Small Object Detection Based on Stereo Vision*

Kazuhisa ISHIMARU Long QIAN Seiichi MITA

Noriaki SHIRAI

Small size objects which dimensions are around 0.15m are one of the major security risks to driving vehicles in the highway. LIDAR and radar are hard to detect this kind of objects due to the sparsity of their detecting signal. Vision based methods are possible to solve this problem because camera can generate dense information. We propose a new method to detect small objects in the highway based on stereo vision. This method uses Multi-Path-Viterbi algorithm to obtain dense depth information of stereo images. Based on the depth information, road surface can be detected. Objects on road can be mapped to the 3D space to determine their size and location, then small objects dangerous to the host Vehicle can be recognized and located.

Key words :

stereo matching, Multi-Path-Viterbi, obstacle detection, autonomous driving

1. Introduction

Sensing the environment around the host vehicle is an essential task of autonomous driving systems and advanced driver assistance systems. Detecting small objects on road is one of the environment sensing tasks which LIDAR has some fundamental difficulties to handle. For example, assume a typical LIDAR, Velodyne HDL-32E, installed on the top of a vehicle with 2.4 meter high. According to the specification of HDL-32E, the angle between two laser lines in vertical direction is 1.2919 degree. Generally we define an object as obstacle only if at least 2 laser lines touch it. Then at about 6m distance, we can easily calculate that the minimum detectable height of obstacle is 0.154m. It means that an object which is smaller than this dimension may be missed by HDL-32E. However, these kind of objects are still dangerous to vehicles driving on the highway and must be detected. On the contrary, stereo vision can generate dense depth information as well as color information, which help to detect very small objects on road easily.

In this paper, we demonstrate a small object detection method based on our novel stereo matching algorithm. This method mainly includes three steps: stereo matching, road surface extraction and small objects detection. (i) The stereo matching algorithm is based on a hierarchical bi-direction Viterbi process constrained by Total Variation (TV)[1] and paired with an auto rectification framework based on optical flow. This algorithm can generate dense depth information in real-time for 3D environment sensing tasks. (ii) The road surface extraction algorithm is mainly based on v-disparity[2]. After having the disparity

＊ Aachhen Colloquium 2015 事務局の了解を得て，一部を加筆して転載

map generated by stereo matching, we calculate the histogram at every columns and rows of the disparity map to transform it to v-disparity map. Then we use Radon transform to detect straight lines in v-disparity space. These lines will be treated as straight roads or curbs after transforming back to image space. For curved roads or curbs, we use Viterbi algorithm to find the optimal path in v-disparity space. (iii) With the detailed road information, we can extract road surface out of the disparity map. The remaining parts within road area can be considered as obstacles. Based on some simple thresholds for volumes and positions after transforming the disparity to depth, the small objects on road can be detected finally. Compared with other object detection methods[3)4)5)] based on stereo vision, our method has the following benefits: (i) Our stereo matching algorithm is not based on image segmentation. Since small objects and complicate scenarios are hardly well segmented by current image segmentation methods, our algorithm is sensitive to edge and has good performance for small objects on road as well as low curbs. (ii) Unlike some segmentation-based or global-optimization-based methods, the running time of our algorithm is not influenced by the image content. For any 640×480 images with maximum 40 disparities, the running time is below 100ms with GTX TITAN GPU and Xeon E5-2620 CPU. This feature is helpful for process scheduling of real-time operating system and data synchronization of multi sensors as well as hardware implementation. (iii) Most stereo matching algorithms are designed for controlled structure environments. In the driving vehicle, luminance variation and nonlinear epipolar line distortion caused by windshield and vibration are two main factors which can degrade the stereo matching performance significantly. Our algorithm uses structural similarity[6)] to improve the robustness to the luminance variation and uses auto-rectification to tolerate the epipolar line distortion. (iv) Many stereo matching algorithms cannot generate dense disparity for non-textured area. For example, Semi Global Matching(SGM)[7)] often makes the oversaturated road area in disparity map have irregular shape and many “holes”. On the contrary, our method combines both Viterbi and TV to generate full dense disparity map. It is approximately equivalent to use a slanted plane to fit the hole at the non-textured area on road, which helps keeping the shape of the road in the disparity map and detecting small objects on the road.

