Recovering a comprehensive road appearance mosaic from video

Ivan Sikirić
Mireo d.d.
Cebini 28, 10000 Zagreb
e-mail: ivan.sikiric@mireo.hr

Karla Brkić, Siniša Šegvić
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb
e-mail: name.surname@fer.hr

***Abstract*—We describe a system that employs a single calibrated camera mounted on a moving vehicle to produce a road appearance map as a comprehensive mosaic of individual orthogonal views. The system first transforms the current image of the road acquired from a driver's perspective into the orthogonal view by inverse perspective mapping. Consequently, the orthogonal image is aligned with previously recovered parts of the mosaic by an exhaustive search optimization technique. Experiments have been performed on videos taken along public roads through Croatian countryside and small cities. The obtained results are provided and discussed.**

I. Introduction

We consider a setup in which video is captured using a single calibrated camera mounted on a moving vehicle. The vehicle is driving down the road for which we try to obtain a comprehensive surface appearance map. We apply inverse perspective mapping [1], [2] to each of the captured video frames in order to obtain orthogonal (bird's eye) view of the road surface [3]. Neighbouring orthogonal views typically have many corresponding road pixels, i.e. pixels which are projected from the same point of the road surface. We strive to employ these correspondences in order to place all orthogonal views in global alignment one by one. Fusion of the aligned orthogonal views results in the desired comprehensive map which we term road appearance mosaic.

A surface appearance map of the road can be useful in a variety of applications. For instance, it can be used to verify appropriate placement of road surface markings, which is critical for traffic safety, especially at the crossings. It can also be used to verify the extent of road maintenance (the surface under new asphalt). Assessing the state of road surface in many countries is still performed manually by human operators, and is a time consuming and cumbersome process [4]. Providing a georeferenced road appearance map would speed up and simplify this process considerably, by enabling the verification of the road markings without the need for on-location measuring. Furthermore, road appearance map could be used to verify the existing cadastre maps against the actual conditions. It is even possible to obtain a vectorized terrain map suitable for GIS databases [5].

II. Related work

Existing approaches for obtaining road appearance mosaics and similar forms of road appearance maps differ in the number and types of sensors installed on the acquisition vehicle and in the level of supervision required.

Given that the appearance, structure and other properties of the road are well constrained, it is possible to use a wide array of sensors which add additional cues about the road and hence improve the end result. Commonly used sensors include stereo cameras, laser scanners, GPS, odometers, etc. Different combinations of sensors call for different approaches and algorithms.

Wang et al [6] describe a system which works with data obtained by a mobile mapping vehicle equipped with an inertial navigation system, dual frequency GPS, 6-12 color cameras and an odometer. Given the GPS information, multi-camera panoramic images and sensor calibration parameters, their algorithm outputs a GIS-database-compatible road geometry information, which consists of a 3D lane line model of all the lane lines observed in the video. For each line, line type and color attributes are also available. To obtain the model they first perform a variant of inverse perspective mapping, which enables them to get an orthogonal view of the road. Orthogonal view of the road is beneficial because it simplifies lane line detection. Inverse perspective mapping relies on the assumption that the road is locally planar, which is not always the case. Hence, pitch correction of the mapping is achieved by modeling the road surface profile using geolocation information. For each frame, exact position of the measurement rack in a geographical coordinate system is known. Using this information, it is possible to estimate the spatial trajectory of the vehicle, which corresponds with the road surface profile. Having obtained the orthogonal image, line segments are extracted, linked, classified and added to the model. The system is fully automated.

Shi et al [7] rely on videos acquired from a vehicle equipped with an odometer, two GPS receivers, two sets of stereo camera systems and three laser scanners. By using laser scanners they obtain range data for road and roadside objects in the form of 3D point clouds. The laser data is then fused with the image data to obtain fully automated spatial positioning of road parts, which results in a road appearance map. Other interesting results with laser scanners are available in [8], [9].

There are approaches that rely on some amount of human interaction. For example, Barinova et al [10] present an algorithm for road mapping which is continuously trained to detect a road with the help of a human operator. The algorithm includes an offline learning stage and an online operation / correction stage. However, strong supervision is not a drawback in this case, as the purpose of the described system is to be used as an interactive

This research has been jointly funded by Croatian National Foundation for Science, Higher Education and Technological Development and Institute of Traffic and Communications under programme Partnership in Basic research, project number #04/20.

