The Fifteenth International Symposium on Artificial Life and Robotics 2010 (AROB 15th '10),
B-Con Plaza, Beppu,Oita, Japan, February 4-6, 2010

Automatic Detection of Pedestrians from Stereo Camera Images

Kazuki Inumaru*[1], Joo Kooi Tan*[1], Seiji Ishikawa*[1] and Takashi Morie*[2]

Department of Mechanical and Control Engineering, Kyushu Institute of Technology[*1]
Department of Brain Science and Engineering, Kyushu Institute of Technology[*2]
{inumaru, etheltan ,ishikawa}@ss10.cntl.kyutech.ac.jp[*1], *morie@brain.kyutech.ac.jp*[*2]

Abstract: we propose a technique for detecting pedestrians employing stereo camera images based on probabilistic voting. From a disparity map, each pixel on the image is voted on a depth map employing a 2-D Gaussian distribution. The region having a peak value of the vote is chosen as the foot of an object. The object is specified by a rectangle on the right image, which is referred to as a region of interest (ROI). This ROI is described by HOG features and it is judged by SVM if it contains a person. With the ROI containing a person, Kalman filter is applied to track the person through successive image frames. Performance of the detection of persons was evaluated employing a ground truth data. The rate of detected persons to the ground truth data, called a recall rate, was 80 %. This is a satisfactory result.

Keywords: Stereo cameras, Pedestrians detection, HOG, Kalman filter.

1. INTRODUCTION

In recent years, intelligence of a car has been advancing, and advanced safe technologies to support driving of a driver using an in-vehicle camera have become much important. Several methods have been put into practical use for controlling a vehicle such as recognizing traffic lanes from a camera, detecting obstacles by radar, etc. In particular, to realize a safe traffic society, various studies and technology development have been enthusiastically performed aiming at protection of pedestrians. As the pedestrian detection method from in-vehicle camera images, Gavrila[1] proposed a matching method employing a hierarchical template of the shape of a pedestrian. But it has a problem that it must perform the matching repeatedly at a rough position on the image. Uchimura [2] uses U-V-disparity and it applies Gabor Filter to an object region to obtain its features and distinguishes them by Support Vector Machine. However, the object detection in the complicated background such as a downtown seems difficult. Other than the use of in-vehicle cameras, Zhao & Thorpe [3] realize pedestrian recognition by a neural network. However, the precision of recognition with this technique is not enough for practical use.

In this paper, we propose a pedestrian detection method using a stereo camera system. Similar technique [1-3] has already been suggested, but our study differs from them in that, after the detection of pedestrians, it aims at judging the degree of risk with each pedestrian and informs a driver in the order of the degree. In this particular paper, however, we concentrate on pedestrian detection: We propose a pedestrian detection method employing a stereo camera system and a voting using Gaussian distribution.

In the following, we present the proposed technique and show experimental results with discussion.

2. PROPOSED TECHNIQUE

2.1 Parallel Stereo Cameras

In this study, we employ a parallel stereo camera system. The two cameras are of the same height with each other. Let us denote the camera lens coordinate system of the left camera and the right camera by O_l-$X_lY_lZ_l$ and O_r-$X_rY_rZ_r$, respectively. The light axes coincide with the Z_l axis and the Z_r axis, respectively, and the X_l axis and the X_r axis are collinear horizontally. A pair of arbitrary corresponding points on the left image and on the right image receives epipolar constraint. In case of parallel stereo, the constraint becomes a horizontal line and we only have to search a partner point on the horizontal constraint line. Let us denote a point on the left image by (u_l,v_l) and its corresponding point by (u_r,v_r) on the right image. Then disparity d is calculated by

$$d = u_l - u_r\,. \qquad (1)$$

The relation between the image coordinate system and camera lens coordinate system *O-XYZ* is given by

$$\begin{aligned} X &= b(u_r - c_u)/d \\ Y &= b(v_r - c_v)/d \\ Z &= bf/d. \end{aligned} \qquad (2)$$

Here f is a focal length and b is a base line length.

2.2 Corresponding Stereo Images

In correspondence point search between two images, we use the technique proposed by Franke & Joos [4]. It performs the correspondence search employing initially a low resolution image, then higher resolution images. This results in the reduction of computational cost to a large extent.

