# In Search of a Car
## *Utilizing a 3D Model with Context for Object Detection*

Mikael Nilsson and Håkan Ardö
*Centre of Mathematical Sciences, Lund University, Lund, Sweden*

Keywords: 3D Model, Foreground/Background Segmentation, Context, Traffic, Camera Calibration, Ground-plane.

Abstract: Automatic video analysis of interactions between road users is desired for city and road planning. A first step of such a system is object localization of road users. In this work, we present a method of detecting a specific car in an intersection from a monocular camera image. A camera calibration and segmentation are utilized as inputs by the method in order to detect a car. Using these inputs, a sampled search space in the ground plane, including rotations, is explored with a 3D model of a car in order to produce output in form of rectangle detections in the ground plane. Evaluation on real recorded data, with ground truth for one car using GPS, indicates that a car can be detected in over 90% of the time with an average error around 0.5m.



# 1 INTRODUCTION

Access to accurate positions of road users is desirable in calibration for simulations, finding potential bottlenecks, and finding potential dangers in existing road networks. The task of localizing each road user, utilizing one or several cameras, is indeed a desirable feature. Previous works with similar problem formulation has been approached in several ways. For example, some papers explore model based approaches (Koller et al., 1993; Ferryman et al., 1997; Tan et al., 1998; Li et al., 2009). Others aim to find an occupancy map from multiple views (Khan and Shah, 2006). Some formulate the problem in a probabilistic framework, for example by combining results from Markov Chain Monte Carlo (MCMC) and a Hidden Markov Model (HMM) (Song and Nevatia, 2007). Detection based methods has recently gained some attention (Pepik et al., 2012; Nilsson et al., 2013). Recently, a method utilizing 3D primitives, and monocular view, presented promising results (Carr et al., 2012). In that work, cars were modeled as boxes and pedestrians as cylinders in order to position objects in the ground-plane. This work proposes a way to search for a 3D model of a car which is more detailed than a box. In addition, a 3D context of the object is proposed to be utilized in order to get a more reliable score. One way to look at our proposal is that we exploit graphics techniques in a brute-force manner (dense sampling and rotation) in order to solve a computer vision problem. The different parts of the proposed method can be found in Fig.1.

# 2 DETECTING A CAR USING A 3D MODEL WITH CONTEXT

The aim of this section is to describe the operations performed in order to localize a specific car model, see Fig. 1. A description of the different setup and processing units for the solution will follow. The camera calibration will be addressed in 2.1, the search space in 2.2, the forgound/background segmentation in 2.3, the 3D model search in 2.4 and 2.5, and finally the non-maximum suppression in 2.6.

## 2.1 Camera Calibration

In order to get a camera calibration from an intersection used in the experiments, manually selected points had their 3D positions measured with a Leica GX1230 GG. These points were also manually positioned in an image from a static mounted camera, placed high up in a water tower, and the corresponding points were used for calibration, see Fig. 2. Calibration was performed using Tsai calibration (Tsai, 1987). The final camera parameters, are then used for mappings between the image frame and the world coordinate system.

## 2.2 Search Space - Sampled Ground Plane and Rotations

The space used to search for a specific 3D model is here chosen as a rectangle in the ground-plane $z = 0$

Nilsson M. and Ardö H..
In Search of a Car - Utilizing a 3D Model with Context for Object Detection.
DOI: 10.5220/0004685304190424
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 419-424
ISBN: 978-989-758-004-8