2. Methods

The diagram of our method is shown in **Fig. 1**. A stereo camera captures a pair of left and right images and sends to computer. Our stereo algorithm called Multi-Path-Viterbi (MPV) performs the correspondence calculation between left and right images for every pixel. The horizontal shifting between the correspondence pixels is stored as a disparity map. Then a histogram transform is applied to both horizontal and vertical direction of the disparity map. The results of histogram transform are called v-disparity. Viterbi processes are applied to v-disparity maps to find the road surface and curbs. After extracting the pixels belonging to the road surface from disparity map, the remained pixels inside road area are projected into 3D space and form a 3D point cloud. Then the small objects on road can be found in the 3D point cloud. In the following sections, we explain every parts in detail.

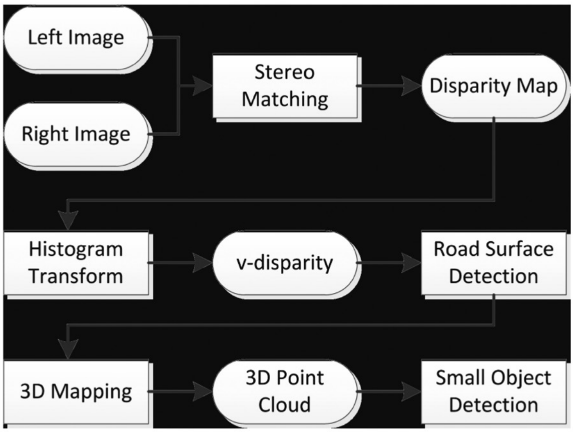

Fig. 1 Algorithm diagram for small object detection

2.1 Stereo Matching

Our novel stereo matching method is called Multi-Path-Viterbi (MPV) algorithm[8], which mainly include two parts. the first part estimates disparity by a Viterbi process[9] and the second part estimates epipolar line distortion by a convex optimization process[10]. Two parts are combined into an online framework to do stereo matching and auto-rectification simultaneously in real-time.

In the first part, the MPV algorithm employs the TV constraint in the Viterbi process as the regularization term. As the matching cost is accumulated in the Viterbi paths with the constraint of TV term, the results of Viterbi process is approximately equivalent to a full 2D convex optimization with TV term. Because TV constraint is applied to all the 4 paths independently, 3D planes at different orientations can be approximately modeled by at least one path. Therefore it can model the 3D objects with one or multiple slanted planes. The TV constraint is useful to smooth some non-textured areas such as road or car body which are common in driving scenes but hard for stereo matching algorithms. Besides that, we also use the intensity gradient information to control the regularization level of the TV constraint and make edges to be sharper. The TV constraint is expressed by defining the energy $E(u)$ on the disparity map u as follows:

$$E(u) = \sum_{p} SSIM(p,u) + \sum_{p' \in L_p} \epsilon_{(p',u') \to (p,u)}$$
$$\epsilon_{(p',u') \to (p,u)} = \lambda e^{-|G|} |u - u'| \qquad (1)$$

where p means pixel, $SSIM(p,u)$ is the structural similarity cost[8] between the pixel p in the left image and pixel p-u in the right image, ε is the TV constraint modified by the gradient information G of the left image, L_p is the neighborhood of p, and λ is the parameter to control the smoothness of disparity map. A 4 bi-directional (horizontal, vertical, and 2 diagonals) Viterbi processes can be used to solve Eq.(1) approximately. We first run the left and right Viterbi processes independently for every scan-line. Then we select the smaller cost of the two parts as the final cost for every possible disparity. Then the up and down Viteribi processes are applied on the previous results for every vertical line. The results of up and down Viterbi are merged by an average to remove noises. Then other diagonal Viterbi processes are performed. The results of the final Viterbi process are the disparities.

