# REAL-TIME 3D MODELING OF VEHICLES IN LOW-COST MONOCAMERA SYSTEMS

M. Nieto, L. Unzueta, A. Cortés, J. Barandiaran, O. Otaegui
*Vicomtech-IK4 Research Alliance, Donostia-San Sebastián, Spain*
{*mnieto, lunzueta, acortes, jbarandiaran, ootaegui*}*@vicomtech.org*

P. Sánchez
*IKUSI, Donostia-San Sebastián, Spain*
*pedro.sanchez@ikusi.com*

Keywords: Computer vision, Monocamera, Traffic flow surveillance, 3D modeling.

Abstract: A new method for 3D vehicle modeling in low-cost monocamera surveillance systems is introduced in this paper. The proposed algorithm aims to resolve the projective ambiguity of 2D image observations by means of the integration of temporal information and model priors within a Markov Chain Monte Carlo (MCMC) method. The method is specially designed to work in challenging scenarios, with noisy and blurred 2D observations, where traditional edge-fitting or feature-based methods fail. Tests have shown excellent estimation results for traffic-flow video surveillance applications, that can be used to classify vehicles according to their length, width and height.

## 1 INTRODUCTION

Counting vehicles is a need for shadow toll road operators, which are paid by governments according to the number of vehicles using the road. Besides, it is also typical to distinguish between the type of vehicles, e.g. heavy or light. For that purpose, vision-based traffic flow surveillance methods have become a major topic in the computer vision community due to the increasing demand of road operators for cost-effective applications.

Compared with other technologies such as radar, ILD (inductive loop detectors), or laser, computer vision can be used to obtain richer information, such as visual features of the vehicles (color, lights), apart from geometric information (vehicle volumes). Nevertheless, computer vision approaches in Intelligent Transportation Systems (ITS) can only compete with radar, ILD and other mature technologies by reducing its costs, and this is typically translated into low-quality cameras and HW with low processing capabilities. Therefore, although there are a huge number of works in the literature related to vehicle classification using computer vision, there is still a lack of solutions which offer a trade-off between accuracy and costs. We have found that the most sophisticated methods use high definition cameras, with no blurring effect and with clear edge information (Pang et al., 2007). Besides, they are typically devised for urban scenarios, where the reduced speed of the vehicles simplifies the classification problem (Buch et al., 2010). Some 3D classification methods have used vehicle models as prior information, such as wireframe fixed models (Haag and Nagel, 2000), which some authors parameterize with car manufacturers data (Buch et al., 2010). However, as a general criticism, in most situations, the fitting accuracy of these methods is much lower than the detail of the wireframe, making uneffective such complex vehicle models. For that reason, most works just assume some minimum and maximum values for the dimensions of the vehicles (Barder and Chateau, 2008).

In this paper we propose a novel method specially devised to classify vehicles according to estimates of their 3D volume in challenging scenarios (due to the low-cost adquisition systems, and the high speed of the vehicles monitorized in motorway scenes as those shown in the examples of Fig. 1). Namely, the main contributions of this work are: (i) a probabilistic dynamic framework that integrates noisy 2D silhouette observations and vehicle model priors; (ii) real-time performance by means of an efficient design of a maximum a posteriori (MAP) method to generate point-estimates of the target posterior distribution; and (iii) provided the calibration of the camera, the system efficiently estimates the lost dimension in the projective

Figure 1: Typical low-quality images of road scenes captured for video surveillance purposes.

process, and thus generates estimates of the dimensions of the vehicle.

## 2 APPROACH OVERVIEW

The target of the method is the estimation of the dimensions of vehicles, which are modeled as rectangular cuboids with width, height and length, in order to classify them as one of a set of predefined vehicle classes. The estimation is done for each time instant, $t$, based on the previous estimations and the new incoming image observations.

