An integrated stereo-based approach to automatic vehicle guidance *

J.Weber, D. Koller, Q.-T. Luong and J. Malik

Abstract

We propose a new approach for vision based longitudinal and lateral vehicle control. The novel feature of this approach is the use of binocular vision. We integrate two modules consisting of a new, domain-specific, efficient binocular stereo algorithm, and a lane marker detection algorithm, and show that the integration results in a improved performance for each of the modules.

Longitudinal control is supported by detecting and measuring the distances to leading vehicles using binocular stereo. The knowledge of the camera geometry with respect to the locally planar road is used to map the images of the road plane in the two camera views into alignment. This allows us to separate image features into those lying in the road plane, e.g. lane markers, and those due to other objects which are dynamically integrated into an obstacle map. Therefore, in contrast with the previous work, we can cope with the difficulties arising from occlusion of lane markers by other vehicles. The detection and measurement of the lane markers provides us with the positional parameters and the road curvature which are needed for lateral vehicle control. Moreover, this information is also used to update the camera geometry with respect to the road, therefore allowing us to cope with the problem of vibrations and road inclination to obtain consistent results from binocular stereo.

*This work is to be presented at the International Conference On Computer Vision, Boston, in June of 1995.

1 Introduction

We propose an approach and develop a system for vision based longitudinal and lateral vehicle control which makes extensive use of binocular stereopsis. Novel aspects include (a) exploitation of domain constraints to simplify the search problem in finding binocular correspondences (b) temporal integration of the results of the stereo analysis to build a reliable depth map of obstacles (c) dealing with crowded traffic scenes where substantial segments of the lane boundaries may be occluded (d) on-line updating of the external camera calibration with respect to the road. The vision system is designed to interface in a modular fashion with the use of non-visual sensors such as magnetic sensors for lateral position measurement and active range sensors (e.g. Doppler radar) for an integrated approach to vehicle control such as that being investigated in the California PATH project.

Longitudinal control — i.e. maintaining a safe, constant distance from the vehicle in front — is supported by detecting and measuring the distances to leading vehicles using binocular stereopsis. A known camera geometry with respect to the locally planar road is used to map the images of the road plane in the two camera views into alignment. Any significant residual image disparity then indicates an object not lying in the road plane and hence a potential obstacle. This approach allows us to separate image features into those lying in the road plane, e.g. lane markers, and those due to other objects. In the absence of this separation, image features due to vehicles which happen to lie in the search zone for lane markers would corrupt the estimation of the road boundary contours. This problem has not yet been addressed by any lane marker based vehicle guidance approach, but has to be taken very seriously, since usually one has to cope with crowded traffic scenes where lane markers are often obstructed by vehicles.

The features which lie on the road are stationary in the scene and appear to move only because of the egomotion of the vehicle. Measurements on these features are used for dynamic update of (a) the camera parameters in the presence of camera vibration and changes in road slope (b) the lateral position of the vehicle with respect to the lane markers. Lane markers are detected and used for lateral control, i.e. following the road while maintaining a constant lateral distance to the road boundary.

For that purpose we model the road and hence the shape of the lane markers as clothoidal curves, the curvatures of which we estimate recursively along the image sequence. These curvature estimates also provides desirable look-ahead information for a smooth ride in the car.

1.1 Related Research

By far the most important and impressive work on a visually guided AVCS has been done in the group of Dickmanns (see [8] and the references cited therein). Their work resulted in a demonstration in 1987 of their 5-ton van, the VaMoRs running autonomously on a stretch of the Autobahn at speeds of up to 100 km/h. Vision was used to provide input for both lateral and longitudinal control on free roads. Further development of this work has been in collaboration with von Seelen's group in Bochum [27] and the Daimler Benz VITA project [32]. For collision avoidance with vehicles in one's lane, model-based techniques are used which exploit heuristics such as symmetry of the bounding box of the vehicle. It seems to us that these techniques are not as reliable and precise as those using binocular stereopsis.

Other examples of the numerous sites where research in this area is being conducted include the CMU NavLab project [28] which is the key university-based project on this theme. The currently favored lane following algorithm in the Navlab project seems to be the ALVINN system [24]. ALVINN is based on training a neural network with input a 30x32 low resolution image and output the desired steering command. The best performance cited is that of a 21.2 mile run at speeds of up to 55 miles per hour. The network performs reasonably well if it is trained with similar road conditions. In order to overcome the problem on which network to use, they recently proposed a connectionist superstructure — MANIAC — which incorporates multiple ALVINN networks, each of which is pretrained for a particular road type ([13]).

A more recent project on road following in the US is a collaboration between NIST and Florida Atlantic University[25]. Lateral control is based on sensing the optical flow at a certain tangent point on the lane and steering so as to make it have no horizontal component. There may be difficulties if the tangent point is occluded.

A basic module in any of the visually guided vehicle control algorithms has to be the detection and tracking of lane boundaries. A leading project is the LANELOK system developed at GM Research. Continuing work has resulted in a real-time implementation reported in [1].

The use of binocular stereopsis for vehicle control has been successfully demonstrated by JPL's planetary robotic vehicle [20] and Nissan's PVS vehicle [23]. Both systems realize a tradeoff between performance time and density of a depth map. For obstacle detection it is actually not necessary to compute a dense depth map neither is it necessary to perform the depth map computation at video rate. A trinocular stereo system is used by [26], where the third camera actually serves as a mean to confirm and refine the results obtained from two cameras.

Research closest related to our obstacle detection approach is described in [33]. They perform first a rectification of the stereo images to achieve zero vertical disparity, before they estimate the disparity by comparing the variances of the difference in a window of the rectified left and right image. Although they exploit the full camera geometry, they do not use the knowledge of the ground plane disparity to reduce the search space, neither does their approach apply to lane marker detection. They furthermore admit that their approach is quite computationally expensive.

Other, more classical approaches for obstacle detection are based on motion stereo or optical flow interpretation ([9, 4]). The key idea of these approaches is to predict the optical flow field model for a moving observer under constraint motion (e.g. planar motion). Obstacles are then detected by a significant difference between the predicted and the actual observed optical flow field. The major drawbacks of these approaches are (a) computational expense and (b) the lack of reliable and accurate optical flow fields and the associated 3D data caused by strong vertical motions a car usually experience while driving on a highway.

A combination of stereo and optical flow is suggested in [5] in order to perform a temporal analysis of stereo image sequences of traffic scenes. They do not explicitly address the problem of obstacle detection in the context of vehicle guidance, but the general problem of object identification. They extract and match contours of significant intensity changes in (a) stereo image pairs for 3D information and (b) subsequent frames to obtain their temporal displacement ([22]). In order to distinguish between

obstacles and road boundaries or lane markers, they also exploit some heuristics like horizontally and vertically aligned contour segments as well as 3D information extracted from the stereo data ([21]).

While the most successful lane marker based approaches for lane following perform quite well in uncrowded traffic scenes, where lane markers are clearly visible and not obstructed by other vehicles, we expect them to fail or at least to perform not so well in crowded traffic scenes, where lane markers *are* obstructed by other vehicles. On the other hand we expect our approach to perform reasonable well even in crowded scenes, since we explicitly distinguish and reason about lane markers and obstacles, lying in the search region for lane markers.

1.2 Control using lane flow

In order to deal with lateral control of the car, we have to obtain an estimate of its motion relative to the road. The approaches based on optical flow are computationally expensive and sensitive to the vertical vibration induced by the car suspension. We study instead a method relying on the global relative motion of the lane markers. If the car is on a straight portion of the road or a a circular arc, the extend of lateral misalignment from the correct trajectory is indicated by the rate and extend of slewing and side-slipping of the lane markers [11]. One of the big advantages of this framework is that it enables us to formulate the problem of road following as one of nulling deviation from the steady state by using visual information that is easy to extract directly from the image (generalizing the approach of [25]). In the coordinate system linked to the car, this is:

- the difference in horizontal offset of the lane marker
- the difference in orientation of the lane marker

In addition to these parameters, we also take into account road curvature (like [8, 15]). Its allows us to improve the control strategies in the sense that predictions for control variables are available that provide a smooth and safe ride.

A model featuring these parameters has already been successfully used by Dickmanns and al [8], but there is a substantial difference between our approach and his. Dickmanns approach relied heavily on an

integrated dynamic model with a large number of parameters including vehicle dynamics and implicit control laws. Since such a model includes already most of the relevant information, the perception phase needed only to be minimal: the vision system just had to consider a window containing the novel section of the road to update the dynamic parameters. This approach therefore required only very little computational resources. By contrast, our approach is to first estimate the lane flow, and then to apply a control system. The computation of lane flow require that at each time the measurement of the entire lane, and not just a small portion of it. Therefore, our approach requires more computational resources. On the other hand, it has a number of advantages. First, since we do a global detection of the lane, and since we are able to use stereopsis, we expect that by exploiting the redundancy, our approach will be more robust. In particular, it deals properly with occlusions caused by other cars, an issue that was not previously addressed. Second, we decouple the problem of estimating lane flow and curvature (which is visual positional information) from the estimation of other dynamical parameters. Therefore, we can experiment with different control strategies, since they form an independent module. Moreover, the vision module can be considered just as one of the sensors (together with magnets and Doppler radars also used in the PATH project), and different multisensorial integration approaches can be considered.

