Vehicle Trajectory Estimation Based on Monocular Vision

Daniel Ponsa and Antonio López

Centre de Visió per Computador, Universitat Autònoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{daniel,antonio}@cvc.uab.es
www.cvc.uab.es/adas

Abstract. This paper proposes a system to estimate the 3D position and velocity of vehicles, from images acquired with a monocular camera. Given image regions where vehicles are detected, Gaussian distributions are estimated detailing the most probable 3D road regions where vehicles lay. This is done by combining an assumed image formation model with the Unscented Transform mechanism. These distributions are then fed into a Multiple Hypothesis Tracking algorithm, which constructs trajectories coherent with an assumed model of dynamics. This algorithm not only characterizes the dynamics of detected vehicles, but also discards false detections, as they do not find spatio-temporal support. The proposals is tested in synthetic sequences, evaluating how noisy observations and miss-detections affect the accuracy of recovered trajectories.

1 Introduction

The research in Computer Vision applied to intelligent transportation systems is mainly devoted to provide them with situational awareness, either to ascertain the state of their driver and passengers or to characterize its external surroundings. Some of the applications required by the automobile industry [1] (automatic cruise control, autonomous stop & go driving, lane change assistance, etc.), rely on determining accurately the position and velocity of other vehicles on the road ahead. First approaches to this problem have been based on active sensors as radar or lidar. However, the research on camera–based solutions has gained popularity, as vision sensors (CCD/CMOS) are passive and cheaper, and provide a richer description of the acquired environment. Different vision-based methods have been proposed to detect vehicles [2], some of them identifying their frontal or rear view on images acquired by a single camera [3,4,5]. These proposals detect vehicles in a frame-by-frame basis, inferring in some cases their position on the road from the 2D image regions where they are detected. From the point of view of a real application, besides the 3D vehicle location, it is very important to estimate its 3D velocity. This allows to predict the location of vehicles in the future, and in that way warning about problematic situations in advance. Thus, the topic of this paper is a method to provide this required dynamics information, based on properly integrating vehicle detections along time. The task

J. Martí et al. (Eds.): IbPRIA 2007, Part I, LNCS 4477, pp. 587–594, 2007.

performed corresponds to the target state initialization that precedes a target tracking process (figure 1). Given detections provided by the vehicle detection module, the objective is to check their spatio-temporal coherence and estimate the vehicle initial state (road position and velocity). Once this state is computed, then it can be efficiently updated along time by a target tracking module, which thanks to assumed system and observation models, localizes vehicles in frames more precisely and efficiently than the target detector can do.

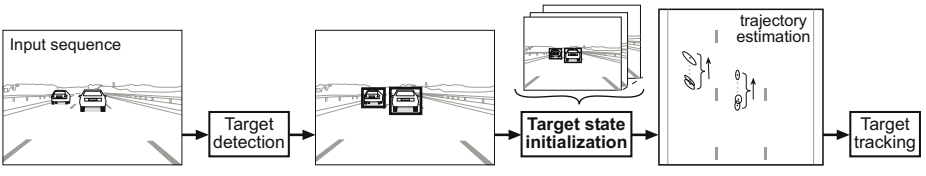

Fig. 1. The method proposed fills the gap between vehicle detection and tracking

The paper proposes a method to construct trajectories from detections, which is useful also to reject spurious false alarms. Indeed, most of false detections are caused by image regions whose appearance is momentarily similar to the one of vehicles. When due to the own vehicle movement the scene is captured from a different viewpoint, these regions are no longer identified as vehicles. In these cases is not possible to construct a coherent trajectory from them (according to expected vehicle dynamics), and this is used to discard them.

The structure of this paper is as follows. Next section details a new proposal to infer 3D vehicle coordinates from the 2D image regions where they are detected. Provided that some assumptions hold, the method combines projective geometry equations with the Unscented Transform to estimate with a Gaussian distribution the road region more likely to contain the detected vehicle. Section 3 describes an algorithm to integrate along time these estimated distributions, building trajectories that corroborate the presence of vehicles and estimate their dynamics. Section 4 evaluates the proposal using synthetic sequences, checking its performance against noisy vehicle detections, and miss-detections. Finally, the paper ends drawing some conclusions.