The method makes estimations of the posterior density function $p(\mathbf{x}_t|Z^t)$, given the complete set of observations at time $t$, $Z^t$, from which determine the most probable system state vector, $\mathbf{x}_t = (w_t, h_t, l_t)^\top$, which models the dimensions of the vehicle. Three main sources of information need to be available: the calibration of the camera (including intrinsic and extrinsic parameters, which can be done offline), 2D image observations of the projection of the volume onto the road, and prior knowledge of vehicle models. Therefore, the proposed method applies on any existing 2D detector, which can be pretty simple, for instance, in this work we have used a traditional background-foreground segmentation based on color and a blob tracking strategy (Kim et al., 2005).

Fig. 2 illustrates an example process that generates the required information.

The proposed solution is based on a Markov Chain Monte Carlo (MCMC) method, which models the problem as a dynamic system and naturally integrates the different types of information into a common mathematical framework. This method requires the definition of a sampling strategy, and the involved density functions (namely, the likelihood function and the prior models). Typically, the complexity of this kind of sampling strategies are too high to run in real time. For that reason we have designed our solution as a fast approximation to MCMC-based MAP methods using a low number of hypotheses. Next sections describe the details of all the abovementioned issues as well as a brief introduction to the MCMC-based methods.

## 3 MCMC FRAMEWORK

MCMC methods have been successfully applied to different nature tracking problems (Barder and Chateau, 2008; Khan et al., 2005). They can be used as a tool to obtain maximum a posteriori (MAP) estimates provided likelihood and prior models. Basically, MCMC methods define a Markov chain, $\{\mathbf{x}_t^i\}_{i=1}^{N_s}$, over the space of states, $\mathbf{x}$, such that the stationary distribution of the chain is equal to the target posterior distribution $p(\mathbf{x}_t|Z^t)$. A MAP, or point-estimate, of the posterior distribution can be then selected as any statistic of the sample set (e.g. sample mean or robust mean), or as the sample, $\mathbf{x}_t^i$, with highest $p(\mathbf{x}_t^i|Z^t)$, which will provide the MAP solution to the estimation problem.

Compared to other typical sampling strategies, like sequential-sampling particle filters (Isard and Blake, 1998), MCMC directly sample from the posterior distribution instead of the prior density, which might be not a good approximation to the optimal importance density, and thus avoid convergence problems (Arulampalam et al., 2002).

The analytical expression of the posterior density can be decomposed using the Bayes' rule as:

$$p(\mathbf{x}_t|Z^t) = kp(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|Z^{t-1}) \tag{1}$$

where $p(\mathbf{z}_t|\mathbf{x}_t)$ is the likelihood function that models how likely the measurement $\mathbf{z}_t$ would be observed given the system state vector $\mathbf{x}_t$, and $p(\mathbf{x}_t|Z^{t-1})$ is the prediction information, since it provides all the information we know about the current state before the new observation is available. The constant $k$ is a scale factor that ensures that the density integrates to one.

We can directly sample from the posterior distribution since we have its approximate analytic expression (Khan et al., 2005):

$$p(\mathbf{x}_t|Z^t) \propto p(\mathbf{z}_t|\mathbf{x}_t)\sum_{i=1}^{N_s} p(\mathbf{x}_t|\mathbf{x}_{t-1}^i) \tag{2}$$

For this purpose we need a sampling strategy, like the Metropolis-Hastings (MH) algorithm, which dramatically improves the performance of traditional particle filters based on importance sampling. As a summary, the MH generates a new sample according to an acceptance ratio, that can be written in our case as:

$$\alpha = \frac{p(\mathbf{x}_t^j|Z^t)}{p(\mathbf{x}_t^{j-1}|Z^t)}\frac{q(\mathbf{x}_t^{j-1}|\mathbf{x}_t^j)}{q(\mathbf{x}_t^j|\mathbf{x}_t^{j-1})} \tag{3}$$

where $j$ is the index of the samples of the current chain. The proposed sample $\mathbf{x}_t^j$ is accepted with probability $\min(\alpha, 1)$. If the sample is rejected, the current state is kept, i.e. $\mathbf{x}_t^j = \mathbf{x}_t^{j-1}$. The proposal density $q(\mathbf{x})$

(a) (b) (c) (d) (e) (f)

Figure 2: Pre-processing steps: (a) original image; (b) correction of lens distortion; (c) detection of orthogonal directions on the plane; (d) rectified road plane; (e) detected 2D blobs using background segmentation; (f) 3D models after applying the proposed method.