1.3 Outline of our approach

In this section, we describe briefly our approach, and refer to subsequent sections of the paper for more details on the individual aspects of our system. It is to be noted that for reasons of space, we are not able to give all the technical details (some of them are in [17]) and rather try to give a description which is sufficient for the reader to understand what we are doing.

The idea behind our approach is to build a reliable and efficient system by exploiting a number of geometric constraints which arise from the configuration of our stereo rig, and from the fact that the road can be modeled locally as a plane. These geometric constraints are detailed in Sec. 2.

At each new instant, we first compute the stereo disparity using an efficient algorithm based on the

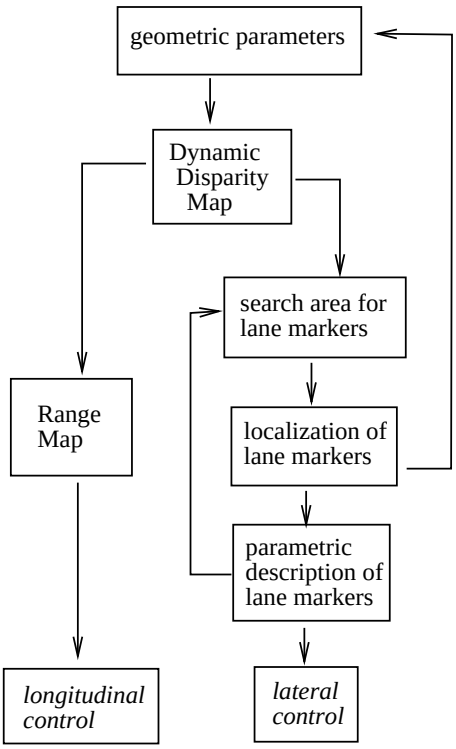


Figure 1: The flow of information in our integrated approach

Helmholtz shear (Sec. 3). The disparity map is used in two ways. First, a 3D obstacle map is dynamically updated over time by tracking identified vehicles and introducing new vehicles which appear. (Sec. 4). This provides the information needed for longitudinal control, ie measuring the distances to leading vehicles. Second, the areas of the image belonging to the ground plane are identified. This ensures that the search area for lane markers (which is defined using the parametric description of the lane markers which was found at the previous instant) is not corrupted by occlusions. Within this area, the lane markers are localized by a specialized feature detector (Sec. 5). From the image positions of the lane markers, we can update the geometric parameters of the stereo rig (Sec. 6) The new parameters will be used to compute the stereo disparity at the next instant, and to map the lane markers to the ground plane, where a parametric description is obtained for them (Sec. 5). This parametric description provides the information needed for lateral control, ie maintaining a constant distance to the road boundary. The flow of information that we just described is summarized in Fig. 1.

2 The geometrical model

2.1 A stereo rig viewing a plane

In our application, the vision system consists of a binocular stereo rig. The road surface plays an important role, since it contains the lane markers to be tracked for lateral control, and since every object which lies above it is to be considered as a potential obstacle. Our key assumption is that this surface can be locally modeled as a plane.

The camera is modeled as a pinhole camera. A 3-D point of world coordinates X, Y, Z is projected to a 2-D point of pixel coordinates u and v according to the equation:

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}_w & \mathbf{T}_w \\ \mathbf{0}_3^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

In this equation $\mathbf{A}$ is a 3×3 matrix containing the camera intrinsic parameters. The image center is the point $[u_0, v_0]$, and if we suppose that the aspect ratio of the image is 1, then the scale factors α_u and α_v are identical, and proportional to the focal length f. $\mathbf{R}_w$ and $\mathbf{T}_w$ are the rotation and translation mapping the world coordinate system to a particular coordinate system attached to the camera, called the *camera coordinate system*. Its origin is the center of projection, the Z-axis is perpendicular to the image plane, the X-axis is parallel to the horizontal directions of the image, and the units are such that the distance between the center of projection and the image plane is 1. Therefore, in this coordinate system, a 2-D point $[x, y]$ is associated to the optical ray $[x, y, 1]$. The intrinsic parameters matrix $\mathbf{A}$ allows us to obtain normalized coordinates from pixel coordinates, and vice-versa. They can be determined by camera calibration [30], and we will generally assume in the sequel of the paper that this is the case.

It is known that (except in some degenerate cases) there is a one-to-one correspondence between

the image plane $\mathcal{R}_1$ and a given plane Π, and that this correspondence is given by the homography:

$$\mathbf{m}_1 = \mathbf{H}_{\Pi 1}\mathbf{M}_\Pi$$

where $\mathbf{m}_1$ are the projective coordinates of a point of $\mathcal{R}_1$ and $\mathbf{M}_\Pi$ are the projective coordinates of a point of Π. This operation allows us to map every measurement of the image of the ground plane into a point of the ground plane.

In the case of two cameras, an identical relation holds for each of the cameras, and therefore, by composition, we get that the two images $\mathbf{m}_1$ and $\mathbf{m}_2$ of a point $\mathbf{M}_\Pi$ on a given plane Π are related by the homographic relation:

$$\mathbf{m}_2 = \mathbf{H}_{12}\mathbf{m}_1$$

This relation means that if we observe the projection $\mathbf{m}_1$ in the first image of a point $\mathbf{M}_\Pi$ of the plane Π, we are then able to *predict* the position of the projection of $\mathbf{M}_\Pi$ into the second image.

2.2 The Helmholtz shear

In a particular case, this relation reduces to what we call the *Helmholtz shear*, a configuration where the process of computing the stereo disparity is tremendously simplified. We have chosen this term to acknowledge the fact that this insight is due to Helmholtz [12] more than a hundred years ago. Helmholtz observed that objectively vertical lines in the left and the right view perceptually appear slightly rotated. This led him to the hypothesis that the human brain performs a shear of the retinal images in order to map the ground plane to zero disparity. The mathematical reasoning is as follows: Under the viewing geometry of parallel axes, disparity for points on the ground plane has a zero value on the horizon and increases linearly with the image plane coordinate in the vertical direction. By applying a constant shear (the amount is a function of the distance between the eyes and height above the ground plane) one can map all points on the ground plane to have zero disparity. Then, any object above the ground plane will have non-zero disparity. This is very convenient because the human visual

system is most sensitive around the operating point of zero disparity. A related approach has been presented in [19].

In the most general situations where the *Helmholtz shear* applies, the correspondence between two views of a point of the road plane can therefore be described by the relation:

$$\begin{cases} u' = u + h_{12}v + h_{13} \\ v' = v \end{cases} \tag{2}$$

which means that the matrix $\mathbf{H}_{12}$ takes the special form:

$$\mathbf{H}_{12} = \mathbf{I}_3 + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & h_{12} & h_{13} \end{bmatrix}$$

It is known that the expression for the general homography is:

$$\mathbf{H}_{12} = \mathbf{A}'(\mathbf{R} + \frac{1}{d}\mathbf{t}\mathbf{n}^T)\mathbf{A}^{-1} \tag{3}$$

In this expression, $\mathbf{A}$ (resp $\mathbf{A}'$) is the matrix of intrinsic parameters of the first (resp. second) camera. The motion parameters $\mathbf{R}$ and $\mathbf{T}$ describe the displacement between the two cameras in the first camera normalized coordinate system. The equation of plane Π is $\mathbf{n}^T\mathbf{M} = d$, where $\mathbf{n}$ is the unit normal vector of the plane and d the distance of the plane to the origin. From this expression, one can easily see that the correspondence $\mathbf{H}_{12}$ is a *Helmholtz shear*, if and only if:

- the intrinsic parameters $\mathbf{A}$ and $\mathbf{A}'$ are the same[1]
- the rotation $\mathbf{R}$ is the identity
- the translation $\mathbf{T}$ has only a component along the X-axis

[1] Actually with the exception of the parameters u_0 and u_0' which may differ. In the sequel, we will just suppose that $\mathbf{A}$ and $\mathbf{A}'$ are totally identical.

