A Pre-Focusing Method For Detecting Pedestrians Running Out Into The Street Using A Mobile Stereo Camera

Takuma Sasaki* and Hiroshi Hanaizumi

Graduate School of Computer and Information Sciences, Hosei University,
Kajino 3-7-2, Koganei, Tokyo 184-8584, Japan

[*takuma.sasaki.3h@stu.hosei.ac.jp](mailto:takuma.sasaki.3h@stu.hosei.ac.jp), hana@hosei.ac.jp

Abstract

For mobile stereo camera systems, constructing three-dimensional (3D) surroundings was an important process to ensure safe navigation; however, it was very time consuming. To decrease the number of traffic accidental deaths, pedestrians entering the street needed to be efficiently detected. Thus, a pre-focusing method (PFM) was proposed to decrease the time required to retrieve matching pixel pairs in stereo pair images. The method PFM provided pre-focused stereo pair images in which an object at a pre-determined distance displays no disparity between images. Pre-focusing was performed by applying a perspective transformation to the image pair, whose coefficients were obtained via a calibration stage of the PFM. An object at the distance was extracted by evaluating the spatial similarity between the pre-focused stereo pair images without scanning in order to find matching point pairs. To avoid errors due to edges and uniform areas, the smooth test, which was a nonparametric test, was introduced to evaluate the spatial similarity. A watching window was introduced to detect the pedestrians. In this study, the method PFM was successfully applied to actual video data captured by a stereo video camera mounted on a bicycle.

Keywords: perspective transformation, smooth test, spatial similarity, nonparametric, histogram

1. Introduction

Various types of autonomous cars are currently being developed and tested on actual roads. They sense their environment to safely reach their destination without human inputs. In the sensing process, sensors such as infrared lasers, cameras, and millimeter-wave radars are used. Computer vision is also used to understand the sensor outputs. In fact, TOYOTA has already developed a practical Pre-Crash Safety System⁽¹⁾. Camera-based systems have also been developed,

i.e., SUBARU's Eyesight and HONDA's Intelligent night vision system⁽²⁾. These systems commonly acquire three-dimensional (3D) environmental road information to control the vehicles without crashing.

The authors believe that another point of view is required in these sensing systems to decrease traffic accident deaths caused by pedestrians entering the street⁽³⁾. Most stereo camera-based sensing systems require a stereo matching process to obtain the 3D environmental road information⁽⁴⁻⁶⁾. The process is, however, very time-consuming. There are at least two possible approaches to solve this problem. One is to use a sub-system, such as GPGPU, to accelerate the processing⁽⁷⁾, and the other is to adopt a smart method to avoid stereo matching. The former is an excellent solution but is very expensive. The latter is a low cost practical solution to detect whether an object exists in the volume of interest; however, it is necessary to develop the method.

In this study, we propose a pre-focusing method (PFM) to detect an object in a 3D region of interest without the stereo matching process, particularly in local mask scanning to find corresponding point pairs. Objects captured in stereo pair images usually have disparities accordingly to their distance from the camera. The stereo matching process calculates the distance via these disparities. In the proposed method PFM, a perspective transformation is applied to one of the stereo pair images to remove the relative disparity for an object at a pre-determined distance. The residual disparities are locally evaluated via an index of the spatial similarity, and high index areas provide detections of an object at a pre-determined distance. The spatial similarity may be affected by shapes of the density histogram in a local area; therefore, we use a nonparametric algorithm to avoid this. To quickly detect pedestrians entering the road, we also introduce a watching window to reduce the processing time.

2. Methodology

The principle of the proposed method PFM is based on the fact that a quadrangle on a plane in a 3D space is perfectly registered onto any quadrangle on another plane in the 3D space. The registration is performed via a perspective transformation whose parameters are determined by coordinate pairs for the four vertices of both quadrangles. After applying the transformation, the object at the pre-determined distance has no positional deviation. Therefore, the object, i.e., a pedestrian, at the chosen distance is detected as an area with no displacement.