2 From 2D Detection to 3D Location Hypothesis

Algorithms to detect objects on single frames consist usually on evaluating a similarity criterion at different rectangular image regions, looking for the object of interest. These regions can be established by a deterministic scanning procedure [5], or by a rough preprocessing step that selects a set of candidate regions to be evaluated [4]. Commonly, real targets are detected in several overlapping regions. In order to provide a single detection per target, a clustering process groups neighboring detections, and returns a list of final detections. Each detection is specified by the parameters of a 2D region: its bottom-left corner

(x^r, y^r), and its width and height (w^r, h^r). Each of these parameters have some uncertainty, due to the inaccuracy of the detection/clustering process, which is commonly modeled as a Gaussian perturbation. Thus, for each detection a Gaussian distribution $\mathcal{N}(\mu^r, \Sigma^r)$ is given, where μ^r details the 2D region parameters, and Σ^r their uncertainty. To extract the 3D vehicle location from this observation, $\mathcal{N}(\mu^r, \Sigma^r)$ has to be retroprojected onto the 3D road coordinates. In general, this retroprojection can not be solved, but by assuming that the road conforms to a flat surface (which is realistic for highways and A roads), and that the camera position with respect to it is known, this is feasible.

2.1 Camera Model

The projective geometry of the acquisition system has been modeled in the following way. Image formation is represented using a pin–hole camera model with zero-skew [6], with effective focals (f^x, f^y) and center of projection (x^0, y^0). This camera is mounted on a *host* vehicle, facing the road ahead at a given height h over the ground, inclined slightly a pitch angle θ (figure 2). Using this model, 3D road points $(x, 0, z)$ are related with their image projection (x^r, y^r) by

$$x = \frac{f^y\, h\, \left(x^0 - x^r\right)}{f^x\, \left(\left(y^0 - y^r\right)\, \cos(\theta) + f^y\, \sin(\theta)\right)} \quad , \tag{1}$$

$$z = -\left(\frac{h\, \left(f^y\, \cos(\theta) - \left(y^0 - y^r\right)\, \sin(\theta)\right)}{\left(y^0 - y^r\right)\, \cos(\theta) + f^y\, \sin(\theta)}\right) \quad . \tag{2}$$

These expressions require knowing the camera extrinsic parameters (h, θ), which unfortunately can vary in every frame due to the action of the host vehicle suspension system. The variation of h can be ignored, as provokes a negligible error on x and z estimations, but a correct θ value is essential to estimate them accurately. θ could be estimated if the width w of the observed vehicle and its rotation φ relative to the Y–axe (Yaw angle) were known. At ground level, $w' = w\cos(\varphi)$ with commonly $\varphi \approx 0$, and w' projects onto w^r image pixels

$$w^r \sim \frac{f^x\, w'}{z\, \cos(\theta) - h\, \sin(\theta)} \quad . \tag{3}$$

Combining properly equations (2) and (3), it is found that

$$\theta = 2\arctan\left(\left(1 - \sqrt{1 + \left(\frac{y^0 - y^r}{f^y}\right)^2 - \left(\frac{w^r h}{w' f^x}\right)^2}\right) \Bigg/ \left(\frac{y^0 - y^r}{f^y} - \frac{w^r h}{w' f^x}\right)\right) \quad ,$$

and this makes feasible using equations (1) and (2) to hypothesize the 3D vehicle location. Unfortunately, to estimate θ the values of w and φ of observed vehicles are required, which are unknown a priori. However, relatively to the initial problem of unknowing θ, now there are some advantages. φ has a narrow range of feasible values, and its value is very close to zero in most of the situations. The range of feasible w values can also be very narrow, if for instance the vehicle detection process identifies the type of vehicle (bus, lorry, sedan, etc.). Thus, estimations more accurate than just by guessing directly θ can be obtained.

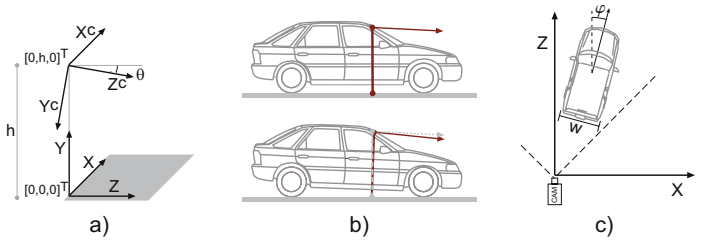

Fig. 2. a) Camera coordinate system relative to the road coordinate system. b) Extrinsic camera parameters vary when the vehicle suspension system actuates. c) Top view of the road coordinate system, detailing the vehicle orientation angle φ.