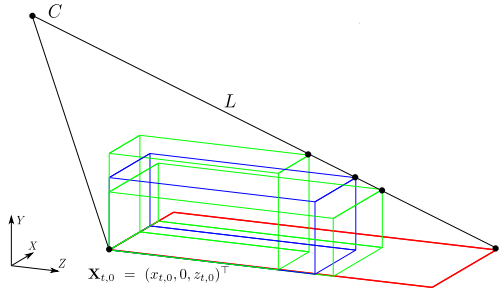


Figure 3: Projective ambiguity: a given 2D observation in the OXZ plane (in red) of a true 3D cuboid (blue) may also be the result of the projection of a family of cuboids (in green) with respect to camera C.

might be any function from which it is easy to draw samples. Typically it is chosen as a normal distribution, which is symmetric, i.e. $q(\mathbf{x}_t^{j-1}|\mathbf{x}_t^j) = q(\mathbf{x}_t^j|\mathbf{x}_t^{j-1})$ and thus the terms depending on the proposal can be removed from eq. 3.

Besides, it is a common practice to select a subset of samples from the chain to reduce their correlation and to discard a number of initial samples to reduce the influence of initialization. Therefore, to obtain $N_s$ effective samples of the chain it is required to generate a total number of samples $N = B + cN_s$, where $B$ is the number of initial samples, and $c$ is the number of samples discarded per valid sample.

## 4 LIKELIHOOD FUNCTION

For each image, the observation is the current 2D silhouette of the vehicle projected into the rectified image. Considering the cuboid-model of the vehicle, and that the yaw angle is approximatedly zero we can reproject a 3D ray from the far-most corner of the projected cuboid and the optical center.

There are infinite points on the ray that are projected in the same image point and therefore correspond to a solution to the parameters of the cuboid, as shown in Fig. 3. Nevertheless, there are a number of constraints that bound the solution to a segment of the ray: positive and minimum height, width and length.

Therefore, the likelihood function must be any function that fosters volume hypotheses near the reprojection ray. For the sake of simplicity, we choose a normal distribution on the point-line distance. The covariance of the distribution expresses our confidence about the measurement of the 2D silhouette and the calibration information. The likelihood function can be written as

$$p(\mathbf{z}_t|\mathbf{x}_t) \propto \exp\left((\mathbf{y}_t - \mathbf{x}_t)^\top S^{-1}(\mathbf{y}_t - \mathbf{x}_t)\right) \qquad (4)$$

where $\mathbf{x}_t$ is a volume hypothesis, and $\mathbf{y}_t$ is its projection onto the reprojection ray. The position of $\mathbf{y}_t$ can be computed from $\mathbf{x}_t$ as the intersection of the ray and a plane passing through $\mathbf{x}_t$ and whose normal vector is parallel to the ray. For this purpose we can represent the ray as a Plücker matrix $L_t = \mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top$, where $\mathbf{a}$ and $\mathbf{b}$ are two points of the line, e.g. the far-most point of the 2D silhouette, and the optical center, respectively. These two points are expressed in the $WHL$ coordinate system. Therefore, provided that we have the calibration of the camera, we need a reference point in the 2D silhouette. We have observed that the point with less distortion is typically the closest point of the quadrilateral to the optical center, whose coordinates are $\mathbf{X}_{t,0} = (x_{t,0}, 0, z_{t,0})^\top$ in the $XYZ$ world coordinate system. This way, any $XYZ$ point can be transformed into a $WHL$ point as $\mathbf{x}_t = R_0\mathbf{X}_t - \mathbf{X}_{t,0}$. Nevertheless, the relative rotation between these systems can be approximated to the identity, since the vehicles typically drive parallel to the $OZ$ axis.