In the second part, we apply a process similar as optical flow to remove the possible distortion between left and right images. Because of the imperfect calibration or rectification, influence of windshield or driving, and hardware installation etc., there are always some small distortions between real stereo image pairs and ideal rectified images. We treat this small distortions as small shifting between the correspondence pixels between left and right images. Because we can build the correspondence by previous Viterbi methods, the shifting can be thought as the optical flow[11] and can be solved by convex optimization method[10]. The optical flow equation is:

$$\hat{v} =_{v} \int_{\mathbb{R}^2} |\bar{I}_1(x, y+v) - I_0(x,y)|^2 + \lambda |\nabla v|^2 \, \mathrm{d}x \, \mathrm{d}y \qquad (2)$$

where v is the shifting between the correspondence pixels, I_0 and I_1 are left and right images, $\bar{I}_1$ is warped right image according to the correspondence obtained by Viterbi method, |·| is L_2 norm and λ is the parameter to control the smoothness of v. Here v only contain vertical component because the horizontal component is much smaller than disparities and discarded.

The problem can be solved according to the convex optimization theory[8] and V can be calculated for any specific frame. We can calculate V for several continuous frames and all the results should be approximately the same according to assumption[1]. Therefore we can distinguish outliers and average all the inliers to improve the robustness. This process does not need to be run in real-time for all frames. Generally running once for several hundred frames is enough to follow the changing of V. After a V matrix is estimated, it can be used to compensate the next hundreds of images obtained by stereo cameras.

2.2 Road Surface Detection

After having the disparity map U of image, we calculate the histogram of disparity in horizontal direction and vertical direction. Let H be a histogram transform to the image I such that $H_x\{I\}$ accumulates the points with the same value that occur on a given horizontal image line and $H_y\{I\}$ for the vertical image line. Let (x,y) denote the image coordination of a pixel in U and (i, j) denote the abscissa and ordinate of a pixel in $H\{U\}$. Then $H_x\{U\}(i, j)$ corresponds to the number of points with same disparity as i at the horizontal image line j in the disparity map U:

$$\mathcal{H}_x\{U\}(i,j) = \sum_{(x,y)\in U} \delta_{y,j}\delta_{u_{(x,y)},i} \tag{3}$$

where δ denotes the Kronecker delta. Similarly, for any pixel (i,j) in $H_y\{U\}$:

$$\mathcal{H}_y\{U\}(i,j) = \sum_{(x,y)\in U} \delta_{x,i}\delta_{u_{(x,y)},j} \tag{4}$$

$H_x\{U\}$ and $H_y\{U\}$ are also known as v-disparity and u-disparity [12][2]. They can be thought as the deformed left view and top view of 3D scene intuitively.

Generally, we can detect the road surface in $H_x\{U\}$ and detect the curb or obstacle in $H_y\{U\}$. To improve the robustness of curb detection, we apply the following filtering techniques to $H_y\{U\}$ with the knowledge of road surface in $H_x\{U\}$. After obtaining the road surface in $H_x\{U\}$, we can have the ordinate h_u of road surface corresponding to the disparity u in $H_x\{U\}$, which represents the vertical 3D coordination of road surface is h_uB/u according to after-mentioned Eq.(8). Let H denote the curb or obstacle's maximum physical height. Then its minimal image coordination is $\left(\frac{h_uB}{u} - H\right)u/B$. It means the image coordination of curb and obstacle is in interval $(h_u - \frac{uH}{B}, h_u)$. Because we only pay attention to these parts in $H_y\{U\}$, we can change Eq.(4) about $H_y\{U\}$ to:

$$\mathcal{H}_y\{U\}(i,j) = \sum_{(x,y)\in U} \delta_{x,i}\delta_{u_{(x,y)},j}[h_u - \frac{Hu}{B} \leqslant y < h_u] \tag{5}$$

where [ ] means Iverson bracket. After applying this modification, the noise in $H_y\{U\}$ will be greatly reduced.

