Vision based Driver Assistance in Urban Traffic

Uwe Franke, Dariu Gavrila, Steffen Görzig
DaimlerChrysler AG, HPC: T728
D-70546 Stuttgart
Phone: (+49)711-17-41420, Fax: (+49)711-17-47054
{uwe.franke, dariu.gavrila, steffen.goerzig}@daimlerchrysler.com

Summary

Future intelligent vehicles will be equipped with cameras for driver information, warning and active assistance. Lane departure warning is at the market and more sophisticated systems for lane keeping on highways and distance control in slow traffic will follow.

In order to further extend the capabilities of the vehicles' new eyes to everyday traffic, we are developing algorithms for understanding urban traffic scenes. In this contribution we describe our approach to build an intelligent real-time vision system for this scenario. This includes stereo vision for depth-based obstacle detection and tracking, a framework for monocular detection and recognition of relevant objects including pedestrians and traffic signs and an attempt to realise such a system without the necessity of a super computer in the trunk. The computational power in our demonstrator car UTA II (Urban Traffic Assistant) is currently limited to three 400 MHz dual-Pentium II PCs.

1. Introduction

Most known autonomous vehicles have been designed for the relatively simple highway scenario where lanes are usually well marked and built with slowly changing curvature, traffic signs are large and clearly visible and other vehicles are the only potential obstacles that need to be considered. At the final presentation of the European Prometheus project, the Daimler-Benz demonstrator vehicle VITA II showed autonomous driving on highways and performed overtaking manoeuvres without any driver interaction [1].

A vision system would be even more attractive for future customers if its use is not limited to highway-like roads, but it also supports the driver in everyday traffic situations, including city traffic. Imagine an *Intelligent Stop&Go* system that is able to behave like a human driver: it does not only keep the distance to its leader constant, as a radar based system would do, but also follows the leader laterally. Moreover, it stops at red traffic lights and stop signs, gives right of way to other vehicles if necessary and tries to avoid collisions with children running across the street. The recognition of pedestrians and other vulnerable traffic participants on the road is a challenging task. The European project PROTECTOR aims to find robust solutions. Chapter 5 sketches the modules developed at DaimlerChrysler so far.

In manual mode, driver assistant systems like rear-end collision avoidance or red traffic light or 4-way stop warning are also of interest. Why not automatically switching the engine off if the car stops in front of a red traffic light and restart it when the light turns green? This saves fuel, reduces air pollution (and wakes the drive up, if necessary).

The most important perception tasks that have to be performed in order to build such systems are:

- The leading vehicle must be detected and its distance, speed and acceleration must be estimated in longitudinal and lateral direction.
- The course of the lane must be extracted even if it is not given by well painted markings and does not show clothoidal geometry.
- Small traffic signs and traffic lights have to be detected and recognised in a highly colored environment.
- Different additional traffic participants like bicyclists or pedestrians must be detected and classified.
- Stationary obstacles that limit the available free space e.g. parking cars must be detected.

2. Stereo based Obstacle Detection and Tracking

For navigation in urban traffic it is necessary to build an internal 3D map of the environment in front of the car. This map must include position and motion estimates of relevant traffic participants and potential obstacles. In contrast to the highway scenario where one can concentrate on looking for rear sides of leading vehicles, our system has to deal with a large number of different objects and must be able to react to obstacles that it has never seen before.

The most natural method to derive 3D-information is binocular stereo vision. The key problem is the correspondence analysis. One can distinguish between two different categories of stereo vision systems: feature based and area (correlation) based approaches. In our first steps towards 3D scene understanding we used an extremely fast feature based approach [2] running an a 200 MHz PowerPC. An early area based system used in a passenger is described in [3]. A 4x4 correlation is running on specialized hardware.