- the component of the normal $\mathbf{n}$ along the X-axis is zero

In such a situation, described in Fig. 2, the stereo rig is entirely characterized by the following parameters:

- intrinsic parameters of the first camera $\mathbf{A}$
- baseline b

The position of the plane with respect to the stereo rig can be described by two parameters (that we will call the geometric parameters), for instance:

- the height of the stereo rig with respect to the road plane d
- the angle of tilt of the stereo rig with respect to the road plane α

The mapping from the image plane to the ground plane can then be expressed as:

$$\mathbf{H}_{1\Pi} = \begin{bmatrix} -t_{wz}/\cos\alpha & t_{wx}\tan\alpha & t_{wx} \\ 0 & t_{wy}\tan\alpha - t_{wz} & t_{wy} + t_{wz}\tan\alpha \\ 0 & \tan\alpha & 1 \end{bmatrix} \mathbf{A}^{-1} \tag{4}$$

where: α is the angle of rotation (along the X-axis) mapping plane frame to camera frame, and $\mathbf{t}_w$ is the translation from plane frame to camera frame. Note that the choice of t_{wx} and t_{wy} is arbitrary, whereas t_{wz} is the distance d just mentioned. Let us now consider a system of two cameras which are in a configuration where the Helmholtz shear applies. All the parameters which appear in (4) have to be identical, except t_{wx} and t'_{wx} whose difference is the baseline b. By combining $\mathbf{H}_{1\Pi}$ and $\mathbf{H}_{2\Pi}$ as obtained from (4), we get the expression of the coefficients of the correspondence $\mathbf{H}_{12}$:

$$\begin{cases} h_{12} = b/t_{wz}\sin\alpha \\ h_{13} = b/t_{wz}\cos\alpha \end{cases} \tag{5}$$

3 Binocular Stereopsis

The major task of our longitudinal control module is to extract distance and relative velocity measurements of objects appearing in front of the car from stereo image pairs. Although proposed algorithms in the literature for computing binocular stereopsis are quite computationally expensive we are able to reduce the complexity considerably by using region-of-interest processing and exploitation of domain constraints.

Using the extracted information for the depth and velocity we can formulate obstacle hypotheses that will be verified by temporal integration, as described in Sec. 4.

3.1 Ground Plane Alignment

Computing the stereo disparity requires solving some correspondences of image features in the left and right view. The process of computing the stereo disparity is tremendously simplified by using the *Helmholtz shear* described in Sec. 2. After applying this very simple transformation to the image, obstacles get mapped to points of non-zero disparity making them very easy to detect. See Fig. 3 where (a) and (b) are a stereo image pair, and (c) and (d) show the points of interest on the sheared left and on the right image. Significant disparities correspond to obstacles. From the residual disparity at a certain image location we obtain the 3D location of the point. Furthermore, the epipolar lines are horizontal lines which makes the procedure for establishing correspondences especially simple.

Disparity Computation We perform the *residual* disparity computation[2] on a reduced set of points-of-interest on horizontal scanlines in the image. These points-of-interest are simply the locations in the image with significant horizontal grey value changes. The disparity is found by computing the normalized correlation between small horizontal windows in the two images at the locations of the

[2]For the sake of performance we actually do not compute a sheared image according to the ground plane disparity but use the amount of the shearing as a predisplacement of the search area.

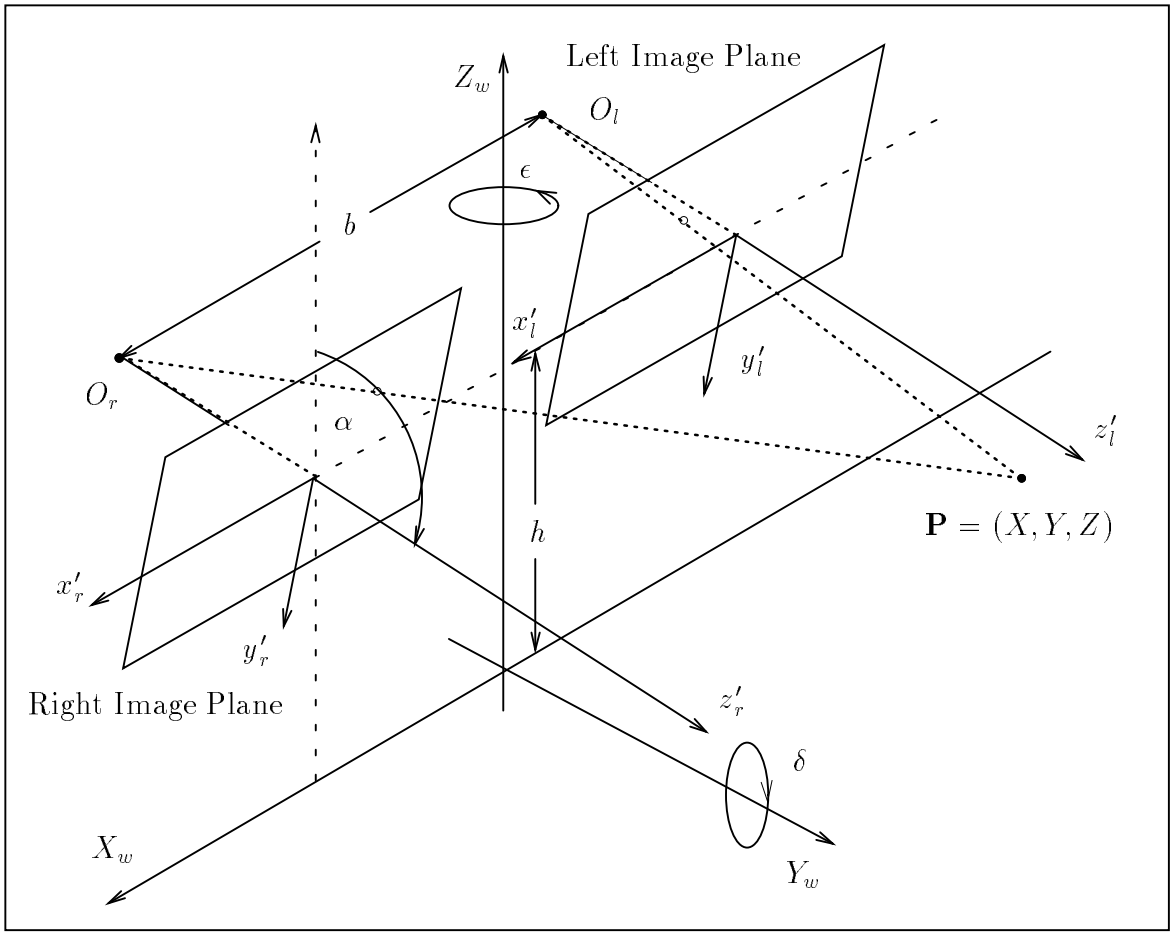


Figure 2: A general view of a stereo camera set-up where the baseline b is parallel to the ground plane and the optical axes of the cameras are parallel with an inclination angle α towards the normal of the ground plane.

points-of-interest. The normalized correlation for disparity shift τ, at horizontal image location x is:

$$\rho_x(\tau) = \frac{\sigma^2_{l,r}(x+\tau, x)}{\sigma_{l,l}(x+\tau, x+\tau)\sigma_{r,r}(x, x)} \tag{6}$$

where the correlations $\sigma_{i,j}(x, y)$ are approximated by summations

$$\sigma^2_{i,j}(x, y) = \sum_{u=-W/2}^{+W/2} g_i(x+u) g_j(y+u) \tag{7}$$

which are calculated over a window of size W.

Points-of-interest are those locations in the right image where the value of $\sigma^2_{r,r}$ is above a threshold.

The normalized correlation function is calulated only in those regions. If the maximal value of the function $\rho(\tau)$ is above a threshold (set at 0.9 in our experiments) we assume that the disparity can be calculated. Sub-pixel disparities are obtained by quadratic interpolation of the function about the maximum τ.

To illustrate the results from stereo correspondence, the maximum correlation disparity and corresponding depth computation for a single stereo image pair are shown in Fig. 3.