For curved roads or curbs, we build Viterbi searching space in the $H\{U\}$ directly. We treat every pixel (i, j) in the $H\{U\}$ as a Viterbi node and the Viterbi process accumulates the value of each node to find a continuous path $j=P(i)$ which has the maximum sum of value at proper straightness constraints. For road detection, the Viterbi equation is:

$$\hat{P} = \arg\max_{P} e(i, P(i)) \tag{6}$$

According to Viterbi algorithm, we have:

$$e(i,j) = \begin{cases} \mathcal{H}_x\{U\}(i,j) + \max\limits_{0<j-j'<\eta} e(i-1,j') & \text{for road surface} \\ \mathcal{H}_y\{U\}(i,j) + \max\limits_{0<i'-i<\eta} e(i',j-1) & \text{for left curb} \\ \mathcal{H}_y\{U\}(i,j) + \max\limits_{0<i-i'<\eta} e(i',j-1) & \text{for right curb} \end{cases} \tag{7}$$

where η is the parameter to control the straightness of roads and curbs.

2.3 Small Object Detection

we can map every points in the G to 3D space with the camera parameters. Let (x,y) denote image coordination of a pixel in the image and u denote its disparity value such that $u=U(x,y)$. Assume its 3D coordination is (X,Y,Z). Given the focal length f and base line length B of calibrated pinhole stereo camera, we have:

$$u/B = f/Z = x/X = y/Y \tag{8}$$

according to the geometry of stereo vision.
Suppose the height of the small object is S_h and the equation of the road surface is $ax+by+cz=d$. Then we can classify a pixel (x,y) as a part of a small object on road if and only if:

$$(x,y) \in G \ \wedge \ 0 < \frac{axB + byB + cfB - du}{\sqrt{a^2u^2 + b^2u^2 + c^2u^2}} < S_h \tag{9}$$

3. Experimental Setup

We evaluated our method on our experimental autonomous car with the stereo camera installed outside as shown in **Fig. 2**. The selected stereo camera is Bumblebee BBX3-13S2C-38. Our algorithm has very few parameters and the parameters except maximum disparities do not need to be changed for almost all normal scenarios. This is one of merits of our algorithm. In all of the following experiments, we set the window size of SSIM as 5×5 pixels and other parameters of SSIM as its original paper. The weight of TV constraint is set to 10 and the maximum disparity is set according to the scenarios. The parameters of SGBM[7] and ELAS[13] are set according to KITTI website[14] .

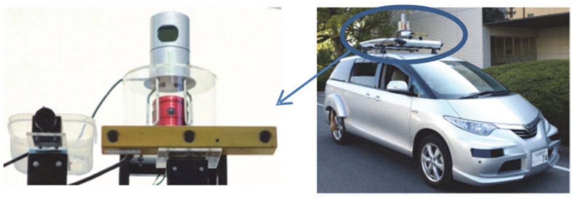

Fig. 2 experimental car and the stereo camera

4. Results

We first tested the performance of our stereo matching algorithm. We did an objective evaluation at the KITTI datasets[14] . We use the KITTI training dataset which includes total 194 images and use the development kit in KITTI website to do the evaluation. The error rates for every image compared with SGBM and ELAS can be found at **Fig. 3**. Our method has 7.38% average error rate compared to SGBM's 12.88% and ELAS's 11.99%.

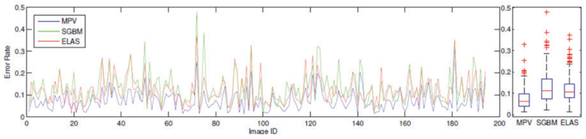

Fig. 3 KITTI Result

We also tested the small object detection performance for our MPV method, SGBM, Velodyne HDL-32E LIDAR, and Ibeo LIDAR. Velodyne has 32 laser layers and installed on the top of vehicle. Ibeo has 4 layers and installed at the most front of vehicle. The size and position of every objects are shown in **Fig. 4**. We first calculated disparity map by our MPV algorithm and SGBM algorithm respectively as shown in **Fig. 5** and **Fig. 6**. Then we translated the disparity map to 3D point cloud by *Eq*.(8). LIDAR can generate 3D point cloud directly. After having the 3D point cloud, we detected small objects on road based on *Eq*.(9). **Fig. 7** shows the comparison between Velodyne HDL-32E and IBEO LIDARs. The position of host vehicle is the origin of the 3D coordinate system. The figure shows Velodyne and Ibeo can only detect 2 (red circle) and 1 (red cross), just as we have analyzed in the introduction part. We also compared our method to SGBM as shown in **Fig. 8** where black dots are MPV results and green plus signs are SGBM results. From **Fig. 8**, we can see both MPV and SGBM can detect all the 4 objects. However, SGBM results have much more noises than MPV results, in the road surface area. It is reasonable because the road area of MPV disparity map is less noisy than the road area of SGBM disparity map.