Thanks to the power of modern standard processors, correlation based stereo analysis can be performed without special hardware nowadays. Of cause, brute force correlation is still computationally much too expansive and remains the domain of specialized hardware. We have realized an efficient multi-resolution correlation schemes, which runs in less than 100 msec on a standard 400 MHz Pentium II PC [4]. Vertical edges mark points of interest for which the disparities are evaluated with subpixel accuracy in a gaussian pyramid. It is based on the sum-of-squared differences error criterion, checks the normalized cross-correlation coefficient

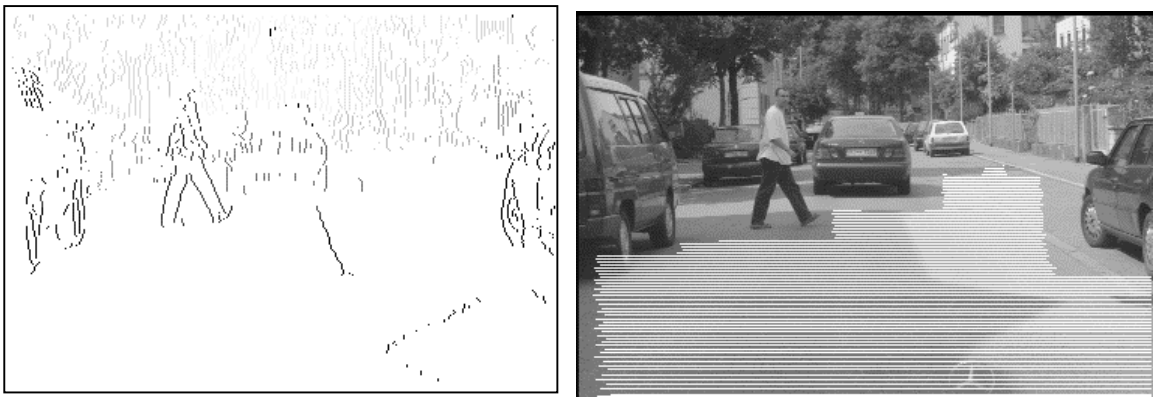

Fig. 1: a) Depth image derived by stereo vision. Darkness is inverse proportional to depth, erroneous correspondences do only occur in the background
1b) Image of the left stereo camera with marked free space

for the best match and applies a left-right test to avoid mismatches. Fig. 1a shows a depth image obtained by this approach.

From this distance image a two-dimensional depth map is derived which allows the detection of obstacles. In addition, the free space in front of the car can be obtained from this map, as shown in fig. 1b overlaid on the image of the left camera. The stereo cameras are 30 cm depart and have a viewing angle of about 40 degrees.

Detected objects are tracked over time in the depth image and their lateral and longitudinal positions and motion parameters i.e. speed and acceleration are estimated. This data is the basis for autonomous Stop&Go as well as for the recognition of pedestrians and vehicles in front.

3. Recognition of the infrastructure

Driving in urban traffic requires knowledge about the infrastructure's local information system (e.g. traffic lights, traffic signs and road markings) to comply with the traffic rules.

3.1 Traffic Light Recognition

A color camera is used to detect traffic lights. The recognition consists of three steps: color segmentation, filtering and classification. Color segmentation uses a simple look-up-table in order to determine image parts in the traffic light colors red, yellow, and green. By means of a subsequent color connected components analysis blob shaped regions are extracted. An efficient query language allows to select those regions which have areas within a certain range as possible traffic light candidates.

A region of interest (ROI) of a size adapted to the blob diameter is then cropped such that it contains not only the luminous part of the traffic light but also its dark box. The ROI is normalized to a uniform size. Eventually, a local contrast normalization by means of a simulated Mahowald retina is carried out additionally.

This pre-processed ROI serves as an input to a three-layer feed-forward neural network that performs the actual object classification task. It is constructed such that a neuron of the second network layer does not „see" the complete underlying image but only a small region of it, i.e. its spatial receptive field. These receptive fields extract local features from the input image that have been learned during the training process. The actual classification takes place in the higher network layers. The network has 2 output neurons, the class „traffic light" and the class „garbage".

The appearance of red, red-yellow, yellow, and green traffic lights is trained separately. On a 400 MHz Pentium II PC, the algorithm runs at a rate of about 8 images per second, depending on the number of traffic light candidates. In experiments, we found recognition rates of above 90%, with false positive rates below 2%. Problems occur, if green lights are too bright and the images are saturated. Investigations revealed that future high dynamic range cameras will overcome this sensor problem.