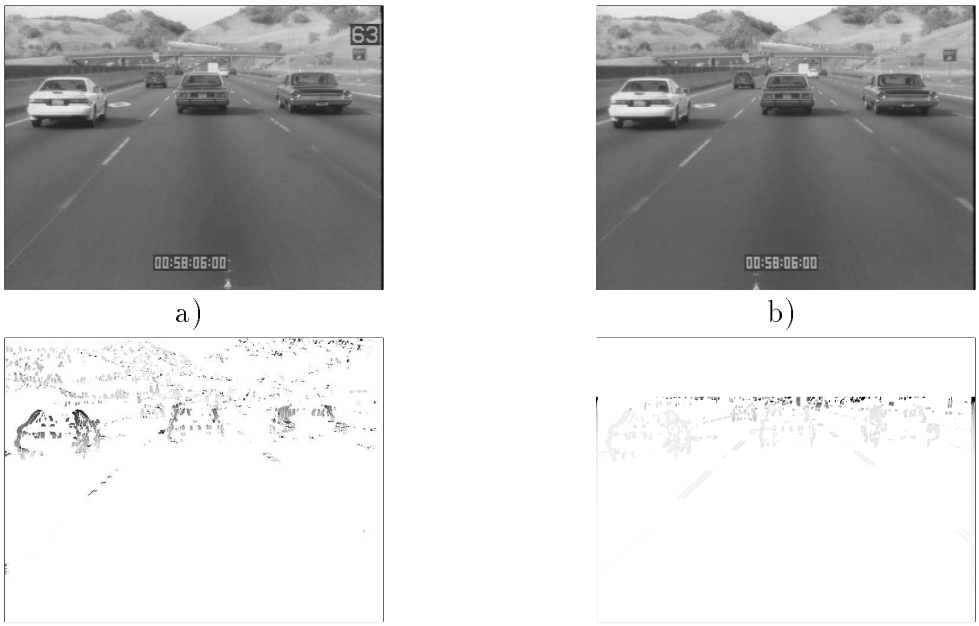


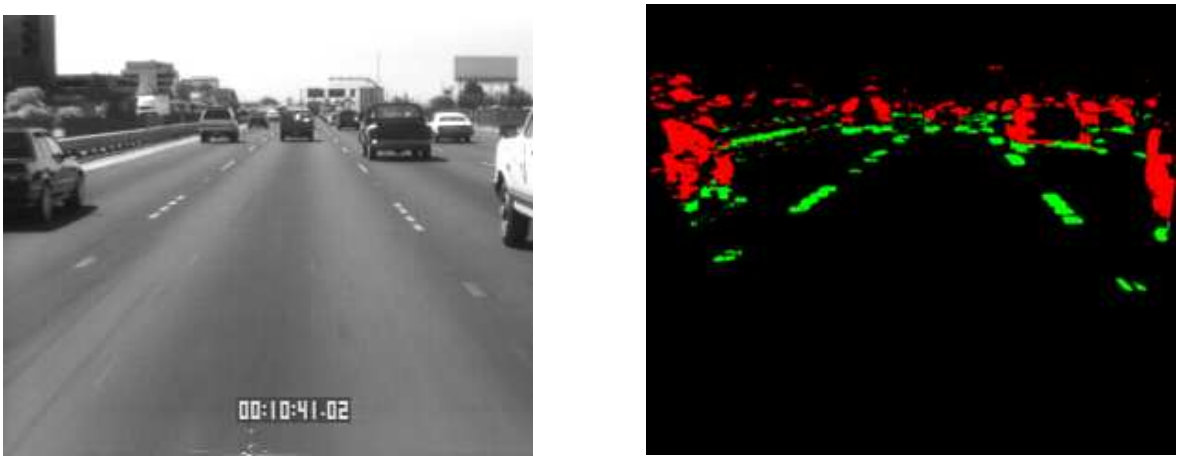
Figure 3: a) left and b) right view; c) disparity map and d) depth map.

Computing the disparity in this way is relatively computationally inexpensive. Finding the points of interest involves a simple convolution and summation which can be performed at frame rate using processors which support vector operations such as DSPs. In addition, computing the correlation function only over a subset of feature points reduces the amount of computation.

3.2 Obstacle Detection

Residual disparities — which appear in the image after the ground plane disparity has been mapped to zero — indicate objects which appear above the ground plane. A simple threshold can be used to distinguish between features lying on the ground plane (e.g. lane markers or other stuff painted on the

road) and features due to objects lying above the ground plane (which may become future obstacles). To make the ground plane/ above ground plane distinction we label those points with residual disparity above a given threshold as existing above the ground plane while those with disparity delow the threshold are assumed to lie on the ground plane. Figure 4 shows the result on a single frame. Objects detected as above the ground plane are colored red and those on the ground plane are colored green. Black regions indicate regions where the disparity could not be accurately recovered. Notice that the tires of the other vehicles as well as their shadows are detected as being located on the road surface.



a) b)

Figure 4: a) left image and b) green indicates objects were detected to be on the road surface, red indicated objects are above the road surface.

The process for obstacle detection is sketched in Fig. 5.

4 Temporal integration

Computing depth from just a pair of images is known to be sensitive to noise. In addition, correctly dealing with occlusion at depth discontinuities may require a computationally expensive algorithm [3]. One can improve the accuracy of the depth estimation by exploiting the temporal integration of information with time using the expected dynamics of the scene. For the case of an autonomous vehicle, the input consists of two video signals providing 30 frames per second. In addition we can exploit the physical constraints of the environment. We can assume we are interested in connected, rigid objects.

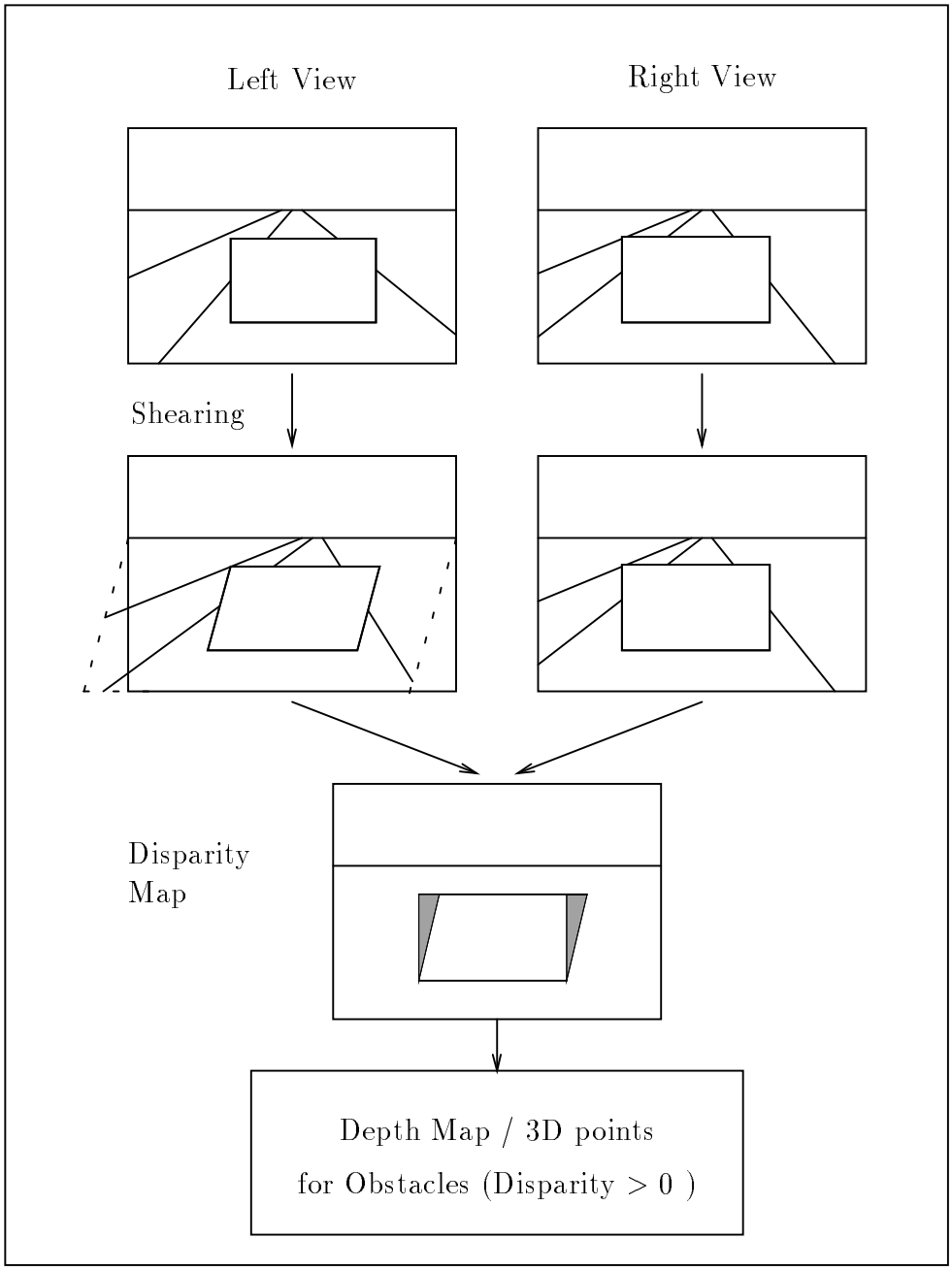


Figure 5: Obstacle detection by detecting residual disparities: to detect obstacles we compare a precomputed disparity according to features on the ground plane and the measured disparity. Any significant differences are due to features lying above the ground plane and are potential candidates for projected features of an obstacle which may preclude the course of the vehicle.