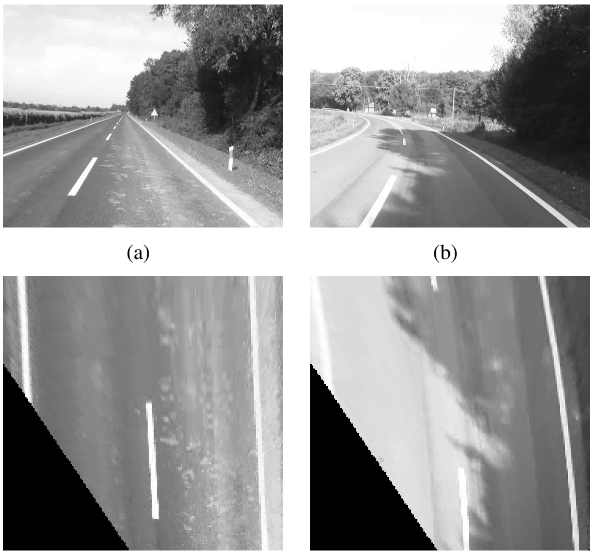

Fig. 1. Images captured from a moving vehicle (a,b) and the corresponding inverse perspective mappings (c,d). Different configurations of the road are shown: a straight road (a) is transformed into the orthogonal image (c), a slightly curved road (b) is transformed into the orthogonal image (d).

tool for examination of road defects.

To summarize, using multiple different sensors yields better maps. However, the overall cost and complexity of obtaining a map increases with the number of sensors. Depending on the application, the systems for road mapping can be fully automated or interactive. The system we would like to build is similar in spirit to the approach of Wang et al [6], as it also produces a road appearance mosaic, however we would like to achieve that using as little sensorial input as possible. To the best of our knowledge, none of the previous research has addressed the problem of recovering road appearance mosaics by only employing a single perspective camera.

III. Inverse perspective mapping

A single captured video frame represents a projection of a 3D scene onto the image plane (cf. figure 1(a), 1(b), figure 2(bottom) and figure 3(bottom)). This process is generally impossible to invert since it is not injective. However, we are only interested in obtaining the image of the road surface from an orthogonal perspective (cf. figure 1(c), 1(d)), and do not need to attempt a full 3D reconstruction of the scene. If we assume that the road surface is contained in a plane, and that it is not occluded by other objects, then we can employ inverse perspective mapping [1], [2], [11] to obtain orthogonal images.

In the following, we shall denote the points of the plane $\mathbf{q} \in \pi$ by homogeneous coordinates [12] such that $\mathbf{q}_i = [x_i, y_i, 1]^\top \forall i$, where (x_i, y_i) denote the usual cartesian coordinates in the Euclidean plane. Denote the points on the road plane as $\mathbf{q}_\mathrm{R}$, their projections to vehicle's camera plane as $\mathbf{q}_\mathrm{P}$ and their projections to the orthogonal plane as $\mathbf{q}_\mathrm{I}$. Then these points can be related by the following bijective mapping [12], [13]:

$$\mathbf{q}_{\mathrm{P}i} = \mathbf{H}_\mathrm{RP} \cdot \mathbf{q}_{\mathrm{R}i}, \forall i \qquad (1)$$

$$\mathbf{q}_{\mathrm{I}i} = \mathbf{H}_\mathrm{RI} \cdot \mathbf{q}_{\mathrm{R}i}, \forall i \ . \qquad (2)$$

The transformations $\mathbf{H}_\mathrm{RP}$ and $\mathbf{H}_\mathrm{RP}$ are planar projective mappings, which are often also referred to as homographies. From (1) and (2) follows:

$$\mathbf{q}_{\mathrm{I}i} = \mathbf{H}_\mathrm{IPM} \cdot \mathbf{q}_{\mathrm{P}i}, \forall i \qquad (3)$$

$$\mathbf{H}_\mathrm{IPM} = \mathbf{H}_\mathrm{RI} \cdot \mathbf{H}_\mathrm{RP}^{-1} \ . \qquad (4)$$

The homography $\mathbf{H}_\mathrm{IPM}$ is often referred to as inverse perspective mapping [1], [2], [11].