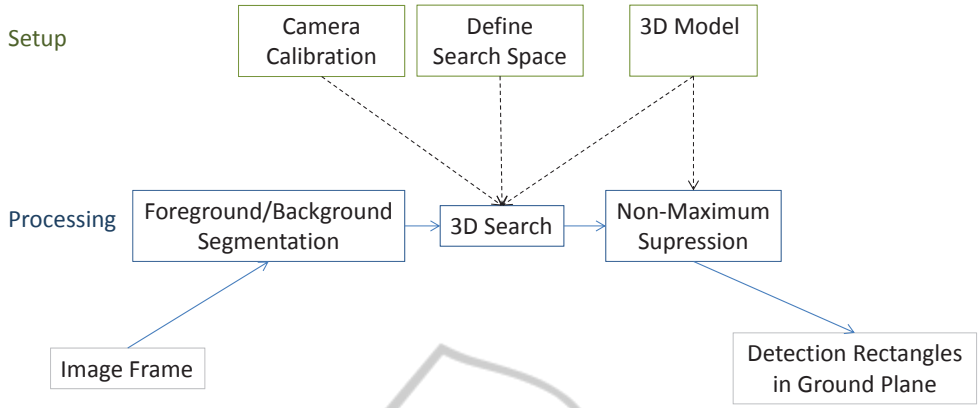


Figure 1: Overview of proposed solution.

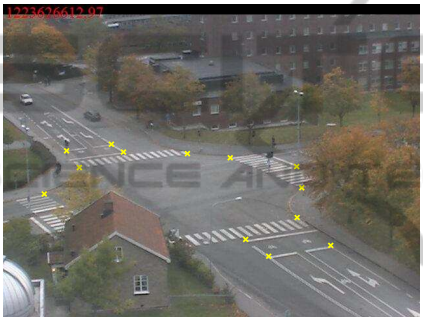

Figure 2: Calibration points used in image. These positions were measured in the world coordinate system using a high precision GPS.

Figure 3: Ground plane with positions for the dense grid (left) and the masked position for the current view (right).

with around ten centimeters steps in each direction. Further, a mask is manually created to remove positions outside the road as well as those that are covered by buildings in the viewpoint, see Fig. 3. At each position, the angle for the object is further explored, here at 22.5 degree steps. Thus, in principle this can be viewed as an occupancy map (Khan and Shah, 2006). By using a max operator for scores found at all angles at each position a visualization similar to occupancy maps can be produced.

## 2.3 Foreground/Background Segmentation

As input a probabilistic background/foreground segmentation algorithm was used (Ardö and Svärd, 2014). It produces an image that in each pixel stores the probability that this pixel currently shows a moving object as opposed to the static background. To make it robust to lighting variations and shadows, it does not utilize the image intensities directly. Instead it preprocesses the input frame by calculating the gradient direction in each pixel, and then the segmentation is based on those preprocessed input frames instead.

Gradient directions is a good feature when the gradient magnitudes are high, but can be very noisy when the magnitudes are low. This means that some gradient orientations are matched with more confidence than others. This uncertainty is estimated, which means that more weight can be put on the confident matches than those with higher uncertainty. This is achieved by using the probability distribution of gradient orientations parameterized by a signal to noise ratio defined as the gradient magnitude divided by the standard deviation of the noise. The noise level is reasonably invariant over time, while the magnitude has to be measured for every frame. Using this probability distribution the segmentation can be posed as a Bayesian classification problem with two classes, background and foreground. The classification yields a probability for each pixel that represents how likely it is that it belongs to each of the classes.

The gradient directions of the current input frame are compared with a background model that is constructed and updated online using recursive quantile estimation (Ardö and Åström, 2009). That model consists of two parts: i. A background image estimated as the median of the latest observed gradient

Figure 4: Image frame (left) and the foregound/background segmentation used (right).

directions ii. A noise level image estimated from the 25%- and 75%-quantiles of the latest observed gradient directions. An example of the segmentation can be found in Fig. 4.

## 2.4 Fast Box Search for Rapid Rejection

Note that the sampled space in the scenario described above results in 8000+ positions with angles to check with the 3D model. In order to speed up this process, a simpler box of size $1.5 \times 1.5 \times 1.5$ meters is used initially and with steps of one meter instead of 10 centimeters and with no rotation applied. This box is defined by 12 triangles. Mapping the triangles from world to image and rasterizing them produces a set of pixels $\mathcal{B}_{x,y}$ for a given position $(x,y)$ in meters from the ground plane, see Fig. 5.

Figure 5: A 3D box model shown as a yellow wire frame in the image (left) and the corresponding rasterized pixels creating a set of pixels $\mathcal{B}_{x,y}$ indicated in black (right).