This would allow us to use spatial coherence in identifying objects from the depth map. Also, objects of interest will be assumed to be either other vehicles on the road or stationary objects connected to the road plane. Since we are anticipating certain dynamics, Kalman Filters can be used to track objects

with time.

We utilize the spatial coherence of objects in order to segment the depth map into objects of interest. First, connected components are found in a $3D$ space consisting of the two image dimensions plus the depth dimension. In the two image dimensions, points are connected if they are one of the 4 nearest neighbors. In the depth dimension they are connected if the difference in depth is less than the expected noise in the depth estimates. This expected noise is simply the variation in depth due to a uniform uncertainty of 0.5 pixel in disparity. Thus more distance objects will have larger uncertainty. Figure 7 gives an example of two objects which are connected in this image/depth $3D$ space.

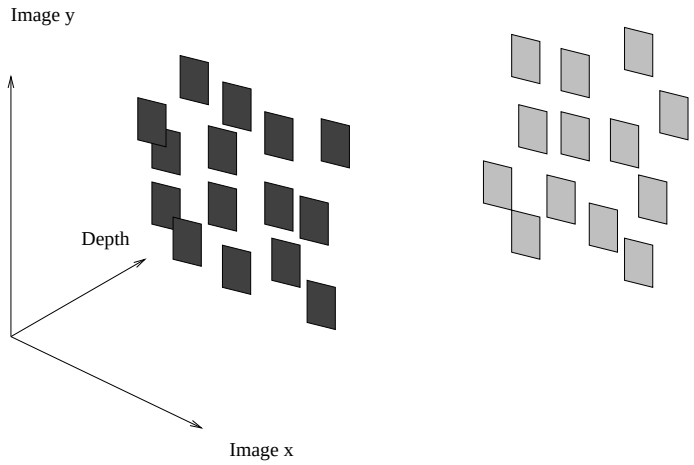


Figure 6: Connected components in image/depth space consist of those pixels which are nearest neigbhors in image coordinates as well as having depth differences less than depth uncertainty.

These connected components form the basis of potential objects which are to be tracked with time. If the same object appears in two consecutive frames, we can initialize a Kalman filter to track its position and velocity with respect to our vehicle. For each frame that the object is visible, the accuracy of our knowledge about that object will improve. Figure 7 show the objects found by this method. The objects have colored boxes around them. The connected components are the colored pixels within the boxes. On the right side of the image is a view from above the road surface showing the locations of the identified objects with respect to our vehicle. Only those objects which are within the lanes of traffic in the same direction as our vehicle are shown. Other vehciles in the opposite direction as well as traffic signs are detected but not shown.

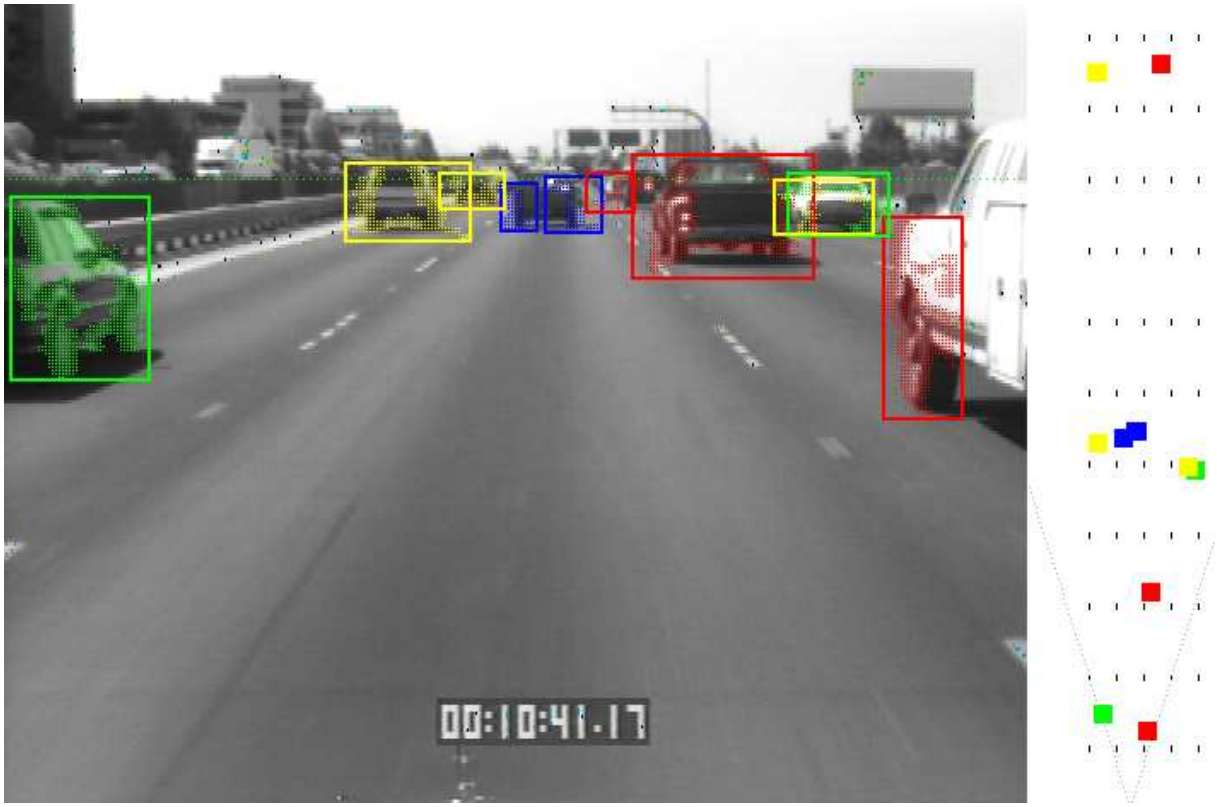


Figure 7: Objects identified as being in the same lanes of traffic as the test vehicle. On the right side of the image is a "birds-eye-view" from above the road surface showing the relative position of the tracked objects with respect to the test vehicle. Other objects such as cars in the other lanes and road signs were identified but are not shown to reduce clutter.

5 Lane marker detection and tracking

The goal of lane marker detection and tracking is to provide us with the information needed for lateral control. In the coordinate system linked to the car, this is:

- the horizontal offset of the lane marker
- the orientation of the lane marker

All the identification is done in a 3D reference coordinate system attached to the car, rather than a 3D coordinate system attached to the road. This allows us to obtain parameters which are directly relevant in terms of control actions. In addition to these parameters, we also take into account road curvature that we want to detect as far ahead as possible.

In order to reduce the computational cost, we use only predetermined sampled positions along vertical axis. These positions are uniformly sampled in the image, so that the density in the ground plane is proportional to the precision of reconstruction.

5.1 Initialization and control structure

When the system is started (or reinitialized in case of detected unconsistencies), we assume that we are in the most usual configuration where the car is on a straight portion of the road. We search for the position of potential line markers as straight lines, using the fact that they are roughly parallel. The image of Gaussian derivatives is back-projected to the ground plane. We select by a voting procedure the common orientation of line markers, and by a clustering procedure the offsets of line markers. The algorithm is fast, since the image of Gaussian derivatives is already available, since it was computed to find the features during disparity computation. An example of lines detected by the algorithm is shown Fig. 9.

In the steady state mode, the tracking is based on a predict and verify procedure. The following loop continually executes:

- *Predict* new parameters for each lane marker. This is done by a locally constant velocity model for each of the parameters.

- Define horizontals *search* bands in the image (an example is given Fig. 9) based on the predicted parameters, their uncertainties, and the previous homography matrices. The information provided by stereopsis is used here to exclude points which are images of obstacles.

- *Localize* within the search zone the center of line markers, using a model of the image intensity produced by these line markers (bright bars). An example of the points found is shown Fig. 9, and details are given Sec. 5.2.

- By comparing the positions of the feature points in the left and right image, *update* the values of the homography matrices for each image, as explained in Sec. 6.

- Backproject the points just found to the ground plane using the updated homography matrices and the predicted parameters. *Fit* a new clothoid model to the backprojected points. An example of a portion of clothoid found is shown in Fig. 10, details are given Sec.5.3.

- *Update* the model parameters

5.2 The 2-D feature detector

The localization of the lane markers is done in two stages. First, we localize the center of the marker by performing a local convolution with an approximation of the elongated 2D filter:

- In the direction perpendicular to the predicted position of the lane marker, the filter is a second derivative of a Gaussian, its width being matched to the expected width of the marker.

- In the direction parallel to the lane marker, the filter is a Gaussian.