Once the matrix $\mathbf{H}_\mathrm{IPM}$ is known, the orthogonal view $\mathbf{I_{orth}}$ is easily recovered from a given perspective image $\mathbf{I_{persp}}$ as follows:

$$\mathbf{I_{orth}}(\mathbf{q}) = \mathbf{I_{persp}}(\mathbf{H}_\mathrm{IPM}^{-1} \cdot \mathbf{q}), \forall \mathbf{q} \in \mathbf{I_{orth}} \ . \qquad (5)$$

There are many ways for recovering the matrix $\mathbf{H}_\mathrm{IPM}$. The simplest one is to manually locate four known points in the perspective image, and to recover the unique mapping as the solution of a homogeneous linear system [12]. We have established an involved but somewhat more practical method whereby it suffices to select the edges of a straight road ahead, which is similar in spirit to what has been proposed in [14]. However, the following two assumptions need to hold: (i) that the internal camera parameters are known [15], and (ii) that the roll of the camera with respect to the road plane is negligible. The matrix $\mathbf{H}_\mathrm{IPM}$ can be calibrated beforehand (this is our current practice) [16], or continuously adapted to the dynamics of the vehicle motion by a suitable optimization procedure [14], [17].

The matrix $\mathbf{H}_\mathrm{IPM}$ could also be recovered by determining the appropriate motion between the physical camera and the virtual camera corresponding to the orthogonal view, using the following equation [13].

$$\mathbf{H}_\mathrm{IPM} = \mathbf{K}_\mathrm{C} \cdot (\mathbf{R} + \frac{\mathbf{T}\mathbf{n}^T}{d}) \cdot \mathbf{K}_\mathrm{C}^{-1} \ . \qquad (6)$$

In the above equation, $\mathbf{K}_\mathrm{C}$ denotes intrinsic camera parameters [15], $\mathbf{R}$ and $\mathbf{T}$ rotation and translation from the physical to the virtual camera, respectively, while $\mathbf{n}$ and d denote the normal of the plane and its distance in the coordinates of the physical camera.

In practice, the assumption about local planarity of the road surface holds for short parts of the road, because extreme slope changes would be dangerous and are hence avoided in road construction. Nevertheless, the vertical orientation of the camera (the tilt angle) can vary slightly due to vehicle dynamics. Even slight errors in determined orientation would produce large errors in the appearance of the parts of the road that are far from the camera. Fortunately, there is little pixel data for those parts of the road, so they must be ignored in any case. For these reasons, we can safely assume the plane of the road is constant throughout the video sequence.

IV. Obtaining the road appearance mosaic

The system we propose has been developed and tested on a subset of a large collection of videos obtained from a vehicle driving the countryside, suburbs and small cities in Croatia [4]. The vehicle is equipped with a single top-mounted camera, an odometer and a GPS sensor (cf. figure 4). Hence, all obtained videos are georeferenced. Additionally, all sensor inputs are synchronized

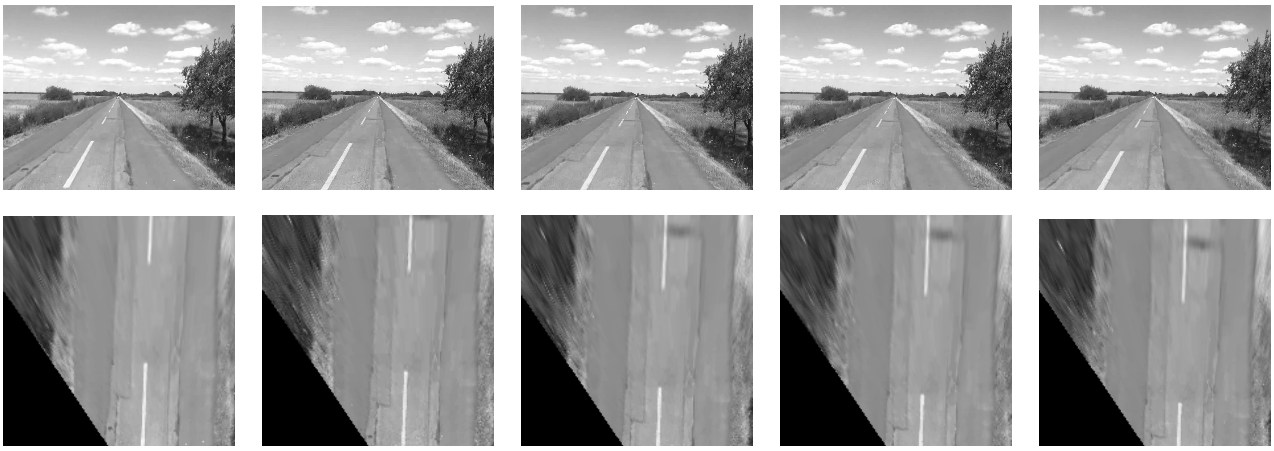

Fig. 2. Five consecutive frames in a video of a straight road (top) and the corresponding orthogonal views (bottom).