2.1 Perspective transformation

Using a perspective transformation, we can map a quadrangle on a plane onto another plane as shown in Fig. 1. The transformation is described as

$$\begin{cases} u = \frac{a_1x + a_2y + a_3}{a_7x + a_8y + 1} \\ v = \frac{a_4x + a_5y + a_6}{a_7x + a_8y + 1} \end{cases} \quad (1)$$

where (x, y) and (u, v) are coordinates on the two planes and a_i ($i = 1-8$) are the coefficients for the transformation. Those coefficients are determined by at least four coordinate pairs, $(x_i, y_i) - (u_i, v_i)$ ($i = 1, \dots, n$). Multiplying the denominator on both sides of Eq. (1), we obtain the matrix equation:

$$U = XA, \quad (2)$$

where

$$U^t = (u_1 \quad v_1 \quad u_2 \quad v_2 \quad \dots \quad u_n \quad v_n), \quad (3)$$

$$A^t = (a_1 \quad a_2 \quad \dots \quad a_8), \quad (4)$$

and

$$X = \begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -u_1x_1 & -u_1y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -v_1x_1 & -v_1y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -u_2x_2 & -u_2y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -v_2x_2 & -v_2y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & 1 & 0 & 0 & 0 & -u_nx_n & -u_ny_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -v_nx_n & -v_ny_n \end{pmatrix}. \quad (5)$$

Equation (2) is solved as

$$A = (X^t X)^{-1} X^t U. \quad (6)$$

In the proposed PFM method, we apply a perspective transformation to one of the stereo pair images so that planar objects at a pre-determined distance have no disparity, as shown in Fig. 2. A checkered pattern is used as the planar object, and its extracted corner points are used to obtain the coefficients of the perspective transformation.

Fig. 1. A perspective transformation.

Fig. 2. Planar objects and a stereo camera.

The pre-focusing is achieved by applying the transformation to one of the stereo pair images. Now, an object at the pre-determined distance has no disparity in the pre-focused stereo pair images. Next, the object at the pre-determined distance is detected by evaluating the local similarity.

2.2 Neyman's smooth test⁽⁸⁾

In the pre-focused stereo pair image, the area with no disparity suggests that there exists an object at the pre-determined distance. To find the area, we locally evaluate the spatial similarity using a small-sized square mask. The similarity can be calculated as the correlation coefficient to fit a straight line to the density scatter-diagram of the pixels in the mask. The correlation coefficient is, however, affected by the distribution of the pixel density, i.e., the shape of the density histogram, in the mask. This requires a unimodal distribution with an adequate dynamic range; however, these conditions are often not satisfied, especially in road scenes.

In the proposed method PFM, we evaluate the spatial similarity by comparing the density histograms between pixels in a local mask on both pre-focused images. The smooth test, in which the goodness of the fit for the mean value and variance of the histogram are tested, is used for the evaluation. The smooth test is a nonparametric test and is based on the fact that a distribution, i.e., a histogram, transferred by a cumulative distribution function forms a uniform distribution. We assume that a distribution A is the

same as another distribution B from which the cumulative distribution function is obtained, when the transferred distribution from A forms a uniform distribution. The first and second moments are used to evaluate the uniformity of the distribution.

Let $f(x)$ and $F(x)$ be a pixel density distribution function in a local mask and its cumulative distribution function, respectively, and let $x_i (i=1, \dots, N)$ be the i -th pixel in the mask. To calculate the first and second moments of the transferred distribution, we define the statistics T and S as

$$T = \sum_{i=1}^N F(x_i) \quad (7)$$

and

$$S = \sum_{i=1}^N \{F(x_i)\}^2. \quad (8)$$

When $F(x_i) (i=1, \dots, N)$ forms a uniform distribution, we obtain the expected values as $E[T] = \frac{N}{2}$ and $E[S] = \frac{N}{3}$. Therefore, the goodness of fit in mean difference is given as $(T - N/2)^2$ and that of the dispersion is given as $(S - N/3)^2$. Therefore, the total goodness of fit is described⁽⁸⁾ as