For reasons of efficiency, this scheme has been slightly modified, in that the image window around the lane is warped (like in [2]) to align the expected orientation with the vertical, instead of using a rotated filter. This allows us to use a separable filter and therefore to avoid having to do 2D convolution.

The candidate positions are obtained as local maxima of the responses. To select the final position, we then use a model-based approach. Within the expected interval of width, we compare the brightness value and variance of the bar and of the background. This allows us to select the final position, as well as to obtain a measure of uncertainty based on contrast, and variance ratio.

5.3 The 3-D road model

The road model that we use is based on the the actual road layouts widely used in civil engineering to produce high-speed roads. Each of the line markers detected is modeled as a plane curve which is characterized by the $N + 2$ parameters illustrated by Fig. 8:

- O: lateral offset of the line in the car's coordinate system
- θ: angle between the direction of the line at the closest position and the line of sight of the car
- $C_1 \ldots C_N$ curvature at predetermined positions, the first one being the closest and the last one the farthest of the zone being processed.

The clothoid model, which is used in road designing, consists in assuming that the curvature along the road is a continuous function of arc length , with a piecewise constant variation a_i: $C(s) = C_1 + a_i s$. The constant is supposed to change only at the predetermined positions, and therefore the depth band that we are processing is split into a number of sections of constant curvature variation. The model can represent accurately straight lines, arc of circles, and the transitions between them. In spite of the small number of parameters, and of these simplifications, the model is more accurate than assuming just zero curvature [31, 6, 14] or parabolic sections [16]. In order to avoid unstabilities in the portions of low curvature, the sign of the curvature is constrained to be identical along the different segments.

The computation of the actual 3D points from the values of the model parameters is quite expensive, since curvature is a second-order quantity and therefore, two integral calculations would be required in the exact case. Even using the first-order approximation [8] is not fast enough, and thus to have a fast

access to the curve from its parameters we use a look-up-table which is precomputed. This look-up table describe a single arc of clothoid with positive curvatures (the negative curvatures are obtained symmetrically) and is indexed by values of orientations, initial curvature and final curvature.

The fitting procedure is based on a non-linear minimization of an error criterion of the form:

$$\min_{O,\theta,C_1 \ldots C_N} \sum_i \{w_i(x_i(O,\theta,C_1 \ldots C_N) - x_i(\text{measured}))^2\} + \lambda \sum_{j=0}^{N} C_j$$

where the x_i are the lateral positions at each vertical sample, the w_i are the confidence measures computed at the feature detection phase, and λ is the relative weight of a regularization term which favors low curvatures. The minimization is done using a gradient approach. Since the variation from frame to frame is small, it usually converges quite fast.

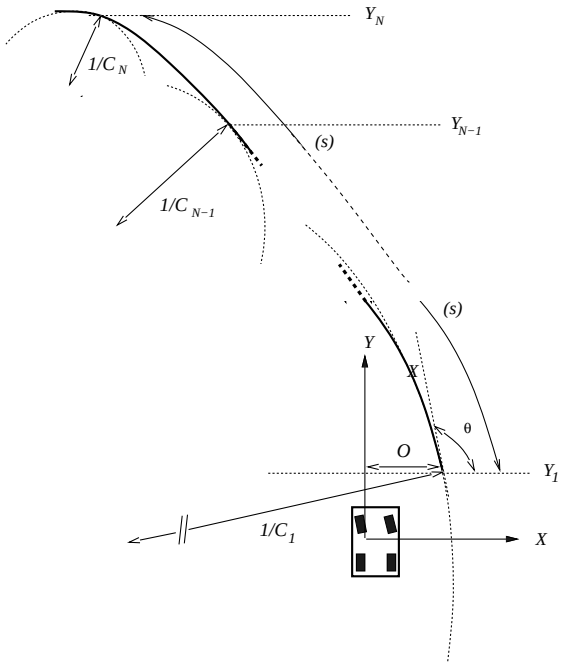


Figure 8: The model of line markers

5.4 Examples

The algorithm has been tested in typical road scenes. The results that we have obtained so far seem satisfactory, even if we have not yet been able to compare them with the ground truth. Although we have not yet introduced any further hypothesis on the line markers such as parallelism, the results of curvature estimation are fairly consistent with what could be expected, as can be seen in Fig. 10. As shown in Fig. 11, occlusion of lane markers might lead to spurious values of the curvature parameters. Removing the occluded points using using the residual disparity with respect to the ground plane enables us to find that the lane being tracked is straight.

Figure 9: The initialization of the algorithm is done by detection of portions of straight lines of common orientation (top left). Within the search zone predicted (top right), a precise localization of line markers points if performed (bottom).

6 Determining the camera geometry

So far, we have supposed that the camera geometry with respect to the road is known, and fixed. However, the movements of the car's suspension and the change in vertical road curvature can affect

parameter	left	right
lateral offset (m)	0.73	4.57
orientation (degrees)	90.6	90.0
initial radius of curvature $\frac{1}{C_1}$ (m)	2074	3910
final radius of curvature $\frac{1}{C_2}$ (m)	309	297

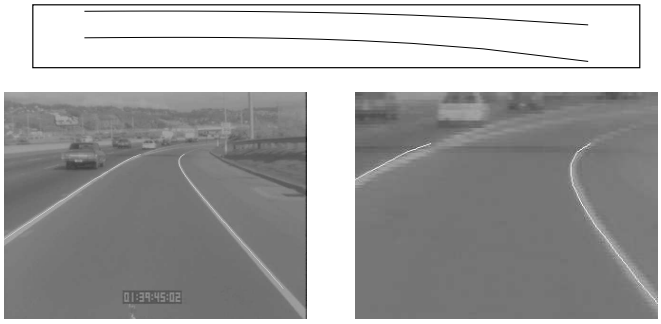


Figure 10: Estimated parameters, top view, and reprojected view of the fitted clothoids. The zoom shows that the fit is quite precise, in spite of the large distance and curvature variation.

parameter	all points	occlusion removal
lateral offset (m)	2.14	2.10
orientation (degrees)	-0.67	-0.45
initial radius of curvature $\frac{1}{C_1}$ (m)	10000	10000
intermediate radius of curvature $\frac{1}{C_2}$ (m)	10000	10000
final radius of curvature $\frac{1}{C_3}$ (m)	250	10000

Figure 11: Estimated parameters and zoom showing the feature points (green) and the fitted lane markers (red). Left: all the feature points are used. Right: the spurious feature points (purple) are removed using the residual disparity with respect to the ground plane

this geometry. Indeed, it has been reported [7] that a small difference in the assumed and actual camera tilt angle with respect to the ground affects the 3D reconstruction significantly. Moreover, the operation of mapping the ground plane disparity to zero is very sensitive to this parameter, as a small error in the inclination angle will cause a significant error on the localization of the ground plane. Therefore, we need a way to update the camera geometry relative to the ground plane.

The idea is to use the measurements of the image of the road to compute the external parameters. Since this turns out to be a most critical task for our application, we describe three different solutions that we have implemented.

6.1 Using residual disparity

The camera parameters most affected by movement of the car's suspension and change in vertical road curvature are the inclination angle α and camera height, h (refer to figure 2). The points-of-interest which exhibit small residual disparities are assumed to lie on the ground plane. This residual disparity can be used to update the inclination angle α and height h. The idea is to minimize with respect to α and h the sum of squares of differences between these measured disparities and the disparity under the ground plane assumption. The values of α and h are then continuously updated over time using a linear Kalman Filter based on the dynamics of α and h. For example, the height h is modeled as a damped harmonic oscillator driven by noise. This is a model consistent with the suspension system of the car. Details can be found in Appendix A.

There are essentially two origins for variations in α and h: a short term variation due to camera vibrations, which requires a large process noise, and a long term variation caused by a change in the slope of the road, which can be captured using a small process noise. An example of the results obtained from a sequence of 210 frames recorded during freeway driving is shown in figure 12.