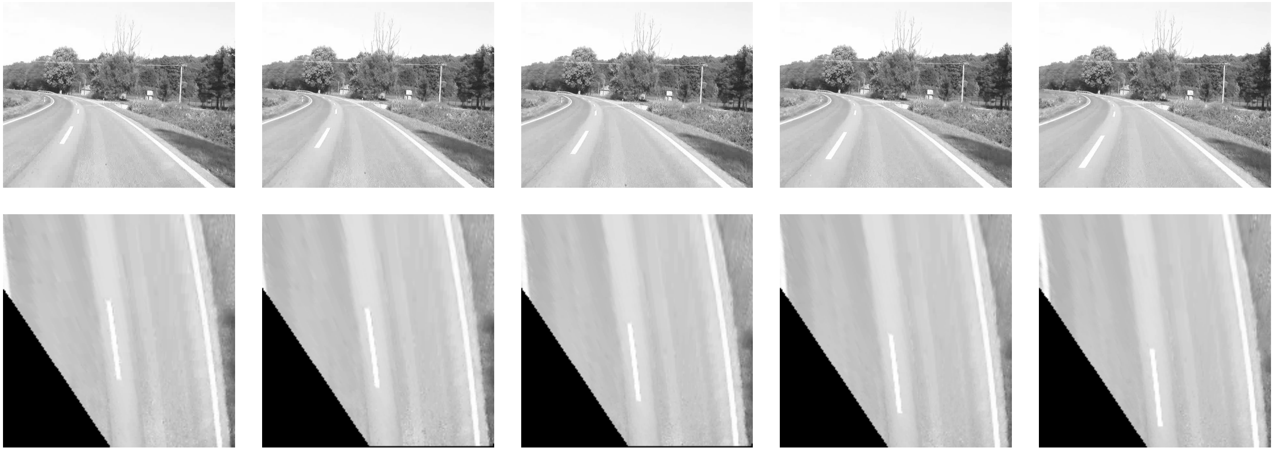

Fig. 3. Five consecutive frames in a video of a curved road (top) and the corresponding orthogonal views (bottom).

with respect to the common clock. However, in this stage of our work we rely exclusively on the video data and discard the information obtained by GPS and odometer.

In the presented system the surface appearance map is obtained by combining two techniques: (i) inverse perspective mapping and (ii) similarity based image alignment. Each frame of the video is first transformed into an orthogonal perspective. Subsequent frames are then joined into a map by examining road pixel data.

V. Aligning orthogonal images

By using inverse perspective mapping, we obtain an orthogonal image for every frame of the video sequence. A single part of the road is represented in multiple consecutive images, as shown in figures 2 and 3. Notice that even though only a small part of the scene is visible in orthogonal images, they still share vast amount of common pixel data. This can be used to determine the relative position of orthogonal images in a plane, which enables us the construct the comprehensive road appearance mosaic. The quality of the pixel data of the common road part in subsequent images is not the same. If the vehicle is moving forward, then the corresponding part of the road has moved closer to the camera, and has better resolution in newer frames. For that reason, newer frames overwrite the pixel data of older ones.

Fig. 4. The vehicle used for acquisition of road videos, equipped with a single camera, a GPS receiver and an odometer. The videos are geo-referenced using an on-board computer equipped with a geoinformation system.

