column “ASLA”), even when our method deliberately uses the image plane although the ground plane is visible. In addition, as demonstrated in Fig. 9, our method also works as good as ASLA when our method is directly evaluated with the general dataset used in Wu et al. [40].

Table 2
Comparison of tracking results using the center location error. The numbers indicate average center location errors in pixels. Red is the best result and bold black is the second-best result. N/A means that a tracking result is not available because a tracking method does not work in a sequence.

Seq.	MS	MC	IVT	OAB	SemiT	MIL	MTT	VTS	TLD	STRUCK	ASLA	KCF	TCS
1	276	301	319	281	80	292	222	249	**14**	181	203	323	**6**
2	99	107	151	59	**19**	52	47	39	36	35	28	52	**17**
3	98	111	108	71	66	100	65	76	**25**	130	71	87	**20**
4	145	250	188	208	**15**	283	263	247	121	246	281	202	**11**
5	162	178	149	65	120	89	54	87	N/A	**48**	57	64	**4**
6	100	99	80	105	139	19	144	74	**10**	126	**9**	40	26
7	89	85	117	31	30	55	51	28	N/A	**23**	27	42	**16**
8	56	57	58	56	26	37	24	127	96	57	**10**	17	**6**
9	30	34	12	28	56	16	13	20	N/A	8	**7**	16	**7**
10	53	37	176	18	155	80	77	**16**	**11**	83	92	20	22
11	91	95	81	58	**28**	56	63	59	50	62	56	131	**15**
12	90	132	200	103	180	78	125	77	**28**	106	182	81	**21**
13	87	101	52	96	18	24	82	53	120	18	**16**	48	**10**
Average	105.8	122.1	130.1	90.7	71.7	90.8	94.6	88.6	**51.1**	86.4	79.9	86.4	**13.9**

Table 3
Comparison of tracking results using the success rate. The numbers indicate the percentage of successfully tracked frames (score > 0.5), where the score is defined by the overlap ratio between the predicted bounding box B_p and the ground truth bounding box B_{gt}: $\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$.

Seq.	MS	MC	IVT	OAB	SemiT	MIL	MTT	VTS	TLD	STRUCK	ASLA	KCF	TCS
1	4	2	3	33	18	24	55	24	47	44	**57**	25	**95**
2	8	5	7	40	41	42	42	48	54	33	**81**	45	**81**
3	9	7	10	8	6	8	8	24	38	8	**46**	33	**55**
4	10	7	8	**12**	2	10	5	4	5	10	4	14	**56**
5	2	3	3	1	9	5	29	14	N/A	**44**	32	55	**84**
6	1	1	2	40	14	1	4	5	11	1	**57**	5	**57**
7	1	1	1	7	2	2	6	2	N/A	6	10	**45**	**18**
8	3	2	4	1	3	14	18	16	18	18	**78**	47	**87**
9	13	12	22	6	6	32	40	26	N/A	36	**54**	45	**73**
10	12	19	12	**44**	18	28	28	40	**74**	18	22	40	34
11	10	7	12	44	36	45	51	51	56	47	**60**	43	**57**
12	2	2	2	2	1	2	2	2	18	2	2	**40**	**26**
13	2	1	1	11	21	25	84	44	15	25	**68**	56	**84**
Average	5.9	5.3	6.7	19.2	13.6	18.3	28.6	23.1	33.6	22.4	**43.9**	37.9	**62.1**

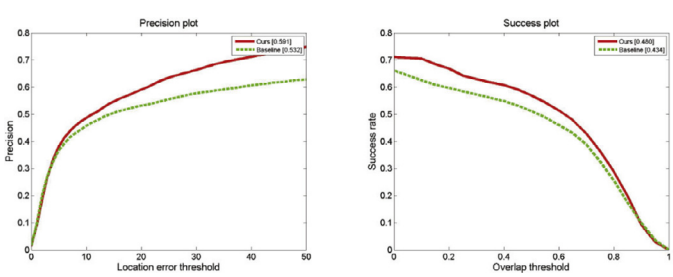


Fig. 9. Tracking error when ground planes are not visible in general test sequences [40]. The evaluation metrics are the precision plot and the success plot introduced in Wu et al. [40].