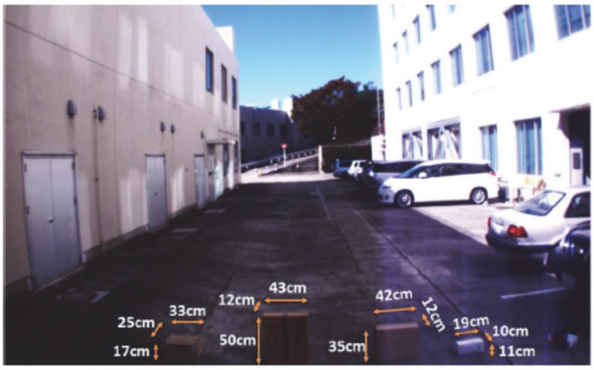

Fig. 4 Experimental small objects on road

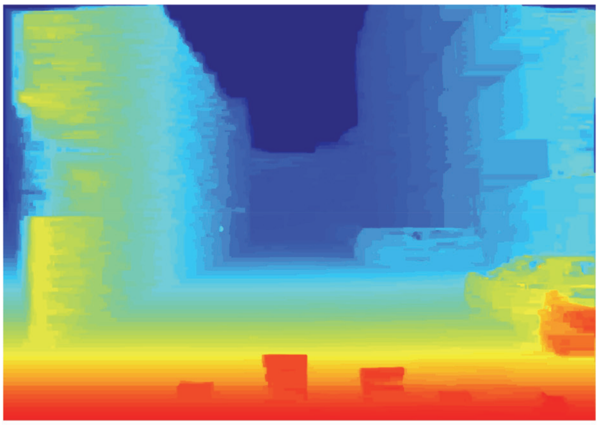

Fig. 5 MPV disparity map

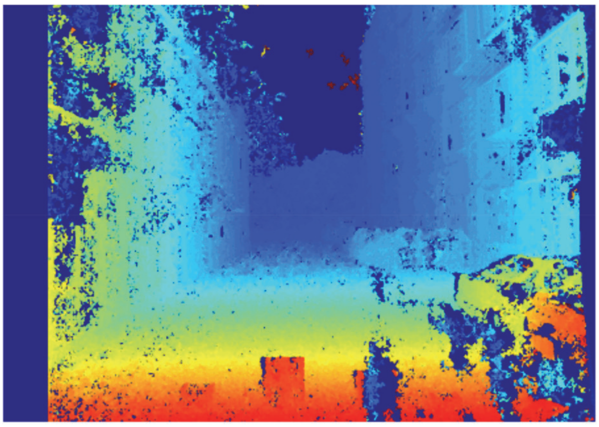

Fig. 6 SGBM disparity map

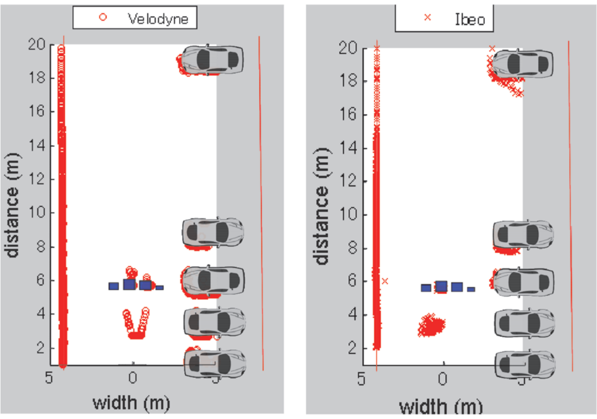

Fig. 7 LIDAR results

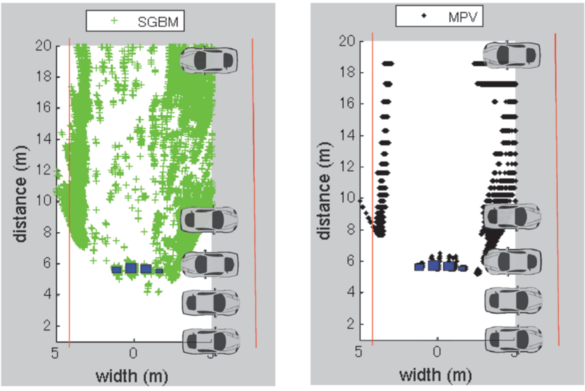

Fig. 8 Stereo results

5. Conclusion

In this paper, a new obstacle detection method based on stereo matching is described. The main benefit of the method is that it can detect very small on-road objects, which is very hard or impossible for normal LIDARs because of their sparse essence. The main contribution of this method is the novel stereo matching algorithm which is optimized specially for intelligent vehicles. It can generate denser result, lower error rate and faster speed compared to the stereo matching algorithm widely used in intelligent vehicles society, such as SGBM and other real-time methods. For example, in KITTI training dataset, our stereo matching algorithm has an improvement of 5.5% to SGBM's pixel error rate. We also compared our object detection performance with two typical LIDAR: Velodyne HDL-32E and Ibeo. The result showed that our method can detect very small objects with the dimension of about 10cm, which is hardly detected by LIDAR.

REFERENCES

1) L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D: Nonlinear Phenomena, vol. 60, no. 1, pp. 259-268, 1992

2) R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," in IEEE Intelligent Vehicles Symposium (IV), vol. 2. IEEE, pp. 646-651, 2002,.

3) K. Y. Lee, G. Y. Song, J. M. Park, and J. W. Lee, "Stereo vision enabling fast estimation of free space on traffic roads for autonomous navigation," International Journal of Automotive Technology, vol. 16, no. 1, pp. 107-115, 2015.