2.2 Retroprojecting a Gaussian Distribution

The vehicle detection procedure generates as result not simple 2D regions, but normal distributions $\mathcal{N}(\mu^r, \Sigma^r)$ of region parameters (x^r, y^r, w^r) (h^r is discarded because is irrelevant in the proposed retroprojection scheme). Estimating the distribution of 3D road points corresponding to a detection is not straightforward, basically because the retroprojection equations (1) and (2) are non–linear. This paper proposes computing a Gaussian approximation of this distribution by applying the Unscented Transform [7]. Given a Gaussian distribution to be transformed, a set of samples (sigma–points) are deterministically chosen which jointly match its mean and covariance. These samples are propagated using the non–linear equations, generating a cloud of transformed points. Computing the sample mean and covariance of these points (properly weighted) the Gaussian distribution fitting the transformed distribution is characterized (figure 3). The uncertainty of the final vehicle location distribution depends on the particular uncertainty in each detection parameter, being w^r the one who affects 3D locations the most.

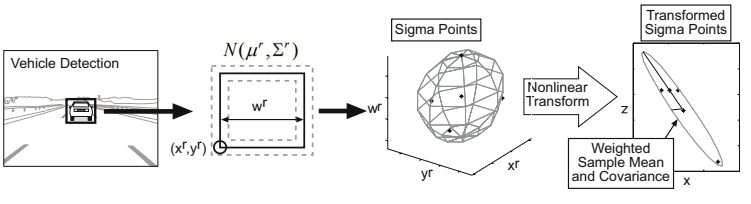

Fig. 3. Process of transforming 2D detections into 3D road coordinates

3 Detection Confirmation and Tracking

The procedure described up to this point analyses sequences in a frame-by-frame basis, and it can not provide information about the dynamics of detected vehicles. It performs like a sensor that perceives an image, and provides as output a noisy list of 3D location distributions where vehicles may lay. These distributions may

correspond to real vehicles or not, so the burden is also sort out false alarms from genuine detections. One possibility to deal with this problem is using Multiple Hypothesis Tracking (MHT) methods [8]. The idea is to connect consecutive detections along time, constructing trajectories or *tracks*. As new observations are collected, they may be integrated in existent trajectories, as well as start new ones. Trajectory construction is not always a trivial task, as several detections can verify a same trajectory at a given instant. One way to deal with this *data association problem* is considering all the feasible hypothesis, constructing a tree of trajectories along time. Trajectories started from false detections are commonly discarded in a few frames, since they will not receive more detections in the future. Trajectories coherent with a given number of observations confirm the presence of real vehicles, and then their dynamics can be estimated.

There exist many different MHT algorithms in the literature, which differ basically in how they deal with the data association problem. In general, the tree of trajectories can grow exponentially with the observations. If occasional miss-detections have also to be considered, the growth in the number of hypothesis is even more severe. So in order to fulfill computational requirements, *suboptimal* algorithms propose different strategies to consider only a (presumably good) subset of all possible trajectories. In the studied application, each observation can be generated only by a single vehicle, what reduces significantly the trajectories to be spreaded. Another important characteristics of vehicle detection is that 3D coordinates of observed vehicles are significantlly apart (they are at least separated by the width of one vehicle). These facts allow to pose a MHT with one of the simplest association methods: the Global Nearest Neighbor (GNN) criterion. It finds the best (most likely) assignment of input observations to existing trajectories, under the constrain that an observation can be associated with at most one track. For observations that remain unassigned, new trajectories are initiated. Due to paper size restrictions, is not possible to detail the algorithm implemented, based on the multiple target Kalman tracker in [9].

3.1 Trajectory Characterization

Trajectories managed in the MHT algorithm collect measurements until a given number N of them is reached. This task is done by a Kalman Filter (KF), estimating the state distribution of the vehicle at each instant t (i.e. $\mathbf{x}_t$, mantaining the vehicle coordinates on the road), using observations available up to t ($\mathbf{y}_{1:t}$). In formal terms, the MHT characterises $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. A maximal use of the N collected observations can be done, by using all of them to estimate the vehicle state at each instant t. Formally, this corresponds to estimate $p(\mathbf{x}_t|\mathbf{y}_{1:N})$ where $N \geq t$. In estimation theory terms, this procedure corresponds to the *smoothing* of the trajectory. The Kalman Smoother (KS) scheme implemented in this paper is based on the Rauch–Tung–Striebel fixed–interval optimal smoother [10]. This algorithm estimates a smoothed trajectory, from where the forward velocity v_t and orientation φ_t of the vehicle are computed. Figure 4 sketches the procedure carried out. The value of φ_t is constrained in the range $[-\pi/2, \pi/2]$, forcing in that way to maintain the vehicle movement direction in the sign of v_t.