Let us define the right image as a base image and the left image as a reference image. In the first place, Sobel filtering and binarization is applied to the base image to obtain the edge image of the original. By the employment of the pyramidal representation of an image, lower resolution images are produced with the base, reference and the edge image of the base image. Correspondence is searched on the low resolution edge images first. If a pixel on an edge is found on the edge image, its corresponding pixel is specified on the lower resolution base image and its partner is searched on the lower resolution reference image. Normalized correlation is employed for the search, since it is robust to illumination change. Once a corresponding pair of pixels is found, the pixel and its 3-neighbor pixels, i.e., the next column pixel of the same row, the next row pixel of the same column and the next column pixel of the next row, on the higher resolution base image are examined correspondence with the higher resolution reference image. The image coordinate of a corresponding point is computed to the sub-pixel order by use of parabola fitting. This procedure is repeated in turn to higher resolution images to obtain the final solution. **Fig. 1** shows a disparity map of a road image.

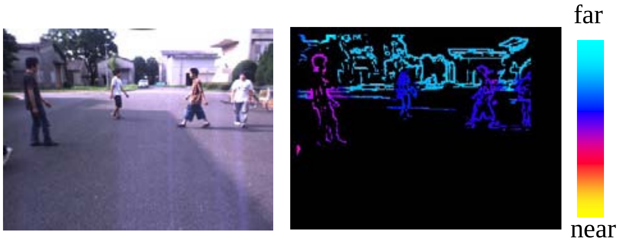
2.3 Object Detection

2.3.1 Depth map

Employing the distance information provided from the disparity image, the 3-D position of the object is specified on the real space, from which the region of the object is identified on the right (base) image. Let us fix the Y coordinate to a certain constant, e.g., Y=0. The XZ plane is then a voting plane which is separated into unit cells. All the pixels on the edge image of the base image receive transform by Eq.(2) and voted on the cells.

2.3.2 2-D Gaussian distribution

The cell where a vertical object exists has a large number of votes. This is, however, not the case in a practical situation, since a longer distance will contain larger errors in parallel stereo. To overcome this difficulty, instead of voting '1' on a certain cell, the voting is done by a 2-D Gaussian distribution whose center is the voted cell.



(a) (b)

Fig.1. Deriving a disparity map: (a) The base (the right) image is given and (b) the depth image is computed from the disparity.

It converts arbitrary point $p(ur, vr; d)$ in provided disparity image into point $\mu(u_r, 0, Z_r)$ in the uZ plane. Here uZ plane is a two-dimensional plane owning horizontal direction u of the image on depth direction Z, the cross axle in a vertical axis. It converts d into Z_r by Eq (2) and v_r is 0, for all pixels p with the same disparity. We vote on the uZ plane with the probability value of the 2-D normal distribution at a provided point $\mu(u_r, 0, Z_r)$. The 2-D Gaussian distribution is defined by

$$f_v(u, Z; \boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{U}^T \Sigma^{-1} \mathbf{U}\right)\right\} \tag{3}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_u \\ \mu_z \end{pmatrix} = \begin{pmatrix} u_r \\ Z_r \end{pmatrix}, \quad \mathbf{U} = \left(u - \mu_u, \quad Z - \mu_z\right), \quad \Sigma = \begin{pmatrix} \sigma_u & 0 \\ 0 & \sigma_z \end{pmatrix}$$

After all voting, we convert the u axial direction into the X-axis direction employing Eq (2). The uZ plane is converted into an XZ plane by this procedure and this yields a depth map.

2.3. 3 Evaluating cell regions and occupation rate

We take a region called a cell region composed of 3×3 cells and evaluate an occupation rate of the cell region. *The occupation rate* is defined as the average of the voted values over the 9 cells. *A depth image* is an image showing the voted values of the cell regions having the occupation rate larger than a threshold. A binarized depth image is called *a grid image*. The depth image receives labeling and clustering and the weighted centroid of each labeled region is calculated on the depth image. The image giving the locations of the centroids is called *a centroid image*. See **Fig. 2** for examples of these images.

2.3. 4 Region of interest

The equation of the road is given by

$$aX + bY + cZ + 1 = 0. \tag{4}$$

Here, a, b, c are road parameters and calculated in

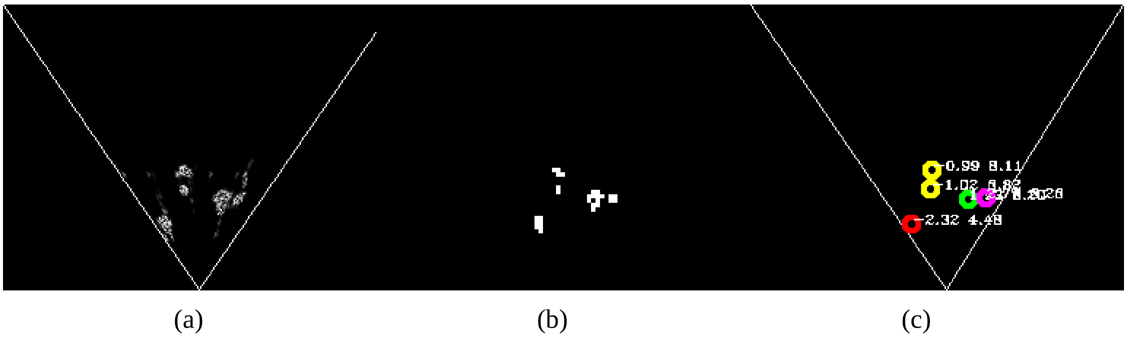


Fig.2. Employed maps: (a) Depth map, (b) grid map, and (c) centroid map.