From the set of pixels $\mathcal{B}_{x,y}$ a box score can be found as

$$b_{x,y} = \frac{1}{|\mathcal{B}_{x,y}|} \sum_{\mathbf{k} \in \mathcal{B}_{x,y}} P(\mathbf{k}). \quad (1)$$

where $P(\mathbf{k})$ is the probabilistic segmentation for a pixel $\mathbf{k} = [i \quad j]^T$. A threshold, $\theta_{box}$, on this box score is then applied in order to see if a more detailed search for the 3D model with rotation should take place around the point checked. Thus, a grid search is applied in the search space. In a way, this can be viewed as a 3D search variant of a sliding window cascade commonly used for object detection in images to quickly reject uninteresting patches (Viola and Jones, 2001; Dollár et al., 2012).

## 2.5 3D Model of the Car and Context

The 3D model used is that of a Toyota Corolla, a sedan car. This car was used in the experiments and also equipped with GPS sensors, which will be exploited as ground truth in the experiments. This car was manually measured with a foot ruler and a 3D model was created using a triangle mesh with 60 triangles, see Fig. 6. Note that the chosen model is more sophisticated than a box model (Carr et al., 2012) but not detailed in comparison to 10000+ triangles meshes not uncommonly used in the gaming industry. The reason for this model trade off is to strike a balance between processing speed and performance. Given a correct model for the sought object, a more sophisticated model than a box will improve the localization accuracy. However, an overly detailed model, for example considering adding wing mirrors, will be a bottleneck performance-wise and not add any significant amount to localization accuracy. This since the automatic foregound/background segmentation is noisy in practice. Furthermore, even with a close to perfect segmenter, the pixel resolution required to extract some details, is not available with the current camera used.

In order to produce some context around the car, a mid point for the 3D car model is found (middle of rectangle in $x$, $y$ and $z$ at the half height of the car) and then scaling the points with a factor $f_{context} > 1$ produces a car larger than the original. Additionally, any scaled point getting a negative z value, i.e. below the ground plane, is set to zero, see Fig. 6.

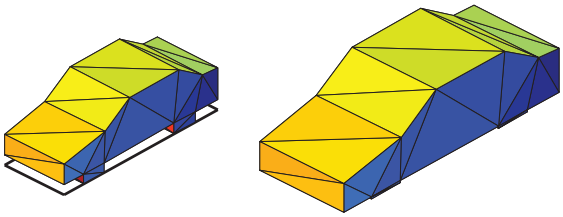

Figure 6: 3D model of a Toyota Corolla using 60 triangles and its rectangle footprint (left) and the enlarged 3D model (right) used to capture context.

Similar to the box described earlier, given a position $(x,y)$ in meters from the ground plane and an angle $a$ in degrees, both the car model and the enlarged car model are placed and rotated. First, the enlarged model is transformed from world to image coordinates and rasterized. Second, the original car model undergoes the same process. Thus, in the image, two sets of pixels are formed, the object set $\mathcal{O}_{x,y,a}$ and enlarged object set $\mathcal{E}_{x,y,a}$. From these two sets the context set $\mathcal{C}_{x,y,a}$ is formed as the difference

$$\mathcal{C}_{x,y,a} = \mathcal{E}_{x,y,a} \setminus \mathcal{O}_{x,y,a}. \quad (2)$$

An example of the object and context set are shown in Fig. 7. An object score, $o_{x,y,a}$, for a given position and rotation is found as

$$o_{x,y,a} = \frac{1}{|O_{x,y,a}|} \sum_{\mathbf{k} \in O_{x,y,a}} P(\mathbf{k}) - \frac{\alpha_{context}}{|C_{x,y,a}|} \sum_{\mathbf{k} \in C_{x,y,a}} P(\mathbf{k}) \quad (3)$$

where $\alpha_{context}$ is a variable used to strike a balance between object and its context. In order to produce a detection, in form of a rotated rectangle in the ground plane, a threshold, $\theta_{object}$, on $o_{x,y,a}$ is employed.