The plane is defined as $\pi_t = (\mathbf{n}_t^\top, D_t)^\top$, where $\mathbf{n}_t = (n_x, n_y, n_z)^\top$ is the normal to the ray $L_t$, and $D_t = -\mathbf{n}_t^\top\mathbf{x}_t$. Therefore, the projection of the point on the ray can be computed as $\mathbf{y}_t = L_t\pi_t$.

## 5 PRIOR FUNCTIONS

The information about the volume of the vehicle can be encoded as the product of two functions, each one modeling two independent sources of information:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathcal{M}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathcal{M}) \qquad (5)$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ represents the dynamic model of the system. In our case, we will assume that a vehicle is

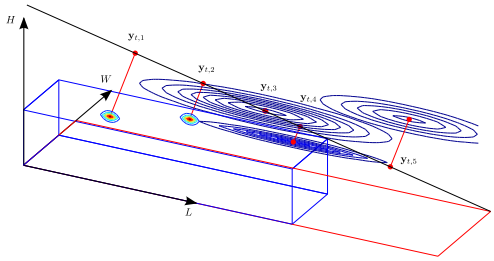

Figure 4: Projection of vehicle prior models into the ray $L$. For a better visualization, each $p(\mathbf{x}_t, \mathbf{x}_m)$ is shown as a point in $WHL$ and a contour slice parallel to $OHL$.

a non-deformable rigid object, such that it does not vary its dimensions through time, and thus

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \propto \exp\left((\mathbf{x}_t - \mathbf{x}_{t-1})^\top (\mathbf{x}_t - \mathbf{x}_{t-1})\right) \quad (6)$$

The second term of eq. (5), $p(\mathbf{x}_t|\mathcal{M})$, contains the information that we have about typical configurations of vehicle dimensions, i.e. typical proportions of vehicles according to a number of models, such as truck, motorcycle, car, etc. Let us represent this information as a set of clusters that can be parameterized as a mixture of normal distributions in the $WHL$ space: $\mathcal{M} = \{\mathbf{x}_m\}_{m=1}^M$. Therefore,

$$p(\mathbf{x}_t|\mathcal{M}) = \sum_{m=1}^{M} p(\mathbf{x}_t, \mathbf{x}_m) \quad (7)$$

where $\mathbf{x}_m = (W_m, H_m, L_m)^\top$ and

$$p(\mathbf{x}_t, \mathbf{x}_m) \propto \exp\left((\mathbf{x}_t - \mathbf{x}_m)^\top S_m^{-1} (\mathbf{x}_t - \mathbf{x}_m)\right) \quad (8)$$

and $S_m = diag\{\sigma_w^2, \sigma_h^2, \sigma_l^2\}$ is the covariance matrix of model $m$.

Table 1 exemplifies a set of vehicle models. The gaussian model ensures that the vehicle models are not rigid nor fixed, in contrast with typical wireframe models, and thus enhances the flexibility of prior information. For instance, trucks can be modeled as a 3D gaussian centered at $(2.0, 2.5, 7)$ with high variance values, since trucks may vary significantly in length or height.

## 6 ALGORITHM COMPLEXITY REDUCTION

Once we have defined the prior and observation models, the complete expression of the MH acceptance ratio is given by:

$$\alpha = \frac{p(\mathbf{z}_t|\mathbf{x}_t^j)}{p(\mathbf{z}_t|\mathbf{x}_t^{j-1})} \frac{\sum_{i=1}^{N_s} p(\mathbf{x}_t^j|\mathbf{x}_{t-1}^i)}{\sum_{i=1}^{N_s} p(\mathbf{x}_t^{j-1}|\mathbf{x}_{t-1}^i)} \frac{\sum_{m=1}^{M} p(\mathbf{x}_t^j, \mathbf{x}_m)}{\sum_{m=1}^{M} p(\mathbf{x}_t^{j-1}, \mathbf{x}_m)} \quad (9)$$

Table 1: Example configuration of vehicle models.