6.3. Comparison with other methods

Quantitative comparison: Table 2 shows quantitative results measured by the center location error. The smaller the number the better the tracking performance. In most sequences, our method shows the best tracking performance. SemiT, TLD, STRUCK, and ASLA shows the second best tracking performance. Although these tracking methods are shown to produce very accurate tracking results, they frequently fail to track the target in our sequences because the sequences include complex camera motions, small targets, and severe occlusions. Our tracker outperforms the second-best tracker with a 40 pixels lower center location error.

Table 3 shows quantitative results measured by success rates, where a higher number means better tracking performance. Our method produces accurate tracking results in all sequences and outperforms the second-best tracker with a 20% points higher success rate. Because OAB, SemiT, and STRUCK cannot handle scale changes of the target, the tracking performance in terms of the success rate is worse than the performance in terms of the center location error. On the other hand, ASLA accurately captures the scale changes of the target and shows the second best tracking performance. The tracking results in terms of both the center location error and success rate consistently demonstrate that our method robustly deals with severe occlusions and complex camera motions and outperforms the conventional tracking methods.

Qualitative comparison: In Figs. 10 and 11, we qualitatively compare our method with the other tracking methods. The orange, blue, white, pink, and yellow boxes represent the tracking results of our method (TCS), ASLA, TLD, STRUCK, and MTT, respectively. Red and green arrows describe the image and ground planes used at a frame, respectively. As shown in Fig. 10(a)–(i), our method successfully tracks the target in spite of severe occlusions. Other tracking methods miss the target because the appearances of occluders at many frames are very similar to that of the target, whereas our method overcomes this problem by using motion smoothness on the ground plane. When the target is occluded by other objects, our method adaptively changes the SSMs and uses the ground plane for visual tracking.