$$g = 3(2T - N)^2 + 5(6S - 6T + N)^2. \quad (9)$$

As the goodness, g , asymptotically approaches a chi-squared distribution as

$$\frac{g}{N} \sim \chi(2), \quad (10)$$

the threshold for the test is determined by a percent point of the chi-squared distribution. When the g/N value obtained from the pixels in a small mask exceeds the threshold, we recognize the area as a part of an object at the pre-determined distance. The small mask is scanned over the pre-focused stereo pair images skipping the edges in the image. For pedestrians entering the street, there is no need to process all the images. Therefore, we set a Region Of Interest (ROI) with some lattice points in the image and apply the smooth test only to those points. The pedestrian is detected when a minimum fixed number of the points give us a larger g/N value than the threshold.

Fig. 3. A stereo camera used in the experiments.

3. Experimental results and discussion

We used a digital camera, FinePix REAL 3D W3 (Fuji Film), as the stereo camera, as shown in Fig. 3. The camera has a capability to record stereo video data. The image size was 1280×720 . The distance between the two lenses was 75 mm. The camera also had a zooming function. The recorded data were extracted and processed as sequential still images after partitioning.

3.1 Coefficients for the perspective transformation

A large-sized checkerboard plane was observed with a zoom of three times by the stereo camera, which slid on a rail so that its optical axis did not change during the series of observations. The observation distance was changed from 5 to 10 [m] with a 1 [m] step. Figure 4(a) shows the color composite image between the originally observed images at a distance of 10 [m], in which the left image is assigned to red and the right image is assigned to green and

(a) Color composite image of the original observations.

(b) Color composite image with pre-focusing.

Fig. 4. Checkerboard images (red: left image, green and blue: right image).

Table. 1 Coefficients for the perspective transformation

	6 m	8 m	10 m
a_1	1.012	1.009	0.991
a_2	-0.008	-0.009	-0.015
a_3	-4.557	14.633	30.448
a_4	0.013	0.013	0.009
a_5	1.008	1.005	0.992
a_6	-3.621	-5.101	-2.591
a_7	0.000008	0.000007	-0.000004
a_8	0.000002	0.000002	-0.000007

blue. We can see some disparities on the checkerboard. The coefficients of pre-focusing were obtained so that the right image rectified well with the left image. Figure 4(b) shows an example of the rectification in which 70 corner points were extracted and used for the calculation. We can see that the disparities in Fig. 4(a) are removed in Fig. 4(b). Some of the coefficient values are listed in Table 1.

3.2 Extraction of the focused area

Figure 5(a) shows one of the original frames in the video data captured by the bicycle-mounted stereo camera. In the frame, a pedestrian near the left edge of the frame was 10 [m] from the camera. We can see that the object disparities depend on their distances from the camera. The 3D surroundings can be obtained by searching for corresponding point pairs over the image. Figure 5(b)

(a) Original stereo pair image (without focusing).

(b) Focused stereo pair images.

Fig. 5. Color composite images before and after focusing.

(a) Result processed by the correlation coefficient method.

(b) Result processed by the smooth test.