Since we assume the road is locally planar, parts of the road appearing in two subsequent orthogonal images can be related by a transformation matrix $\mathbf{T}$ with 4 degrees of freedom: translation along the line of movement (i.e. direction of road), slight translation orthogonally to the line of movement, rotation along a vertical axis, and scaling (which approximates the effects of vehicle tilting). Finding these parameters is an optimization problem. We use sum of squares of differences of corresponding road pixel intensities as the objective function. We denote the orthogonal image that we are trying to place as $\mathbf{I_{cur}}$, its predecessor as $\mathbf{I_{prev}}$ and the pixels of the predecessor as $\mathbf{q}_i$. The objective function is:

$$F(\mathbf{T}) = \sum_i \left(\mathbf{I_{prev}}(\mathbf{q}_i) - \mathbf{I_{cur}}(\mathbf{T} \cdot \mathbf{q}_i)\right)^2 \tag{7}$$

It is important to note that we consider only pixels $\mathbf{q}_i$ for which intensity of their mapping $\mathbf{I_{cur}}(\mathbf{T} \cdot \mathbf{q}_i)$ is defined (falls within the current orthogonal image). If these pixels form less than 30% of the total number of pixels in the image, than the value of the function is set to infinity instead. This avoids the trivial case of transformations which result in zero overlap of the images. The threshold of 30% was chosen by ad-hoc testing, and invites futher discussion.

The transformation matrix $\mathbf{T}$ is obtained by solving:

$$\mathbf{T} = \arg\min_{T}\{F(\mathbf{T})\} \tag{8}$$

Our system currently uses exhaustive searching to solve this problem. It would perform poorly if it were to search the entire state space (because evaluation of the objective function is time consuming). Additional knowledge is required to reduce the search space of the parameters. A simple model of vehicle motion is used to constrain the translation and rotation parameters, since the vehicle's ability to change velocity and direction is limited. The perceived change of scale parameter has been low in considered videos, so we chose to ignore it at this stage. The obtained transformation matrix $\mathbf{T}$ is used to generate the road surface appearance map in the following way:

1) Start with an empty road mosaic and set the orthogonal image of the first video frame as the current orthogonal image.
2) Place the current orthogonal image at the origin of the road mosaic image.
3) If the current image has no successor, end.
4) Obtain the tranformation matrix $\mathbf{T}$ for the current orthogonal image and its successor.
5) Apply this transformation to the successor image.
6) Combine the current and the successor image, store the result as new current image, and repeat from step 2.

VI. Results and conclusion

The proposed approach yields encouraging results under some constraints. The image alignment performs well if the movement of the vehicle is reasonably smooth. Good results are obtained if there is little to no rotation of the vehicle (cf. figure 5). Shadows, patches of newer asphalt and surface markings ensure good convergence of the image alignment algorithm, while bad convergence has been occasionally noted on textureless parts of the road. Those parts of the road, however, are the least interesting, because absence of texture usually indicates there are no defects in road surface. Due to heavy computational load of the exhaustive search, the system performs at about 10 seconds per frame on a modern machine.

VII. Future work

We plan to derive additional constraints from the use of motion sensors, such as GPS. We have developed a working visual odometry subsystem which has not yet been integrated with the road mapping framework. We have promising results with using steerable filters to obtain road lane lines. The road lane lines can be used to recover initial approximations of the alignment transformation parameters. If any lane line is present, we can calculate the rotation parameter. Dashed lane line can be used to impose constraints on translation. Road detection simplifies removal of non-road pixels, which would speed up the evaluation of the objective function, by taking into account only pixels projected from the road plane. More advanced modelling of vehicle movement would be very useful. It would reduce the state space for alignment optimization process, and it could be used to detect errors occurring in other subsystems (such as visual odometry).