4) N. Bernini, M. Bertozzi, L. Castangia, M. Patander, and M. Sabbatelli, "Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey," in Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on. IEEE, pp. 873-878, 2014.

5) A. Broggi, S. Cattani, M. Patander, M. Sabbatelli, and P. Zani, "A full-3D voxel-based dynamic obstacle detection for urban scenario using stereo vision," in 16th International IEEE Conference on Intelligent Transportation Systems, 2013.

6) Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

7) H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 30, no. 2, pp. 328-341, 2008.

8) Q. Long, Q. Xie, S. Mita, H. Tehrani, K. Ishimaru, and C. Guo, "Real-time Dense Disparity Estimation based on Multi-Path Viterbi for Intelligent Vehicle Applications," in British Machine Vision Conference (BMVC), M. Valstar, A. French, and T. Pridmore, Eds. Nottingham, UK: BMVA, p. 141, 2014. [Online]. Available: http://www.bmva.org/bmvc/2014/papers/paper126/index.html

9) G. D. Forney Jr, "The viterbi algorithm," Proceedings of the IEEE, vol. 61, no. 3, pp. 268-278, 1973.

10) S. P. Boyd and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.

11) B. K. P. Horn and B. G. Schunck, "Determining optical flow," Artificial Intelligence, vol. 17, pp. 185-203, 1981.dd

12) Z. Hu and K. Uchimura, "UV-disparity: an efficient algorithm for stereovision based scene analysis," in Intelligent Vehicles Symposium, 2005. Proceedings, IEEE pp. 48-54, 2005.

13) A. Geiger, M. Roser, and R. Urtasun, "Efficient Large-Scale Stereo Matching," in Asian Conference on Computer Vision (ACCV), 2010.

14) A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3354-3361, 2012. [Online]. Available: http://www.cvlibs.net/datasets/kitti/index.php

著者

石丸 和寿

いしまる かずひさ

株式会社日本自動車部品総合研究所
研究２部 21 研究室
前方認識カメラの開発に従事

龍 潜

ロン チャン

株式会社日本自動車部品総合研究所
研究２部 21 研究室　博士（工学）
前方認識カメラの開発に従事

三田 誠一

みた せいいち

豊田工業大学
スマートビークル研究センター
特任教授 博士（工学）
自動運転，信号処理，画像処理の研究に従事

白井 孝昌

しらい のりあき

走行安全技術１部
前方認識カメラの開発に従事