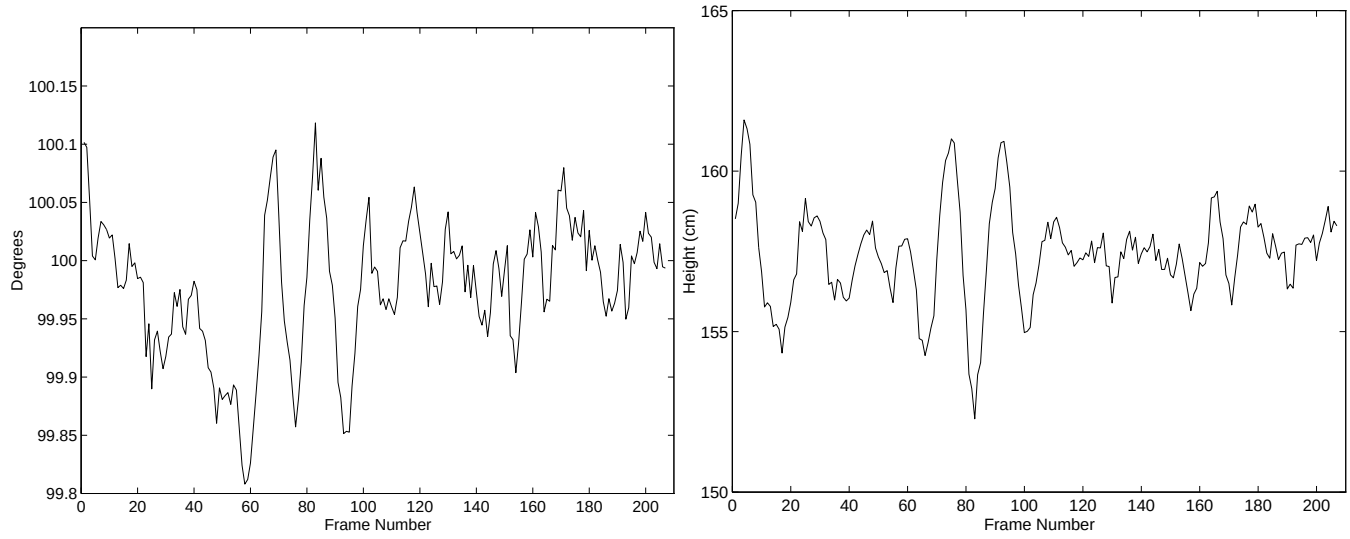


Figure 12: Camera inclination angle and camera height estimated from ground plane disparities for a freeway driving sequence. The 210 frames are from 7 seconds of video.

6.2 Using the binocular correspondence of lane markers

One way to identify features that actually lie on the road plane, is to consider the lane markers. The advantage is that the lane markers are detected and localized in each view using a model-based approach. Moreover, the positional parameters are obtained from several measurements. Therefore, we are able to identify points of the road plane in a precise and reliable way, in each camera.

In the most general case, the correspondence of at least four different points is needed to compute the homography $\mathbf{H}_{12}$ between the two images [18]. The homography can then be used to separate the obstacles from the ground plane. Moreover, the knowledge of the intrinsic parameters allows one to decompose the homography matrix according to (3) and to get the relative position and orientation of the two cameras, as well as the position of the plane with respect to the cameras [29]. From these parameters, the homographies between the images and the ground plane $\mathbf{H}_{1\Pi}$ and $\mathbf{H}_{2\Pi}$ can be obtained. In the situations where the *Helmholtz shear* applies, the calculations are more simple, and this is what we describe next.

When the *Helmholtz shear* applies, the correspondence between two views of a point of the road plane can be described by the two parameters h_{12} and h_{13} as in (2). Therefore, having the correspondence of two points which lie on the ground plane is sufficient to determine this correspondence. In practice, we

want to exploit redundancy in data, and there are two ways to do so. The first way is to compute the coefficients of the correspondence from the global parameters describing a lane marker. Each of these global parameters are obtained by a least-squares fit, and therefore are more reliable than individual points.

In the (often encountered) case of straight lines markers, the computation is particularly easy. If the line of equation $u = Av + B$ in the first image is in correspondence with the line of equation $u = A'v + B'$ in the second image, then, we obtain:

$$h_{12} = A - A' \quad , \quad h_{13} = B - B'$$

This method allows us to compute the homography between the two images of the ground plane. Please note that this does not assume at this stage any camera calibration, which was to be expected, since the identification of a plane is an operation which can be done in un-calibrated perspective images, a fact already exploited for obstacle detection by [10].

However, to obtain the metric information (consisting of position and orientation of markers, and curvature) which is required for lateral control requires the knowledge of the internal parameters of the stereo rig, as well as the geometric parameters: height and angle of tilt of the stereo rig with respect to the road plane. The geometric parameters can be computed from the correspondence if the internal parameters are known. In our model, these parameters are the intrinsic parameters of each camera, and the baseline b, and we can assume that their variation can be neglected compared to the variation of the geometric parameters. They are related to the correspondence coefficients h_{12} and h_{13} introduced previously by (5). Therefore:

$$\alpha = atan(h_{12}/h_{13}) \ , \ t_{wz} = b/\sqrt{h_{12}^2 + h_{13}^2}$$

We then obtain the homographies mapping the image plane to the ground plane by (4).

6.3 Using monocular clues

The geometric parameters can also be determined from a single view, provided that we have some a priori knowledge about the lane markers. We have studied one situation which is of practical importance because it happens often in the context of road following. In this particular situation, we view at least two lane markers which are parallel lines.

and suppose that their separation is also known.

The set of parallel lines intersect at a vanishing point v. The homography between the image plane and the road plane maps this vanishing point to a point which must lie at infinity on the road plane, which means that the third component of $\mathbf{H^{-1}v}$ must be zero. Writing this relation using (4) yields the value of the vanishing point as a function of the camera intrinsic parameters and the tilt angle α:

$$\mathbf{v} = [0, -1/\tan\alpha]^T$$

The determination of the camera height t_z can be done if further knowledge is introduced. For instance, if the separation of two parallel lane markers is known, then we can obtain t_z by solving a quadratic equation.

7 Conclusion

We have proposed an integrated approach for vision based longitudinal and lateral vehicle control. In this approach, the vision module provides the following information to be used for any control system:

- detection of the leading vehicles and measurement of their distance,
- estimation of the flow of lane markers and of road curvature at a distance

The main originality of our approach is the extensive use of binocular stereopsis, and its integration with lane marker detection.

We have presented a new stereo algorithm which exploits domain constraints to achieve efficiency in computing the instantaneous disparity map, as well as a new method to update dynamically a 3-D obstacle map, which is well suited to the nature of our problem. Our results illustrate the fact that by combining these two algorithms, it is possible to obtain a reliable and relatively dense obstacle map at a small computational cost.

We have then shown that the lane tracking task, which is traditionally carried on monocularly, can benefit significantly from a stereo based-approach. In particular, we are able to deal with crowded traffic scenes where substantial segments of the lane markers may be occluded. The binocular measurement of the lane markers enables us to perform an on-line updating of the external geometric parameters of the stereo rig with respect to the road. We can therefore deal with change in slope of the road, as well as variations in inclination of the car caused by bumps or accelerations and decelerations.

Acknowledgements

This research was supported by CALTRANS through PATH MOU 94 and 131.

A Dynamical Camera Update Using Residual Disparity

In Section 6 we outlined how the camera geometry can be updated with time as it changes with road surface and vehicle suspension dynamics. One of the ways to track camera geometry parameters is through the disparity of objects known to lie on the ground plane.

From equations (2) and (5) we saw that the disparity, $\Delta x'$, for objects on the ground plane can be written as a function of screen position y' and camera geometry. Camera geometry is defined by the inclination angle α, height of the cameras t_z, baseline b and focal length f. We assume that the cameras remain horizontal and the the baseline and focal length are fixed. The angle α and height h however can be affected by vehicle motions. The disparity for ground plane objects is:

$$\Delta x' = \frac{b\ f\ \cos\alpha}{t_z} - \frac{b\ \sin\alpha}{t_z}\ y' \tag{8}$$

We assume that variations from this are due to noise in the estimates of the camera geometry. We write an error function $E(\alpha, t_z)$ which we want to minimize with respect to α and t_z:

$$E(\alpha, t_z) = \sum_i \|\Delta x'_i + \frac{b}{t_z}(y'_i \sin\alpha - f\cos\alpha)\|^2. \tag{9}$$

The summation is carried out over all feature points which are assumed to be on the ground plane.

We take partial derivatives of this cost function with respect to both camera parameters and set them equal to zero. This assumes that the values of α and t_z are very close to their correct values since we are linearizing a non-linear function.

Differentiation with respect to t_z produces an equation polynomial in t_z^{-1}:

$$\frac{\partial E}{\partial t_z} = -\frac{2b}{t_z^3}F - \frac{2b}{t_z^2}G \tag{10}$$

where

$$\begin{aligned} F &= \sum_i y_i'^2 \sin^2\alpha + f^2\cos^2\alpha - 2fy_i'\sin\alpha\cos\alpha \\ G &= \sum_i \Delta x_i'(y_i'\sin\alpha - f\cos\alpha) \end{aligned} \tag{11}$$

Under the assumption that t_z is non zero, equating the partial derivative to zero gives:

$$t_z = -\frac{G}{bF} \tag{12}$$

Differentiation $E(\alpha, t_z)$ with respect to α and setting to zero yields the following trigonometric equation:

$$A\ \cos\alpha + B\ \sin\alpha + C\ \sin 2\alpha + D\ \cos 2\alpha = 0, \tag{13}$$

with

$$\begin{aligned} A &= \sum_i \Delta x_i' y_i', \\ B &= \sum_i \Delta x_i' f, \\ C &= \frac{1}{2}\sum_i \frac{b}{t_z}(y_i'^2 - f^2) = \frac{b}{2t_z}\sum_i y_i'^2 - \frac{b\,f^2}{2t_z}\,N, \\ D &= -\sum_i \frac{f\ b}{t_z} y_i'. \end{aligned}$$

where N is the number of measurement points.