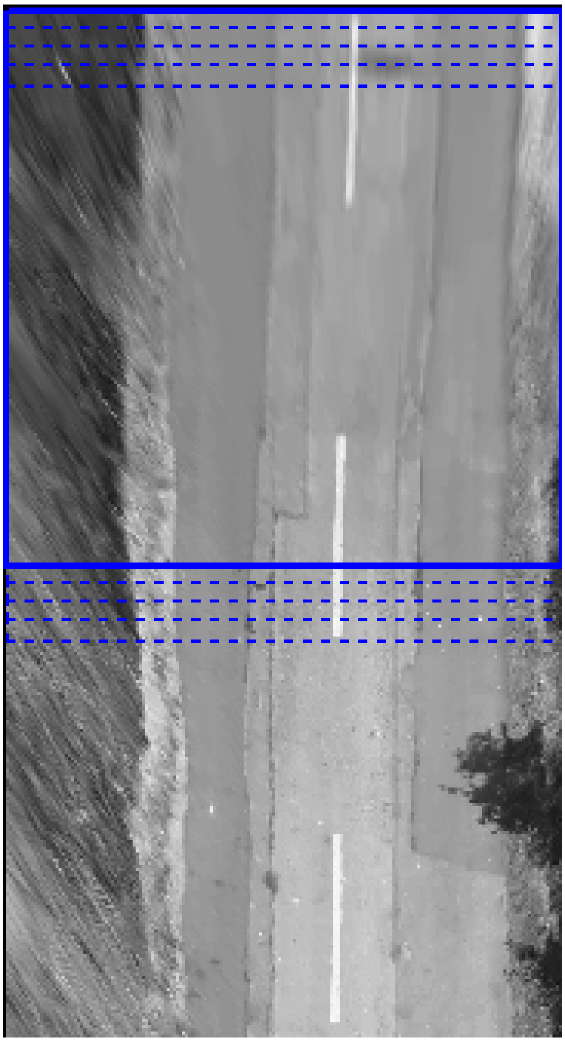

Fig. 5. The obtained road appearance mosaic. The locations of the last five orthogonal images are highlighted.

References

[1] A. Guiducci, "3d road reconstruction from a single view," *Computer Vision and Image Understanding*, vol. 70, no. 2, pp. 212–226, 1998.

[2] J. McCall and M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, pp. 20–37, Mar. 2006.

[3] N. Simond, "Reconstruction of the road plane with an embedded stereo-rig in urban environments," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, (Tokyo, Japan), pp. 70–75, June 2006.

[4] S. Šegvić, K. Brkić, Z. Kalafatić, V. Stanisavljević, D. Budimir, and I. Dadić, "Towards automatic assessment and mapping of traffic infrastructure by adding vision capabilities to a geoinformation inventory," in *Proceedings of MIPRO'09*, pp. 276–281, May 2009.

[5] Y. Wang, L. Bai, and M. Fairhurst, “Robust road modeling and tracking using condensation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, pp. 570–579, Dec. 2008.

[6] C. Wang, T. Hassan, N. El Sheimy, and M. Lavigne, “Automatic road vector extraction for mobile mapping systems,” in *Proc. of ISPRS*, 2008.

[7] Y. Shi, R. Shibasaki, and Z. Shi, “Towards automatic road mapping by fusing vehicle-borne multi-sensor data,” in *Proc. of ISPRS*, 2008.

[8] J. K. G. Hunter, C. Cox, “Development of a commercial laser scanning mobile mapping system - streetmapper,” in *Proceedings of 2nd International Workshop of the future of Remote Sensing*, 2006.

[9] J. Talaya, E. Bosch, R. Alamus, E. Bosch, A. Serra, and A. Baron, “Geomobil: the mobile mapping system from the icc,” in *Proceedings of 4th International Symposium on Mobile Mapping Technology*, 2004.

[10] O. Barinova, R. Shapovalov, S. Sudakov, A. Velizhev, and A. Konushin, “Efficient road mapping via interactive image segmentation,” in *Proc. of CMRT*, 2009.

[11] S. Tan, J. Dale, A. Anderson, and A. Johnston, “Inverse perspective mapping and optic flow: A calibration method and a quantitative analysis,” *Image and Vision Computing*, vol. 24, no. 2, pp. 153 – 165, 2006.

[12] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[13] Y. Ma, S. Soatto, J. Košecká, and S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. New York, USA: Springer-Verlag, 2004.

[14] M. Nieto, L. Salgado, F. Jaureguizar, and J. Cabrera, “Stabilization of inverse perspective mapping images based on robust vanishing point estimation,” in *Proceedings of the IEEE Intelligent Vehicles Symposium*, (Istanbul,Turkey), pp. 315–320, June 2007.

[15] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern recognition and Machine Intelligence*, vol. 22, pp. 1330–1334, Nov. 2000.

[16] A. Guiducci, “Camera calibration for road applications,” *Computer Vision and Image Understanding*, vol. 79, no. 2, pp. 250–266, 2000.

[17] A. Catala Prat, J. Rataj, and R. Reulke, “Self-calibration system for the orientation of a vehicle camera,” in *Proceedings of the ISPRS Commission V Symposium Image Engineering and Vision Metrology*, (Dresden, Germany), pp. 68–73, Sept. 2006.