3.2 Traffic Sign Recognition

Our original traffic sign recognition system developed for the highway scenario was based on color images [5]. This approach has two drawbacks: first, the perceived color changes with the color of the illumination which has negative effects on the robustness of the algorithm and secondly, color cameras are more expensive than grayscale cameras. Therefore, we developed a shape-based traffic sign detection system for grayscale images. It works on edge features, rather than region features. The approach uses a hierarchical template matching technique based on distance transforms (DTs) [6]. DT-based matching allows the matching of arbitrary (binary) patterns in an image. These could be circles, ellipses, triangles but also non-parameterized patterns, for example outlines of pedestrians.

The pre-processing step involves computing a thresholded edge image and computing its distance image. The distance image has the same size as the binary edge image; at each pixel it contains the image distance to the nearest edge pixel of the corresponding binary edge image.

The matching step involves correlating a binary shape pattern with the distance image; at each

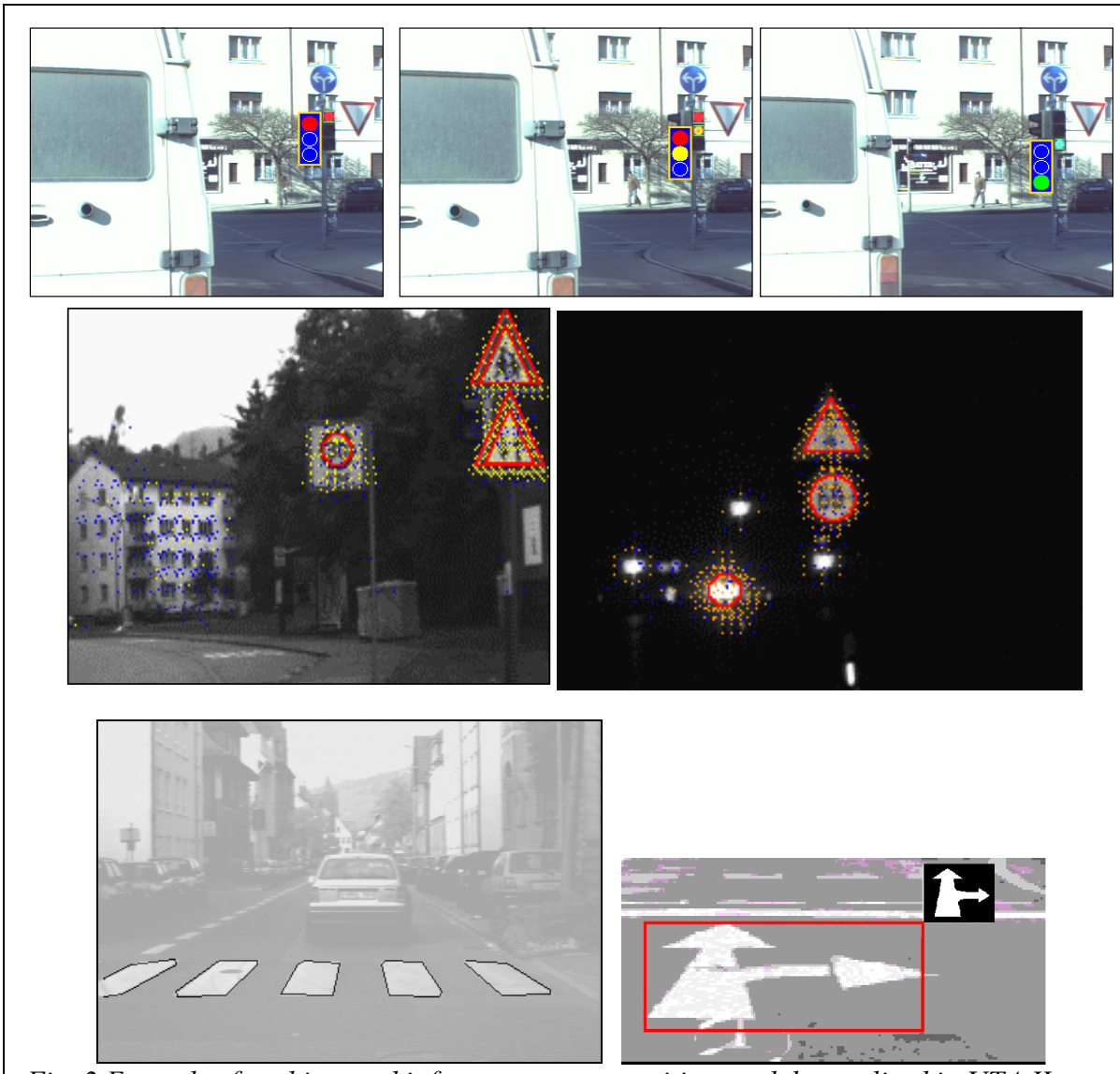

Fig. 2 Examples for object and infrastructure recognition modules realized in UTA II First row: recognition of traffic lights, second row: recognition of traffic signs, third row: recognition of pedestrian crossings and arrows

checked template location a correlation measure gives the sum of nearest-distance of template points to
image edge points. A low value denotes a good match (low dissimilarity), a zero value denotes a perfect match. If the measure is below a user-defined threshold one considers a pattern detected.

Currently, we aim to detect circular, triangular (up/down) and diamond-shaped traffic signs, as seen in urban traffic and on secondary roads. This includes stop-signs which are detected as circles. In order to optimize for speed, we chose to scale the templates (off-line), rather than scale the image (on-line). A further speed up is obtained by a coarse-to-fine search strategy.

Under good visibility conditions, we obtained high single-image detection rates, typically, of over 95%, when allowing solutions to deviate by 2 pixels and by radius 1 from the values obtained by a human. At this rate, there were a handful detections per image which are not traffic signs, on average. Most of these false positives are rejected in a subsequent verification phase, where a radial basis function network is used as pictograph classifier.

3.3 Road Recognition

Although the primary goal in Stop&Go is to follow the leading vehicle, the restrictions given by road markings, arrows and crosswalks have to be considered. Lane changes of the leader should be detected and crosswalks require special care since pedestrians have the right of way.