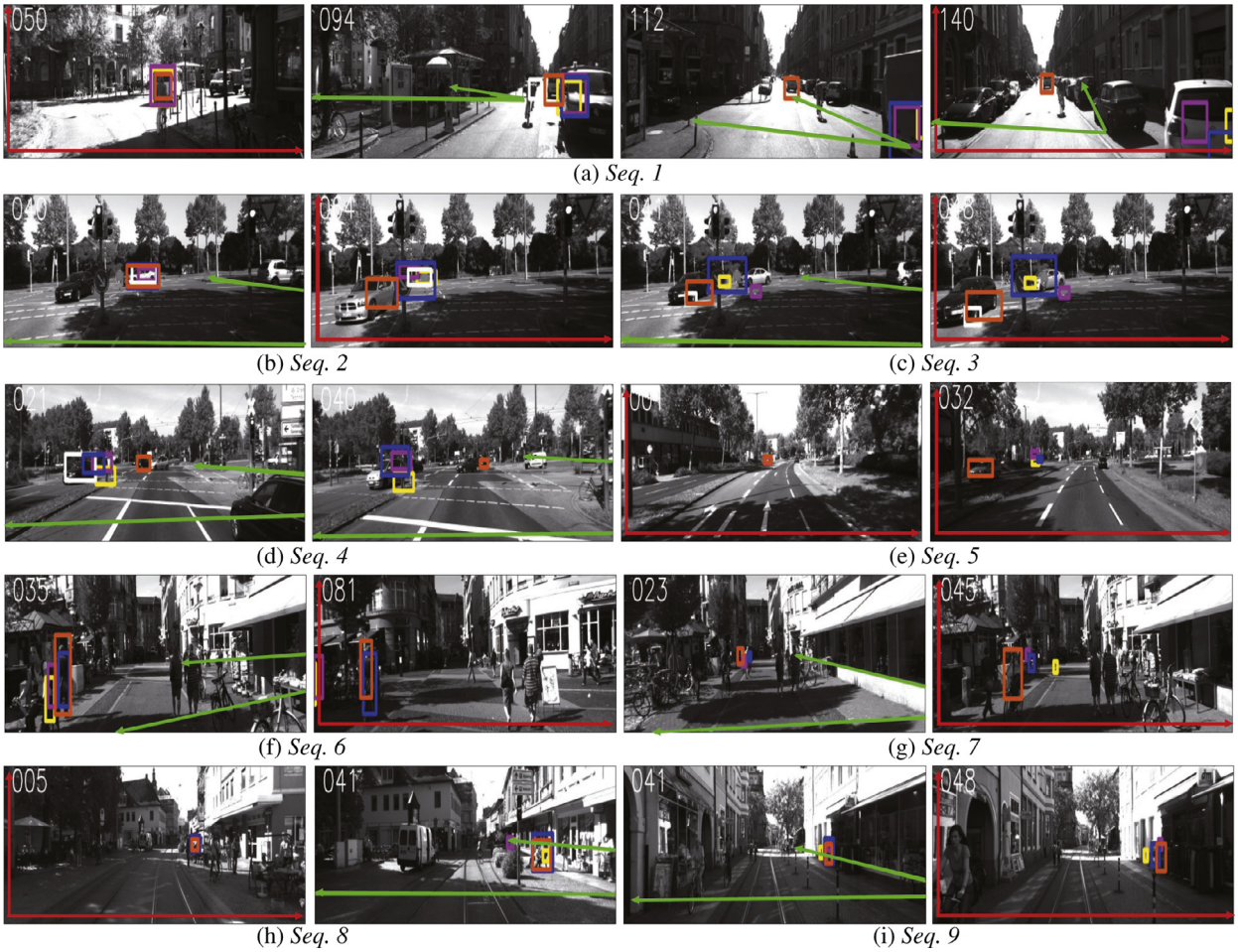


Fig. 10. Qualitative comparison of tracking results using KITTI dataset sequences. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

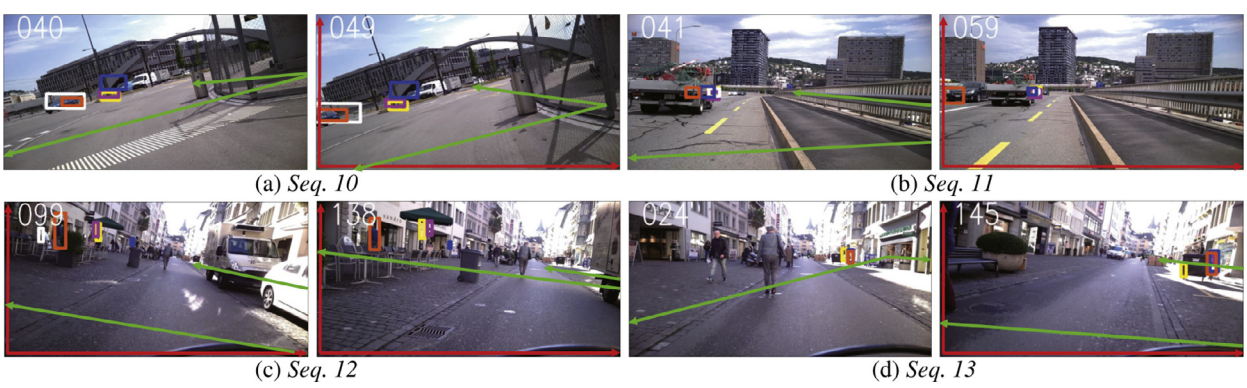


Fig. 11. Qualitative comparison of tracking results using the dataset from Dragon and Van Gool [9]. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

As shown in Fig. 10(a), our method robustly tracks the target although there are non-linear camera motions. Conventional tracking methods at frame #94 in *Seq. 1* suffer from severe changes in the camera pose, wherein the combination of target and camera motions produces very complex trajectories in the videos. However, our method robustly tracks the target because the camera motions are removed and only target motions are considered on the ground plane.

As shown in Fig. 10(e), (g)–(i), our method accurately tracks the target in spite of large scale changes in *Seqs. 5*, *7*, *8*, and *9*. In these sequences, the size of the target significantly increases as time goes on. If the initial size of the target is very small, a target model cannot include sufficient information about the target appearance, which makes visual tracking very difficult. Nevertheless, our method does not miss the target by the help of target motion clues obtained from the ground plane. On the other hand, trajectories of other trackers easily drift into background because the trackers highly depend on the initial appearance of the target.

In Fig. 10(a), 11(a), (c), and (d), our method averages state predictions predicted by the image and ground planes for visual tracking. We analyze this situation in Section 6.2.