| Vehicle type | $W_m$ | $H_m$ | $L_m$ | $\sigma_w$ | $\sigma_h$ | $\sigma_l$ |
|---|---|---|---|---|---|---|
| Car | 1.6 | 1.5 | 4 | 0.1 | 0.1 | 0.2 |
| Motorbike | 1.6 | 1.5 | 2 | 0.1 | 0.1 | 0.2 |
| Truck | 2.0 | 2.5 | 7 | 0.2 | 0.3 | 1.0 |
| Trailer | 1.6 | 1.5 | 7 | 0.1 | 0.1 | 2.0 |
| Bus | 2.0 | 2.5 | 10 | 0.2 | 0.3 | 1.0 |

By drawing $N_s$ effective samples using the MH algorithm we have the approximation of the posterior distribution as in eq. (2). Hence, we can compute point-estimates of the state vector $\mathbf{x}_t$ and thus estimate the volume of the 3D cuboid at each time instant. For instance we can use the sample mean as the simplest statistic, which is valid enough since the posterior distribution can be assumed to be unimodal.

Nevertheless, the generation of the Markov chain implies a significant amount of computations, since the computational complexity is $O(NN_s)$. The reason is that for each proposed sample $\mathbf{x}_t^j$, the complete set of previous samples $\{\mathbf{x}_{t-1}^i\}_{i=1}^{N_s}$ has to be evaluated to compute the acceptance ratio.

To reduce to linear time operation, i.e. $O(N)$, we can instead select a single previous sample, $\mathbf{x}_{t-1}^*$, from the set. Khan et al. (Khan et al., 2005) propose to select a random sample from the set, although we have observed much better performance selecting the point-estimate of the previous time instant. The acceptance ratio expression is then simplified to:

$$\alpha = \frac{p(\mathbf{z}_t|\mathbf{x}_t^j)}{p(\mathbf{z}_t|\mathbf{x}_t^{j-1})} \frac{p(\mathbf{x}_t^j|\mathbf{x}_{t-1}^*)}{p(\mathbf{x}_t^{j-1}|\mathbf{x}_{t-1}^*)} \frac{\sum_{m=1}^{M} p(\mathbf{x}_t^j, \mathbf{x}_m)}{\sum_{m=1}^{M} p(\mathbf{x}_t^{j-1}, \mathbf{x}_m)} \quad (10)$$

Regarding the specific nature of our problem, an additional great reduction of the complexity of the sampling step can be achieved if we force the samples to belong to the ray defined by the likelihood model. This is equivalent to reduce the problem to a one-dimensional search on the ray. On the one hand, the proposal density can be now defined as a one-dimensional normal distribution that draw samples on the ray, as well as the dynamic model. Therefore, the samples are now drawn based on the simplified expression of the acceptance ratio:

$$\alpha = \frac{p(\mathbf{x}_t^j|\mathbf{x}_{t-1}^*)}{p(\mathbf{x}_t^{j-1}|\mathbf{x}_{t-1}^*)} \frac{\sum_{m=1}^{M} p(\mathbf{x}_t^j, \mathbf{x}_m)}{\sum_{m=1}^{M} p(\mathbf{x}_t^{j-1}, \mathbf{x}_m)} \quad (11)$$

subject to $\mathbf{x}_t^j \in L_t$.

The implementation of the algorithm can be as well simplified if the state space is reduced to a discrete number of states, namely $\{\mathbf{y}_m\}_{m=1}^M$, i.e. the projections of the vehicle models on the observed ray.