The recognition of the lane boundaries, stop lines and crosswalks is based on polygonally approximated contour images [7]. A database organizes the polygons with respect to their length, orientation, position and mutual spatial relations such as parallelism and collinearity. It provides fast filters for these attributes. Arbitrary combinations of properties can be specified to detect possible road structures. A recognized pedestrian crossing is shown in figure 3. Regions where obstacles have been detected by the stereo vision system of chapter 2 are masked and ruled out as possible positions of road structures.

The recognition of arrows on the road follows the two-step procedure, which is common for most of our recognition modules: detection and classification. It uses shape and brightness cues in a region-based approach. The detection step consists of a brightness segmentation step and a filtering step.

The color (i.e. grayscale) segmentation step involves reducing the number of colors in the original image to a handful. In this application, this reduction is based on the minima and plateaus of the grayscale histogram. Following this grayscale segmentation a color connected components (CCC) analysis is applied to the segmented image. The algorithm produces a database containing information about all regions in the segmented image. Among the computed attributes are the area, bounding box, aspect ratio, length and smoothness of contour as well as additional features which are determined on a need-basis, due to computational cost. We have developed a query language for this region database, dubbed „Meta-CCC", which allows queries based on attributes of single regions as well as on relations between regions (e.g. adjacency, proximity, enclosure and collinearity). The filtering step thus involves formulating a query to select candidate regions from the database. The resulting set is normalized for size and given as input to a radial basis function classifier. Fig. 3 shows the original and the obtained result.

4. Recognition of moving objects

The most important objects that we have to recognize in urban traffic are the traffic participants. On the one hand we have the vehicles (cars, trucks, busses), on the other hand are pedestrians and bicyclists. At the moment, all these objects are detected by the stereo module. Since we measure size and speed of all tracked objects, we can select those that are candidates for the mentioned object types.

For the recognition of vehicles, which should have a width in the range of 1.4 to 2.5 meters, the neural net used for the traffic light recognition has been adapted. Regions containing potential vehicles are normalized to 32x32 pixels before the classifier is applied.

The recognition of the most vulnerable traffic participants, the pedestrians, is the most challenging task in urban computer vision. At the moment, UTA II has three modules to classify the obstacles, described in the following [8].