7. Derivation

7.1. Derivation of (12)

By using Bayes rule,

$$\begin{aligned} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m) &= \frac{p(\mathbf{Y}_{1:t}|m)}{p(\mathbf{Y}_{1:t-1})} \propto p(\mathbf{Y}_{1:t}|m) \\ &= \int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m)p(\mathbf{X}_t|m)d\mathbf{X}_t \\ &\propto \int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m)d\mathbf{X}_t, \end{aligned}$$

where $p(\mathbf{X}_t|m)$ has a uniform distribution.

7.2. Derivation of (13)

Because $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}) = \frac{\sum_{i\in\mathbf{W}_t} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m=i)p(m=i|\mathbf{Y}_{1:t-1})}{p(\mathbf{W}_t|\mathbf{Y}_{1:t-1})}$ by substituting $\mathbf{Y}_t$ for $\mathbf{X}_t$ in (5), α_t in (9) then becomes

$$\begin{aligned} \alpha_t &= \frac{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)p(\mathbf{W}_t|\mathbf{Y}_{1:t-1})}{\sum_{i\in\mathbf{W}_t} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m=i)p(m=i|\mathbf{Y}_{1:t-1})} \\ &= \frac{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)}{\sum_{i\in\mathbf{W}_t} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m=i)\frac{p(m=i|\mathbf{Y}_{1:t-1})}{\sum_{j\in\mathbf{W}_t} p(m=j|\mathbf{Y}_{1:t-1})}}, \end{aligned}$$

where $p(\mathbf{W}_t|\mathbf{Y}_{1:t-1}) = \sum_{j\in\mathbf{W}_t} p(m=j|\mathbf{Y}_{1:t-1})$. By using the Markov Chain in (10), we implement α_t, as follows:

$$\alpha_t = \frac{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)}{\sum_{i\in\mathbf{W}_t} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m=i)\frac{p(m=i,w|\mathbf{Y}_{1:t-1})}{\sum_{j\in\mathbf{W}_t} p(m=j,w|\mathbf{Y}_{1:t-1})}}.$$

By substituting (12) for $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, m)$, α_t is obtained by

$$\alpha_t \propto \frac{\int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m)d\mathbf{X}_t}{\sum_{i\in\mathbf{W}_t} \int p(\mathbf{Y}_{1:t}|\mathbf{X}_t, m=i)d\mathbf{X}_t \frac{p(m=i,w|\mathbf{Y}_{1:t-1})}{\sum_{j\in\mathbf{W}_t} p(m=j,w|\mathbf{Y}_{1:t-1})}}.$$

8. Conclusion

In this paper, we propose an efficient tracking framework that adaptively switches the SSMs over time. By this, our method takes advantages of all SSMs and robustly tracks the target although videos include complex camera motions and severe occlusions. Experimental results demonstrate that our method outperforms the recent state-of-the-art tracking methods.

Our method can be naturally extended to the multiple target tracking problem. In this problem, each target uses a different SSM according to the tracking environment of each target. There are two challenges. First, a whole state space for all targets significantly increases by multiple SSMs. Second, the interaction between targets is difficult because of using a different SSM for each target. To overcome these challenges, our method can employ more advanced sampling techniques in Andrieu et al. [2] to reduce the state space size. Our method can define a reference state space for the interaction of targets, after each target is tracked independently by using a different SSM.

Acknowledgment

This work was supported by the ERC Advanced Grant VarCity (#273940).