Figure 7: 3D car model as yellow wire frame in image (left) and the corresponding rasterized pixels creating a set of object pixels $O_{x,y,a}$ indicated in black and context pixels $C_{x,y,a}$ in gray (right).

The importance of context is to aid the decision from other objects in the scene. For example, consider the case when a bus, truck or a larger than a car vehicle is present. Not utilizing context would imply that the highest score possible might be when extracting a score within this larger object and creating several detections within it with the car model. Thus, the context where pixels should be close to zero will aid this case and produce a lower score.

### 2.6 Non-maximum Suppression on Rotated Rectangles in the Ground Plane

Output from the search of the specific object is rotated rectangles in the ground plane. Typically multiple overlapping detections for each instance of a car. The Non-Maximum Suppression (NMS) method employed here is similar to the non-rotated bounding box suppression (Felzenszwalb et al., 2010), but here rotations have to be considered also. Detections are sorted according to their score and are greedily removed if the bounding boxes are more than 0% covered. That is, no cars are allowed to overlap in the final output, by a bounding box of a previously selected detection. The overlap check for the rotated boxes can be performed using a general polygon clipper or the separating axis theorem. An example of the outputs (red) from applying threshold $\theta_{object}$ to the object score in Eq. 3, and the corresponding result (yellow) after NMS, can be found in Fig. 8.

Figure 8: Non-Maxmimum Supression (NMS) of rotated rectangles in orto-view. Red rectangles are detections from the 3D model search and yellow rectangles are the result after NMS.

## 3 EXPERIMENTS

The setup used has been explained throughout Section 2. Briefly, the 13 3D points measured at the intersection as well as manually positioned in the image, see Fig. 2, were used to calibrate the camera. The search space was defined and occluded areas removed, see Fig. 3. Foreground/background segmentation was perform on frames of video from the camera. The 3D box and car model was used and detections passing thresholds $\theta_{box}$ and $\theta_{object}$ undergoes non-maximum suppression of rotated rectangles in the ground plane. The parameters used in the experiments can be found in Table 1.

Table 1: Default values used in experiments.

| Parameter | Value |
|---|---|
| $\theta_{box}$ | 0.45 |
| $\theta_{object}$ | 0.2 |
| $\alpha_{context}$ | 1 |

In order to evaluate a specific detection within the scene, the car used was equipped with two GPS sensors. The sensors were placed in the front and at the back of the car. Given these positions, it is possible to sync the GPS to the camera in time and to find an expected rectangle footprint in the ground plane at every frame. An example from a single frame can be found in Fig. 9.

A video sequence with mixed traffic and with the GPS-equipped car performing a left turn was investigated. During this turn, the middle position of some rectangle detected by the proposed system managed to be located inside the GPS rectangle 91.4% of the time, a decent result considering monocular view and the complexity of the mixed traffic. Furthermore, the mean difference between the middle points of the

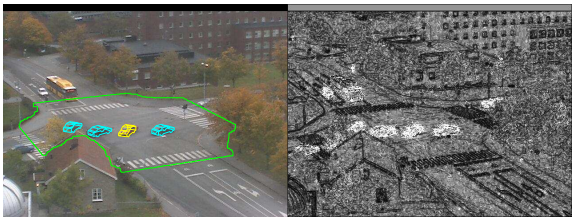

Figure 9: Examples of detections (cyan) and the detection that match to GPS position of the car (yellow) and the corresponding foreground background segmentation (right). The green border indicates the area in which the search takes place.

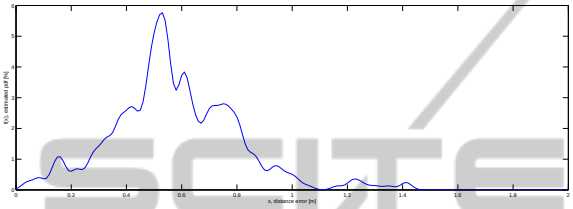

Figure 10: Parzen estimate ($h = 0.02$) of density for euclidian distances between middle of rectangles from detection and GPS.

rectangles (detected and GPS), in the cases where it was considered to be detected, was around 0.5m. A Parzen window estimated density (Parzen, 1962) for this distance can be found in Fig. 10.