Single image classification: The mentioned neural net has been adapted to distinguish between pedestrians and other objects, too. Images of stereo objects that have the typical size of pedestrians are normalized to 64x32 pixel here. The whole classification requires only one millisecond for each tested object. If the neural net is not sure about its output, the stereo box is subsequently analyzed by more computational expensive modules.

Image sequence classification: If the object of interest is moving, we try to verify the pedestrian hypothesis by looking for a characteristic walking pattern. The neural network used for traffic light recognition can be extended into the temporal dimension; the input then consists of grayscale image sequences. This results in a time delay neural network (TDNN) architecture with spatio-temporal receptive fields [9]. In our current approach, eight scaled pictures of the tracked object form the input vector to the neural network. Fig 3a shows a crosswalk situation, the stereo box and the 8 scaled images of the person's legs.

Shape classification: Still pedestrian candidates that have not clearly been recognized in the first stage are given to a shape based recognition module based on the distance transform also used for the traffic sign recognition.

Fig.3: Left: Pedestrian recognition based on motion. The characteristics of walking human legs becomes visible in the 8 pictures of reduced resolution shown above. They are the input to the TDNN-classifier. The stereo-box is plotted additionally.
Right: Pedestrian recognition based on shape. The silhouette that fits best to the gradient image of the pedestrian is overlaid.

We compiled a database of about 1250 distinct pedestrian shapes at a given scale; this number doubles when mirroring the templates across the y-axis. On this set of templates, an initial four-level pedestrian hierarchy was built. Five scales were used, covering a size variation of 50%. Our preliminary experiments on a database of a few hundred pedestrian images (distinct from the sequences used for training) resulted in a detection rate of about 75-85% per image, with a handful false-positives per image. These numbers are for un-occluded pedestrians. A feedback of positive results improves the results in the tracking phase. Fig. 3b shows the matching result on the image also used above.

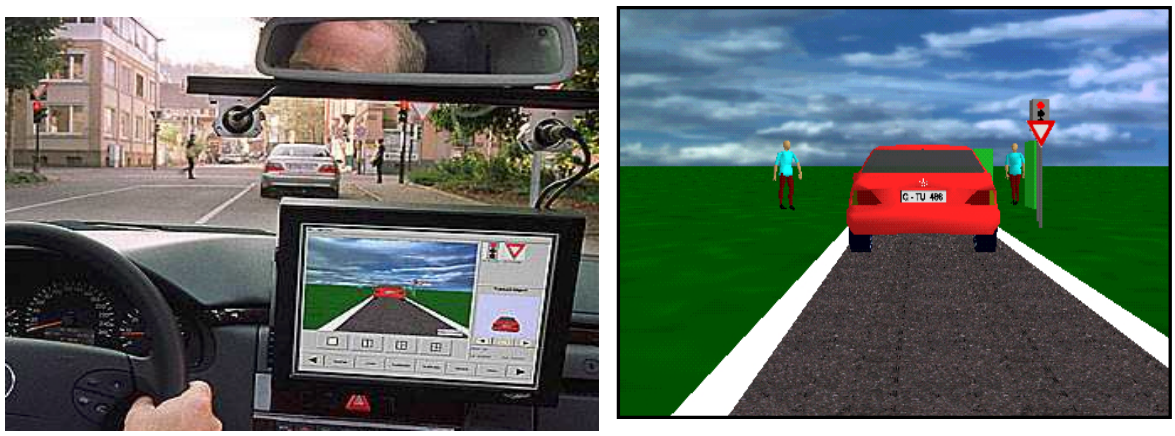

Fig.4 : View out of UTA II: the screen displays the objects detected and recognized by the vision system. The enlarged animation on the right illustrates that the geometry of the scene has been correctly analyzed

6. The UTA II demonstrator vehicle

The described recognition modules have been integrated in our UTA II (Urban Traffic Assistant) Mercedes-Benz E-class demonstrator. The stereo camera system and the color camera are connected to three dual-Pentium II PCs.