References

[1] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, CVPR, 2006.
[2] C. Andrieu, A. Doucet, R. Holenstein, Particle Markov chain Monte Carlo methods, J. R. Stat. Soc. B 72 (3) (2010) 269342.
[3] B. Babenko, M. Yang, S. Belongie, Visual tracking with online multiple instance learning, CVPR, 2009.
[4] J. Berclaz, F. Fleuret, P. Fua, Robust people tracking with global trajectory optimization, CVPR, 2006.
[5] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, Multiple object tracking using k-shortest paths optimization, PAMI 33 (9) (2011) 1806–1819.
[6] L. Cehovin, M. Kristan, A. Leonardis, An adaptive coupled-layer visual model for robust visual tracking, ICCV, 2011.
[7] R.T. Collins, Y. Liu, M. Leordeanu, Online selection of discriminative tracking features, PAMI 27 (10) (2005) 1631–1643.
[8] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, CVPR, 2000.
[9] R. Dragon, L. Van Gool, Ground plane estimation using a hidden Markov model, in: CVPR, 2014.
[10] W. Du, J. Piater, Multi-camera people tracking by collaborative particle filters and principal axis-based integration, in: ACCV, 2007.
[11] M. Godec, P.M. Roth, H. Bischof, Hough-based tracking of non-rigid objects, ICCV, 2011.
[12] H. Grabner, M. Grabner, H. Bischof, Real-time tracking via on-line boosting, BMVC, 2006.
[13] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, ECCV, 2008.
[14] B. Han, L. Davis, On-line density-based appearance modeling for object tracking, ICCV, 2005.
[15] S. Hare, A. Saffari, P.H.S. Torr, Struck: structured output tracking with kernels, ICCV, 2011.
[16] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, PAMI 37 (3) (2015) 583–596.
[17] P. Hlinomaz, L. Hong, A multi-rate multiple model track-before-detect particle filter, Math. Comput. Model. 49 (12) (2009) 146162.
[18] M. Isard, A. Blake, Icondensation: unifying low-level and high-level tracking in a stochastic framework, ECCV, 1998.
[19] X. Jia, H. Lu, M.-H. Yang, Visual tracking via adaptive structural local sparse appearance model, CVPR, 2012.
[20] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, PAMI 34 (7) (2012) 1409–1422.
[21] Z. Khan, T. Balch, F. Dellaert, MCMC-based particle filtering for tracking a variable number of interacting targets, PAMI 27 (11) (2005) 1805–1918.
[22] K. Kim, L.S. Davis, Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering, ECCV, 2006.
[23] W.M. Koolen, D. Adamskiy, M.K. Warmuth, Putting bayes to sleep, NIPS, 2012.
[24] C. Kreucher, A.O.H. III, K. Kastella, Multiple model particle filtering for multi–target tracking, in: ASAP, 2004.
[25] S. Kwak, W. Nam, B. Han, J.H. Han, Learning occlusion with likelihoods for visual tracking, ICCV, 2011.
[26] J. Kwon, K.M. Lee, Tracking by sampling trackers, ICCV, 2011.
[27] X. Li, C. Shen, A. Dick, A. van den Hengel, Learning compact binary codes for visual tracking, CVPR, 2013.
[28] L. Liang-qun, X. Wei-xin, H. Jing-xiong, Multiple model Rao–Blackwellized particle filter, in: ICSP, 2008.
[29] X. Mei, H. Ling, Robust visual tracking using l1 minimization, ICCV, 2009.
[30] K.P. Murphy, Switching kalman filters, in: Technical Report, 1998.
[31] S. Oron, A. Bar-Hillel, D. Levi, S. Avidan, Locally orderless tracking, CVPR, 2012.
[32] P. Peursum, S. Venkatesh, G. West, Observations witching linear dynamic systems for tracking humans through unexpected partial occlusions by scene objects, in: ICPR, 2006.
[33] H. Possegger, S. Sternig, T. Mauthner, P.M. Roth, H. Bischof, Robust real-time tracking of multiple objects by volumetric mass densities, CVPR, 2013.
[34] D.A. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, IJCV 77 (1) (2008) 125–141.
[35] J. Santner, C. Leistner, A. Saffari, T. Pock, H. Bischof, Prost: parallel robust online simple tracking, CVPR, 2010.
[36] L. Sevilla-Lara, E. Learned-Miller., Distribution fields for tracking, CVPR, 2012.
[37] S. Sternig, T. Mauthner, A. Irschara, P.M. Roth, H. Bischof, Multi-camera multi--object tracking by robust hough-based homography projections, ICCV Workshop, 2011.
[38] P. Wang, Q. Ji, Multi-view face tracking with factorial and switching hmm, WACV, 2005.
[39] Y. Wu, T.S. Huang, A co-inference approach to robust visual tracking, ICCV, 2001.
[40] Y. Wu, J. Lim, M.-H. Yang, Online object tracking: a benchmark, CVPR, 2013.
[41] X. Xu, B. Li, Adaptive raoblackwellized particle filter and its evaluation for tracking in surveillance, TIP 16 (3) (2007) 838–849.
[42] W. Zajdel, A. Cemgil, B. Krose, Online multicamera tracking with a switching state-space model, ICPR, 2004.
[43] T. Zhang, B. Ghanem, S. Liu, N. Ahuja, Robust visual tracking via multi-task sparse learning, CVPR, 2012.