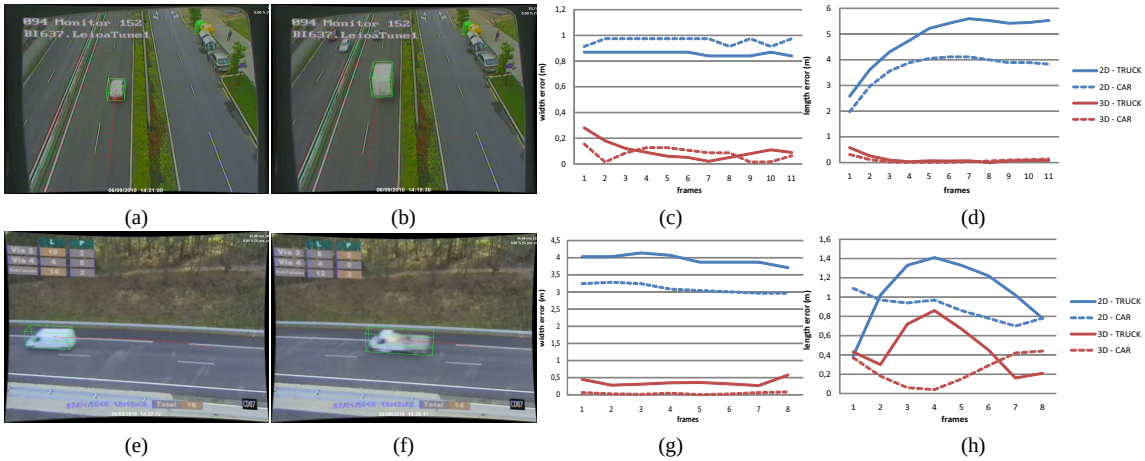


Figure 5: Examples of the error of the 2D and 3D methods for different perspectives and type of vehicles.

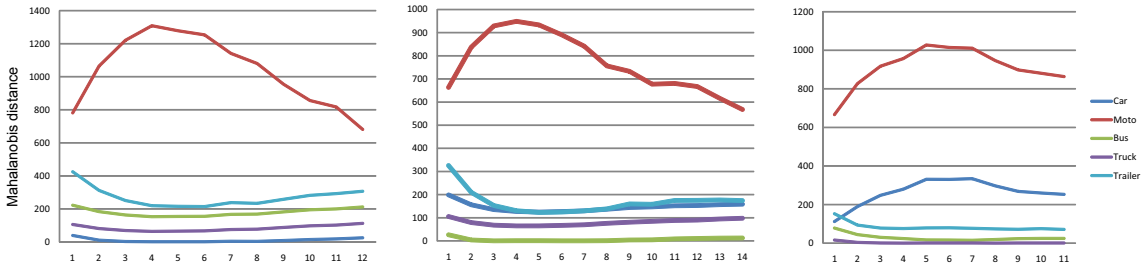


Figure 6: Mahalanobis distance for all the defined classes for three example sequences of a car, a bus and a truck.

Under this assumption, the algorithm computes the posterior probability of each $\mathbf{y}_{t,m} = L_t\pi$ as proportional to $p(\mathbf{y}_{t,m}|\mathbf{y}^*_{t-1})p(\mathbf{y}_{t,m},\mathbf{x}_m)$, and determines the MAP point-estimate of $p(\mathbf{x}_t|Z^t)$ as the most likely projection $\mathbf{y}_{t,m}$.

## 7 RESULTS

The proposed system overcomes the problems of 2D strategies that aim to measure the dimensions of the vehicles for classification purposes in perspective images. Fig. 5 shows some examples of the error committed by the proposed 3D estimation method and the base 2D estimation strategy when computing the width and length of a vehicle with known dimensions. As shown, the perspective distortion causes that the 2D strategies incurr in severe estimation errors. For instance, the images of the upper row of Fig. 5 depict a situation in which the perspective of the camera makes that 2D estimation of the length of the vehicle are greatly incorrect, while the estimation obtained by the proposed 3D module dramatically reduces the error. Analougously, the bottom row of Fig. 5 shows a case where the perspective affects mostly the 2D estimation of the width of vehicles, while the proposed method again achieves great reductions of measurement error. As a consequence, the proposed method helps to improve the reliability of a system that aims to classify vehicles according to their dimensions, which is in turn quite typical in tolling applications.

Finally, we exemplify the classification quality of our approach in Fig. 6, which corresponds to three example sequences of a car, a bus, and a truck (with typical dimensions). This figure shows the values of the Mahalanobis distance of each model $\mathbf{x}_m$ with respect to their projections into the ray $L$. The classification is correct as the “car”, “bus” and “truck” classes obtain that lowest error along their corresponding sequences. As far as the instantaneous estimations are coherent from one frame to another, the application of the motion prior strengthens the classification.