## 4 CONCLUSIONS

A system searching for a specific 3D shape, a car in this case, has been presented. The proposed methodology utilizes camera calibration, a defined search space in the ground-plane, and foreground/background segmentation. Given this, the 3D object, with additional context, is proposed to be utilized in order to find a score for detection. Further, a non-maximum suppression on rotated rectangles in the ground plane is conducted to yield final detections. The system has been applied to real data with mixed traffic. Ground truth for one car in this data could be extracted by the use of a GPS. Experiments on this real data indicate that the car could be detected in 91.4% of the time it was visible and inside the search area. Furthermore, detections matching the ground-truth has an average error of 0.5m.

## 5 DISCUSSION AND FUTURE WORK

While the results are promising, improvements to the proposed framework to handle more complexity and improvement of accuracy is here discussed. For starters, currently only one model has been used, a sedan car, this should be extended with more relevant 3D shapes (vans, trucks, pedestrians, bicyclists etc). A straight forward way to perform this is to use the system described up to the Non-Maximum Suppression (NMS) for serval 3D shapes and then perform NMS for all objects.

Another extension is to place more cameras to better handle occlusions. Different approaches could be adopted here. One way could be to run the whole system up to NMS for all views. This way a score fusion could be adopted before NMS, possibly with some weighting, to produce scores taking into account scores from all views.

The system propped here does not perform any temporal processing. One possibility is to extend the system with a following tracking and thus making temporal assignments and smoothing. Given tracks to an object, yet another extension could be to adjust a detected 3D model further by optimizing the position, the angle, and the 3D shape. For example, by allowing the 3D points, which defines the shape, freedom to move with some constraints.

## REFERENCES

Ardö, H. and Åström, K. (2009). Bayesian formulation of image patch matching using cross-correlation. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8.

Ardö, H. and Svärd, L. (2014). Bayesian formulation of gradient orientation matching. *Submitted to CVPR 2014*.

Carr, P., Sheikh, Y., and Matthews, I. (2012). Monocular object detection using 3d geometric primitives. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision ECCV 2012*, volume 7572 of *Lecture Notes in Computer Science*, pages 864–878. Springer Berlin Heidelberg.

Dollár, P., Appel, R., and Kienzle, W. (2012). Crosstalk cascades for frame-rate pedestrian detection. In *Proceedings of the 12th European conference on Computer Vision - Volume Part II*, ECCV'12, pages 645–659, Berlin, Heidelberg. Springer-Verlag.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645.

Ferryman, J., Worrall, A., Sullivan, G., and Baker, K. (1997). Visual surveillance using deformable models of vehicles. *Robotics and Autonomous Systems*, 19(34):315 – 335.

Khan, S. M. and Shah, M. (2006). A multiview approach to tracking people in crowded scenes using a planar

homography constraint. In *In European Conference on Computer Vision*.

Koller, D., Danilidis, K., and Nagel, H.-H. (1993). Model-based object tracking in monocular image sequences of road traffic scenes. *Int. J. Comput. Vision*, 10(3):257–281.

Li, Y., Gu, L., and Kanade, T. (2009). A robust shape model for multi-view car alignment. In *The IEEE International Conference on Computer Vision and Pattern Recognition*.

Nilsson, M., Ardö, H., Laureshyn, A., and Persson, A. (2013). Reduced search space for rapid bicycle detection. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):pp. 1065–1076.

Pepik, B., Stark, M., Gehler, P., and Schiele, B. (2012). Teaching 3d geometry to deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*, Providence, RI, USA. accepted as oral.

Song, X. and Nevatia, R. (2007). Detection and tracking of moving vehicles in crowded scenes. In *Motion and Video Computing, 2007. WMVC '07. IEEE Workshop on*, pages 4–4.

Tan, T. N., Sullivan, G. D., and Baker, K. D. (1998). Model-based localisation and recognition of road vehicles. *Int. J. Comput. Vision*, 27(1):5–25.

Tsai, R. (1987). A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *Robotics and Automation, IEEE Journal of*, 3(4):323–344.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518.