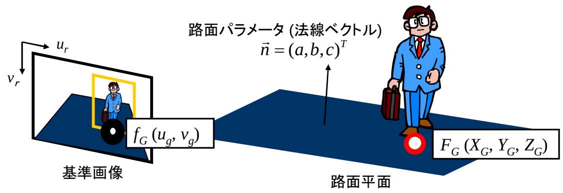


Fig. 3. Diagram of a region of interest.

advance. Using $G(X_G,\ 0,\ Z_G)$ and Eq(3), a contact point $P_G(X_G,\ Y_G,\ Z_G)$ between the road and the object is identified. Y_G is given by

$$Y_G = \frac{-1}{b}\left(aX_G + cZ_G + 1\right). \tag{5}$$

Once $P_G(X_G,\ Y_G,\ Z_G)$ is identified, it is converted by Eq. (2) reversely and the object and its grounding point $p_G\ (u_g,\ v_g)$ is identified so that it may be put on the image of the road. The Region of Interest is expressed with a rectangle, and its size is given by

$$\begin{aligned} Height &= sf\ /\ Z_G \\ Width &= Height\ /\ 2 \end{aligned} \tag{6}$$

Here s is a scale factor and the aspect ratio is fixed to 2:1. **Fig 3** shows a diagram of the ROI.

2.4 Pedestrians Recognition and Tracking

The HOG feature [5] is calculated within the ROI in the frame and the region is judged whether or not it contains a human by the employment of the support vector machine (SVM).

From the position recognized as a pedestrian, we employ Kalman Filter [6] to track it in successive frames.

3. EXPERIMENTAL RESULTS

Experimental results using the proposed method are shown in this section. In experiment 1 (Exp1-1), a stereo camera system is fixed and it take images of the scene where pedestrians pass in front of the cameras, and the scene (Exp1-2) where pedestrians moved around to various orientations. In experiment 2 (Exp2), pedestrians in a zebra crossing are captured images from a moving stereo camera system.

The proposed method was applied to these videos. **Fig.4** shows the results of Exp1-2 and Exp2. Red boxes show the ROIs judged as containing a human, whereas a blue box is the ROI judged as not containing a human. The result of tracking is depicted by respective colors. The size of the processed images is 320*240 pixels. The employed PC is Intel Core2 (CPU:2.39GHz, memory: 4GB).

4. DISCUSSION

4.1 Evaluation Criteria

We consider that not only the result of the detection but also the range of an ROI containing a pedestrian should be evaluated. A detected ROI is therefore compared to ground truth data. It is evaluated what extent of an ROI is included in a ground truth data and, in the same way, what extent of a ground truth data is detected. For this purpose, we define two indices called *cover* and *overlap* given by the following equations.

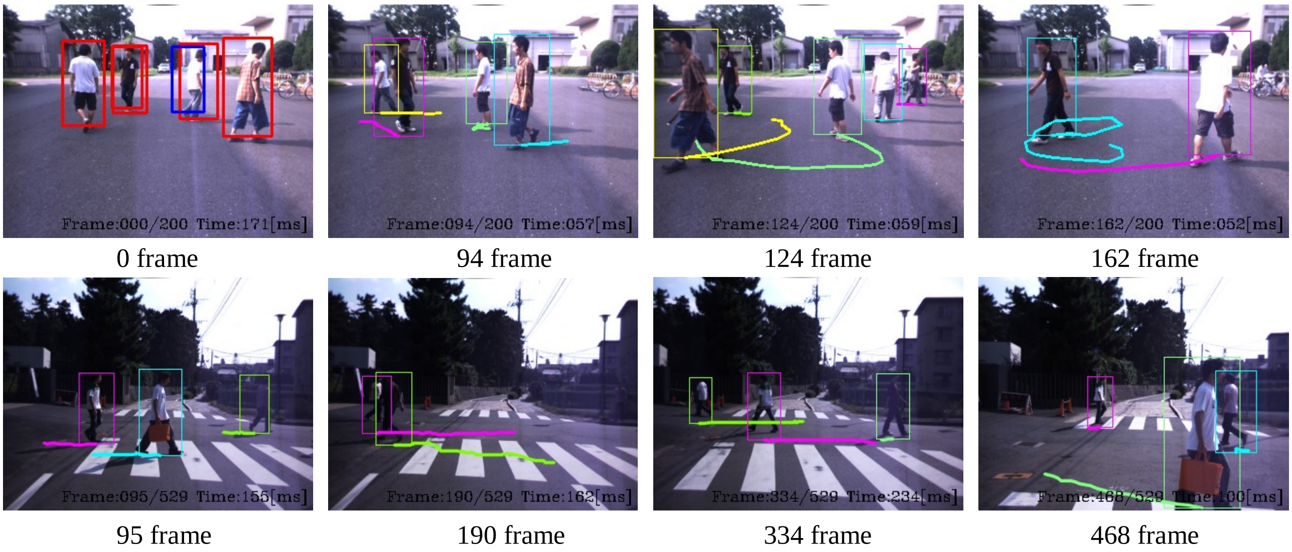
$$\begin{aligned} Cover &= \frac{GT \cap OA}{GT} \\ Overlap &= \frac{GT \cap OA}{OA} \end{aligned} \tag{7}$$