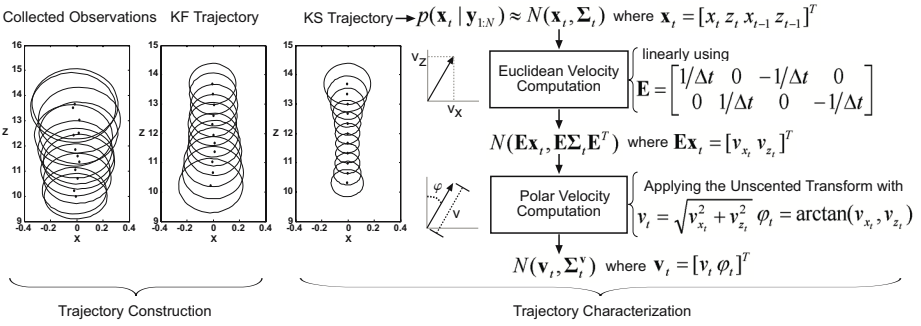

Fig. 4. Process to estimate the vehicle 3D velocity

4 Performance Evaluation

The proposal presented shows qualitatively a very good performance on real sequences, specially in discarding false vehicle detections. However, a quantitatively analysis of its operation is necessary to validate it objectively. Due to the lack of ground truth information in the available real sequences (i.e, knowledge about the real 3D location and velocity of vehicles), the system has been evaluated using synthetic data. Its accuracy depends on many factors (the 3D position of vehicles, their velocities, the noise perturbing detections, inaccurate assumptions, etc.). In this paper, an experiment has been designed to evaluate the system performance when assumptions hold, but vehicles are noisely detected and occasionaly missdetected. Synthetic vehicle trajectories have been simulated, and the parameters of their corresponding 2D image regions computed. An artificial variance has been given to these parameters, emulating the one of real detections. Three different situations have been analyzed, showing a vehicle at three different starting points, with a velocity relative to the host of 6 Km/h. The vehicle detection parameters (x^r, y^r, w^r) have been randomly distorted by different amounts of noise, corresponding to disturbances of the real vehicle position and width between $[-0.15, 0.15]$ meters. Missdetection events with probabilities between 0 and 40% have been considered. For each considered noisy situation, 100 different random sequences have been generated and processed, computing the disparity between the real and recovered trajectory, forward velocity, and orientation. The performance criterion used is the average of the Mean Square Error (MSE) in the 100 experiments. The presence of vehicles is confirmed, when their corresponding trajectories collect 12 observations. Figure 5 shows results obtained.

With respect to recover the vehicle trajectory, the system performs similarly wherever the trajectory starts. Accuracy depends critically on the presence of miss-detections along the trajectory. In the estimation of the vehicle velocity $[v, \varphi]^T$, better results are obtained for closer vehicles, as the uncertainty in the 3D road locations composing the trajectory is smaller. Concerning the x position of vehicles, centered ones provide a more accurate estimation of φ, while the