In order to evaluate the performance of the vehicle classification, we have tested the proposed solution for a set of videos of different roads and perspectives, with an aggregate duration of more than 5 hours. The total number of detected vehicles in the video sequence is 2551/2585 (98.7%). The target application required the classification of vehicles into two broad

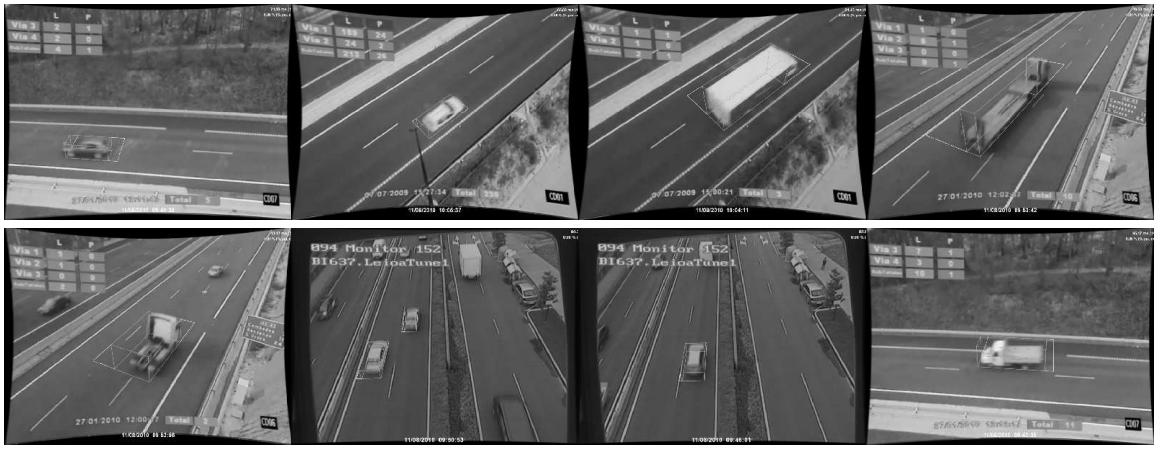


Figure 7: Example results of 3D vehicle modeling, including different size vehicles and type of perspectives.

categories: light and heavy. Considering the detected vehicles, the system correctly classified $2214/2248$ light vehicles, and $337/337$ heavy vehicles according to their volume. Some example images of the renderization of the estimated 3D model are shown in Fig. 7. As shown, in most situations, the cuboid fits correctly the volume occupied by the vehicles (with some unaccuracy due to insufficient perpsective distortion or excessively long vehicles), and thus allow to classify vehicles in the required categories.

## 8 CONCLUSIONS

This paper introduces a real-time method to augment 2D vehicle detections into 3D volume estimations by using prior vehicle models and projective constraints. The solution is described as a MCMC-based MAP method, on which several assumptions and simplifications are applied in order to dramatically reduce the complexity of the algorithm. Tests have shown excellent classification results under different perspectives in the presence of vehicles with heavily varied dimensions and shapes.

## ACKNOWLEDGEMENTS

The authors would like to thank the Basque Government for the funding provided through the ETORGAI strategic project iToll.

## REFERENCES

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

Barder, F. and Chateau, T. (2008). MCMC particle filter for real-time visual tracking of vehicles. In *IEEE International Conference on Intelligent Transportation Systems*, pages 539–544.

Buch, N., Orwell, J., and Velastin, S. A. (2010). Urban road user detection and classification using 3d wire frame models. *IET Computer Vision Journal*, 4(2):105–116.

Haag, M. and Nagel, H.-H. (2000). Incremental recognition of traffic situations from video image sequences. *Image and Vision Computing*, 18:137–153.

Isard, M. and Blake, A. (1998). CONDENSATION– conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28.

Khan, Z., Balch, T., and Dellaert, F. (2005). MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1819.

Kim, K., Chalidabhongse, T. H., Harwood, D., and Davis, L. (2005). Real-time foreground-background segmentation using codebook model. *Real-time Imaging*, 11(3):167–256.

Pang, C., Lam, W., and Yung, N. (2007). A method for vehicle count in the presence of multiple occlusions in traffic images. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):441–459.