Here GT is the area of a ground truth region and OA is that of a detected ROI. If each value exceeds 0.5, it is judged as detected exactly. **Fig.5** shows a conception diagram of the cover and the overlap.

4.2 Evaluation Results

If we denote the number of all the employed GT data by A, the ratio of the number of correctly detected GT data to A is called *Recall* and is defined by

$$Recall = TP\ /\ A. \tag{8}$$


Frame:000/200 Time:171[ms]
Frame:094/200 Time:057[ms]
Frame:124/200 Time:059[ms]
Frame:162/200 Time:052[ms]
Frame:095/529 Time:155[ms]
Frame:190/529 Time:162[ms]
Frame:334/529 Time:234[ms]
Frame:468/529 Time:100[ms]

0 frame 94 frame 124 frame 162 frame

95 frame 190 frame 334 frame 468 frame

Fig. 4. Experimental results: Pedestrians detection and tracking; a fixed camera case (the upper row) and a mobile camera case (the lower row).

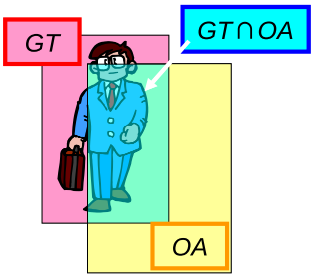


Fig. 5. Diagram of the cover and the overlap.

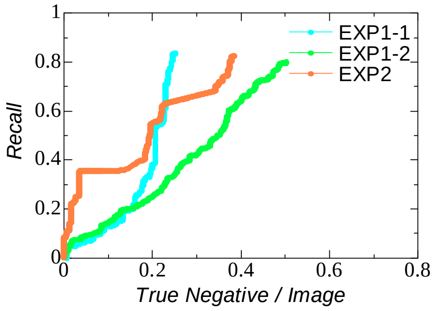


Fig. 6. An evaluation graph.

Here *TP* means true positive. **Fig.6** shows a graph which indicates the relation between the *Recall* values and *TN* values with respect to every experiment. Average processing time with each experiment is 50.5, 61.7 and 150 [ms/frame], respectively.

5. CONCLUSIONS

This paper proposed a method of detecting a pedestrian from a video image taken from a stereo camera system. In order to extract an object from an image, voting of distances of the points detected on the image was done by the employment of a 2-D Gaussian distribution. The detected object region was represented by HOG features and it was judged if the region contains a human by SVM. Experimental results showed satisfactory performance of the method.

The degree of danger with various human activities will be defined and recognized in the next step of this study. To raise the precision of the disparity map and detecting multiple people also remain for further study.

REFERENCES

[1] D. M. Gavrila: “Pedestrian detection form a moving vehicle”, *Proc. European Conf. on Computer Vision*, vol.2, pp.37- 49, 2000.

[2] K. Matsushima, H. Zhencheng, K. Uchimura: “Pedestrian recognition using stereo sensor”, Information Processing Society, Research report, pp.49-54, 2006.

[3] L. Zhao, C. E. Thorpe: “Stereo and neural network based pedestrian detection”, *IEEE Transactions on Intelligent Transportation System*, Vol.1, No.3, pp.148-154, 2000.

[4] U. Franke, A. Joos: “Real time stereo vision for urban traffic scene understanding”, *Proc. IEEE Intelligent Vehicles Symposium* 2000.

[5] N. Dalal, B. Triggs: “Histgrams of oriented gradients for human detection”, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 886-893, 2005.

[6] G. Welch, G. Bishop:“An Introduction to the Kalman Filter”, Technical Report TR95-041, University of North Carolina at Chapel Hill, 1995.