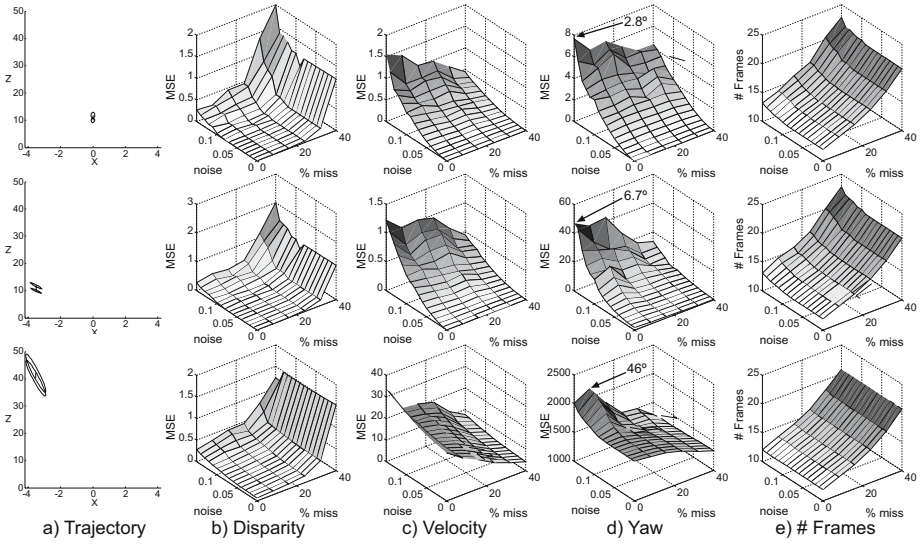

Fig. 5. a) Extremes of the trajectories considered. b–d) Respectively, MSE between real and recovered 3D trajectory locations, forward velocities (v), and orientations (φ). e) Average number of frames required to confirm a vehicle detection.

estimation of v is a little bit more imprecise. Note that the factor that affects more critically the estimation of vehicle dynamics is the noise perturbing observations, while surprisingly, miss-detections clearly favour its better estimation. This results from the fact that due to miss-detections, it takes a longer time to obtain the N observations that compose a *complete* trajectory, and this favours having observations more distant (in spatial terms) between each other. This attenuates the effect of the uncertainty of 3D locations in the velocity estimation (figure 6). The presence of false positives in the observations has also been evaluated. In each frame, false detections has been added, distributed randomly on the acquired road following a uniform distribution. Results obtained are extremely similar to the ones in figure 5, showing the great capacity of the proposal to discard uncorrelated false detections.

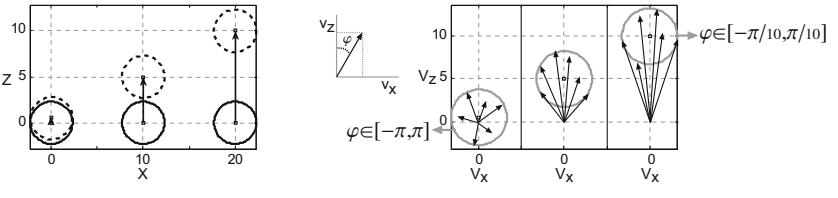

Fig. 6. Left) Consecutive observations of three trajectories. Right) Distribution of their corresponding velocity. The more distant observations are, the less uncertain φ is.

5 Conclusions

A system has been proposed to characterize the 3D position and velocity of vehicles, from 2D detections in images obtained from a camera on a mobile platform. Combinig the UT with projective geometry, vehicle detections are converted into Gaussian distributions detailing the road region where vehicles may lay. These distributions are fed into a MHT algorithm, which evaluates its spatio-temporal coherence in order estimate the 3D vehicle velocity and reject false detections. The performance of the proposal has been evaluated with synthetic sequences, in order to compare results obtained against ground truth data. Results show that miss-detections affect mostly the accuracy of recovered trajectories, while the precision on the 3D velocity estimated degrades with the noise in the observations. False detections are satisfactorily discarded.

Acknowledgments. This research has been partially funded by Spanish MEC project TRA2004-06702/AUT.

References

1. Dickmanns, E.: The development of machine vision for road vehicles in the last decade. In: Int. Symp. on Intelligent Vehicles, Versailles (2002)
2. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(5), 694–711 (2006)
3. Betke, M., Haritaoglu, E., Davis, L.: Realtime multiple vehicle detection and tracking from a moving vehicle. Machine Vision & Applications, (12), pp. 69–83 (2000)
4. Khammari, A., Lacroix, E., Nashashibi, F., Laurgeau, C.: Vehicle detection combining gradient analysis and adaboost classification. In: IEEE Conf. on Intelligent Transportation Systems, pp. 1084–1089 (2005)
5. Ponsa, D., López, A., Lumbreras, F., Serrat, J., Graf, T.: 3D vehicle sensor based on monocular vision. In: IEEE Conf. on Intelligent Transportation Systems, pp. 1096–1101 (2005)
6. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge, ISBN: 0521540518 (2004)
7. Julier, S.J.: The spherical simplex unscented transformation. In: Proceedings of the American Control Conference, Denver, Colorado, pp. 2430–2434 (2003)
8. Blackman, S.B.: Multiple hypothesis tracking for multiple target tracking. IEEE Aerospace and Electronic Systems Magazine 19(1), 5–18 (2004)
9. Davies, D., Palmer, P., Mirmehdi, M.: Detection and tracking of very small low constrast objects. In: British Machine Vision Conference, pp. 599–608 (1998)
10. Gelb, A., et al.: Applied Optimal Estimation. The MIT Press, Cambridge (1974)