In order to solve this nonlinear equation in α we linearize it using a Taylor expansion of the trigonometric functions around an initial value α_0, which we obtain from the static camera calibration This is equivalent to solving this nonlinear equation using the first iteration of the Newton-Raphson method. We set:

$$\alpha = \alpha_0 + \delta\alpha \tag{14}$$

and approximate:

$$\begin{aligned}
\sin(\alpha) &= \sin(\alpha_0 + \delta\alpha) \approx \sin\alpha_0 + \delta\alpha\cos\alpha_0, \\
\cos(\alpha) &= \cos(\alpha_0 + \delta\alpha) \approx \cos\alpha_0 - \delta\alpha\sin\alpha_0,
\end{aligned}$$

and obtain the solution for $\delta\alpha$:

$$\delta\alpha = \frac{A\cos\alpha_0 + B\sin\alpha_0 + C\sin(2\alpha_0) + D\cos(2\alpha_0)}{A\sin\alpha_0 - B\cos\alpha_0 - C\cos(2\alpha_0) + D\sin(2\alpha_0)}. \tag{15}$$

The inclination angle α and camera height t_z are then consistently updated over time using linear Kalman Filters. The model dynamics for the angle α are based on the assumption that α is constant with a small random process noise term. The variable t_z is assumed to be a damped harmonic oscillator driven by a small noise term. This is a valid approximating model for a car's suspension system. The natural frequency of the suspension was set to $2.038Hz$. This value came from finding the dominant spectral component of the camera's height as a function of time from sensor data. It is also consistent with the expected dynamics of the test vehicle's suspension.

References

[1] O.D. Altan, H.K. Patnaik, and R.P. Roesser. Computer architecture and implementation of vision-based real-time lane sensing. In *Proc. of the Intelligent Vehicles '92 Symposium*, pages 202–206, 1992.

[2] D. Aubert and C. Thorpe. Color image processing for navigation: two road trackers. Technical Report CMU-RI-TR-90-09, Carnegie Mellon University, 1990.

[3] Peter N. Belhumeur. A baysian treatment of the stereo correspondence problem using half-occluded regions. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 506–511, Champaign, Illinois, June 15-18, 1992.

[4] S. Carlsson and J.-O. Eklundh. Object detection using model based prediction and motion parallax. In *Proc. First European Conference on Computer Vision*, pages 297–306. Antibes, France, Apr. 23-26, 1990, O. Faugeras (ed.), Lecture Notes in Computer Science **427**, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[5] S. Chandrashekar, A. Meygret, and M. Thonnat. Temporal analysis of stereo image sequences of traffic scenes. In *Proc. Vehicle Navigation and Information Systems Conference*, pages 203–212, 1991.

[6] J. Crisman. *Color Vision for the Detection of unstructured Roads and Intersection.* PhD thesis, Carnegie-Mellon-University, 1990.

[7] D. DeMenthon and L.S. Davis. Reconstruction of a road by local image matches and global 3d optimiztion. In *Proc. International Conf. on Robotics and Automation*, pages 1337–1342, 1990.

[8] E.D. Dickmanns and B.D. Mysliwetz. Recursive 3-d road and relative ego-state recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:199–213, 1992.

[9] W. Enkelmann. Obstacle detection by evaluation of optical flow fields from image sequences. In *Proc. First European Conference on Computer Vision*, pages 134–138. Antibes, France, Apr. 23-26, 1990, O. Faugeras (ed.), Lecture Notes in Computer Science **427**, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[10] F. Ferrari, E. Grosso, G. Sandini, and M. Magrassi. A stereo vision system for real time obstacle avoidance in unknown environment. In *Proc. IROS*, pages 703–708, Tsuchiura, Japan, 1990.

[11] D.A. Gordon. Perceptual basis of vehicular guidance. *Public Roads*, 34(3):53–68, 1966.

[12] H.v. Helmholtz. *Treatise on Physiological Optics* (translated by J.P.C. Southall), volume 1–3. Dover, NY, 1925.

[13] T.M. Jochem, D.A. Pomerleau, and C.E. Thorpe. Maniac: A next generation neurally based autonomous road follower. In *Image Understanding Workshop*, pages 473–479, Washington, DC, April 18-23, 1993, 1993.

[14] S.K. Kenue. Lanelok: Detection of lane boundaries and vehicle tracking using image-processing techniques: Part i+ii. In *SPIE Mobile Robots IV*, 1989.

[15] K. Kluge and C. Thorpe. Representation and recovery of road geometry in YARF. In *Proc. Intelligent Vehicles*, pages 114–119, Detroit, MI, 1992.

[16] K. Kluge and C. Thorpe. Representation and recovery of road geometry in yarf. In *Proc. Intelligent vehicules symposium*, pages 114–119, 1992.

[17] D. Koller, Q.-T. Luong, and J. Malik. Binocular stereopsis and lane marker flow for vehicle navigation: lateral and longitudinal control. Technical Report UCB/CSD-94-804, University of California at Berkeley, March 1994.

[18] Q.-T. Luong and O.D. Faugeras. Determining the Fundamental matrix with planes: unstability and new algorithms. In *Proc. Conference on Computer Vision and Pattern Recognition*, pages 489–494, New-York, 1993.

[19] H.A. Mallot, H.H. Bulthoff, J.J. Little, and S. Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological cybernetics*, 64(3):177–185, 1991.

[20] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal of Computer Vision*, 8:71–91, 1992.

[21] A. Meygret and M. Thonnat. Object detection in road scenes usng stereo data. In *Pro-Art Workshop on Vision*, Sophia Antipolis, April 19–20, 1990.

[22] A. Meygret, M. Thonnat, and M. Berthod. A pyramidal stereovision algorithm based on contour chain points. In *Proc. Second European Conference on Computer Vision*, pages 83–88, S. Margherita, Ligure, Italy, May 18-23, 1992, G. Sandini (ed.), Lecture Notes in Computer Science **588**, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[23] M. Ohzora, T. Ozaki, S. Sasaki, M. Yoshida, and Y. Hiratsuka. Video-rate image processing system for an autonomous personal vehicle system. In *IAPR Workshop on Machine Vision Application*, pages 389–392, Tokyo, Japan, Nov. 28–30, 1990, 1990.

[24] D.A. Pomerleau. Progress in neural network-based vision for autonomous robot driving. In *Proc. of the Intelligent Vehicles '92 Symposium*, pages 391–396, 1992.

[25] D. Raviv and M. Herman. A new approach to vision and control for road following. In *Conference on Computer Vision and Pattern Recognition*, pages 217–225, Lahaina, Maui, Hawaii, June 3-6, 1991.

[26] Bill Ross. A practical stereo vision system. In *Conference on Computer Vision and Pattern Recognition*, pages 148–153, Seattle, WA, June 21-23, 1993.

[27] M. Schwartzinger, T. Zielke, D. Noll, M. Brauckmann, and W.v. Seelen. Vision-based car-following: Detection, tracking, and identification. In *Proc. of the Intelligent Vehicles '92 Symposium*, pages 24–29, 1992.

[28] C. Thorpe, editor. *Vision and Navigation: The Carnegie-Mellon Navlab.* Kluwer Academic Publishers, Norwell, Mass, 1990.

[29] Roger Tsai and Thomas S. Huang. Estimating Three-dimensional motion parameters of a rigid planar patch, II: singular value decomposition. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 30, 1982.

[30] R.Y. Tsai. Synopsis of Recent Progress on Camera Calibration for 3D Machine Vision. In Oussama Khatib, John J. Craig, and Tomás Lozano-Pérez, editors, *The Robotics Review*, pages 147–159. MIT Press, 1989.

[31] M.A. Turk, D.G. Morgenthaler, K.D. Gremban, and M. Marra. Vits — a vision system for autonomous land vehicle navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:342–361, 1988.

[32] B. Ullmer. Vita - an autonomous road vehicle (arv) for collision avoidance in traffic. In *Proc. of the Intelligent Vehicles '92 Symposium*, pages 36–41, 1992.

[33] Y. Zheng, D.G. Jones, S.A. Billings, J.E. W. Mayhew, and J.P. Frisby. Switcher: a stereo algorithm for ground plane obstacle detection. *Image and Vision Computing*, 8:57–62, 1990.