Special care has been spent on the software architecture of the complex image understanding system. Autonomous driving is challenging not only for the image processing, but also for the software architecture point of view. The software architecture has to deal with several modules running in parallel, has to coordinate the different applications and to adapt the system to different hardware environments. The architecture must also consider important issues like portability, scalability, reuse and maintenance. In UTA II a multi-agent system (MAS) is used as software architecture for controlling multiple software modules. The "Agent NeTwork System" (ANTS) offers components for several purposes: the integration and cooperation of modules, reuse of old software, distributed computing, the development of applications and test environments for system components [10].

The ANTS architecture consists mainly of two components: modules and administrators. Modules encapsulate the basic functionalities of the system. They can perform certain tasks, e.g. lane tracking on highways. The modules are controlled by administrators. An Administrator continuously determines the modules, which are currently most appropriate to achieve a specific application. This allows to configure the system during runtime and to activate the vision modules according to the current situation. For example, the recognition of traffic lights and inner-city traffic signs is switched off when UTA II is on a highway. Instead, a spe-

cial lane recognition system for such roads is activated and the stereo system is parameterized for large distances.

Figure 4 shows a view out of UTA II. You can see the stereo camera system mounted behind the windshield. The visualization on the monitor enlarged on the right displays results of the image processing algorithms: the detected lane, an obstacle in front classified as car, obstacles classified as pedestrians, a traffic light and a traffic sign. Note that the animated picture is geometrically correct since the stereo vision system has measured the 3D-positions of all relevant objects.

The main application of UTA II is autonomous Stop&Go driving in an inner-city environment. Once the car in the visualization turns red, the driver can switch the system on. From now on, the own car follows the car in front laterally and longitudinally.

7. Conclusions

UTA II shows, that vision based driver assistance is no longer limited to highways. Vision based Stop&Go is possible and can take into account traffic signs and lights, giving the car a bit of extra intelligence. Pedestrian recognition has been shown in principle, but much work remains on improving the performance of this and the other vision modules. The European project Protector aims at the recognition of those vulnerable traffic participants. However, as we have introduced the "Spurassistent" (lane departure warning system) as the first vision based system to the truck market in this year, we are convinced that image understanding systems will become standard in our future vehicles.

References

[1] B.Ulmer: „VITA II - Active collision avoidance in real traffic", Intelligent Vehicles '94, Paris, 24.-26. Oct.. 1994, S.1-6

[2] U.Franke, I.Kutzbach: "Fast Stereo based Object Detection for Stop&Go Traffic", Intelligent Vehicles '96, Tokyo, pp.339-344

[3] K.Saneyoshi: „3-D image recognition system by means of stereoscopy combined with ordinary image processing", Intelligent Vehicles '94, 24.-26. Oct. 1994, Paris, pp.13-18

[4] U.Franke, D.Gavrila, A.Gern, S.Goerzig, R.Janssen, F.Paetzold and C.Wöhler: „From door to door – principles and applications of computer Vision for driver assistant systems", *in Intelligent Vehicle Technologies: Theorie and Applications*, Arnold, 2000

[5] R.Janssen, W.Ritter, F.Stein and S.Ott: „Hybrid Approach for Traffic Sign Recognition", Proc. of Intelligent Vehicles Conference, 1993.

[6] D. Gavrila. Traffic sign recognition revisited. Mustererkennung 1999, eds. W. Förstner et al., Springer Verlag, 1999

[7] F.Paetzold, U.Franke; "Road Recognition in Urban Environment", IEEE Conference on Intelligent Transportation Systems, October 1998, Stuttgart

[8] U.Franke, D.Gavrila, A.Gern, S.Goerzig, R.Janssen, F.Paetzold and C.Wöhler: „From door to door – principles and applications of computer Vision for driver assistant systems", *in Intelligent Vehicle Technologies: Theorie and Applications*, Arnold, 2000

[9] C.Wöhler, J.K.Anlauf: "An Adaptable Time-Delay Neural-Network Algorith for Image Sequence Analysis", IEEE Trans. on Neural Networks, Vol.10, No.6, Nov. 99

[10] S.Görzig, U.Franke: „ANTS – Intelligent Vision in Urban Traffic", in IEEE Conference on Intelligent Transportation Systems, October 1998, Stuttgart