Fig. 6. Extraction of the focused area.

shows the result using pre-focusing at 10 [m]. We can see that only the pedestrian at the pre-determined distance was well rectified while the other objects still had some disparities. To extract the focused area, we evaluated the spatial similarity between the pre-focused stereo pair images. We compared the performance of the smooth test (Fig. 6(b)) with that of the correlation coefficient method (Fig. 6(a)). We used a 5×5 square mask to process both results. The extractions of Figs. 6(a) and 6(b) took approximately 0.4 and 1.4 [s], respectively, using a PC (CPU: i7-4790 with a clock of 3.6 GHz and 32GB memory). The areas were extracted when the correlation coefficient exceeded 0.7 in Fig. 6(a) and when the g/N value was below 100 in Fig. 6(b). Those thresholds were experimentally selected so that the ratio of the correctly extracted areas to the incorrect ones was maximized. Comparing the results, we see that the correlation coefficient method resulted in poor detection performance for the focused pedestrian and larger incorrect areas at the edges and uniform density regions in the scene. Conversely, the smooth test resulted in higher detection performance and fewer errors. The edges in the scene affected both the correlation coefficient and the g value in the smooth test. To reduce the incorrect detection at the edges, we eliminated them from the processing regions. The edges were detected by a Prewitt filter with a threshold of 20. Figure 7 shows the color composite result in which the areas detected by

Fig. 7. Color composite image. Red shows the result via the smooth test, green one via the correlation coefficient, and blue the edges.

(a) Extracted pedestrian at a pre-determined distance.

(b) No pedestrian extracted at the pre-determined distance.

Fig. 8. Eliminating edges when extracting the pedestrian. the smooth test are assigned to red, while those detected by the correlation coefficient method are assigned to green, and the edges are assigned to blue. The white regions in Fig. 7 indicate the edges that have high spatial similarity; therefore, their elimination is expected to reduce the noise in the background. In actuality, the performance when detecting the focused area was significantly improved, as shown in Fig. 8.

3.3 Detection of pedestrian at pre-determined distance

To quantitatively evaluate the performance of the proposed PFM method when detecting a pedestrian entering a street, we took a short video movie using a stereo camera attached to a bicycle. The focused distance was 10 m. For safety, we measured the pedestrian at a distance. We

Fig. 9. A watching window with 15 lattice points.

Fig. 10. The number of lattice points on which an object area was extracted.

introduced a watching window with 15 lattice points and tentatively set the window near the left edge of the scene; however, the position needs to be carefully designated in an actual system. Figure 9 shows the position of the window. The video data were divided into 130 stereo pair images, and each pair was processed by PFM with a 5×5 pixel mask. Figure 10 shows the number of lattice points on which an object area was extracted in each image pair. In the video data, we captured a pedestrian near the end. Therefore, the object area was part of the pedestrian. The result shown in Fig. 10 indicates that we detected the pedestrian by counting the number of lattice points.

3.4 Focusing performance

We captured another pedestrian crossing the street at a distance of 8 m, as shown in Fig. 11. Using the stereo pair image, we evaluated the focusing performance in the proposed PFM method. We applied transformations with focusing coefficients for 5 to 10 m to the image pair and applied the area extraction using the smooth test as in the previous result. The performance was evaluated using the number of lattice points on which an object area was extracted, as shown in Fig. 12. The result indicated that the number of lattice points reached a maximum at the pre-

Fig. 11. A pedestrian at 8 [m] from the camera with the watching window (red shows left image, and green and blue show right image).

Fig. 12. Focusing performance of PFM.

determined distance of 8 [m] and that the in-focus region extended a few meters in distance. The extension of the in-focus region resulted from the small baseline (75 mm) of the stereo camera. A longer baseline will provide higher focusing performance. The peak value of the curve had only 8 of the 15 points because the pedestrian was dressed in dark texture-less clothes and the points on the clothes did not contribute to the detection. Our PFM method does not work well for a pedestrian entirely dressed in black, as is the case with other former 3D matching methods.

4. Conclusions

In this paper, we proposed a pre-focusing method, PFM, to detect pedestrians entering the street using a stereo camera. This method modifies one of the stereo pair images so that no relative disparity is found for an object at a pre-determined distance. The object is detected as an area that has a high spatial similarity between the pre-focused stereo pair images. The similarity was non-parametrically evaluated using the smooth test to avoid errors due to edges and/or constant density areas, such as the road surface. An edge eliminating technique was also used to reduce the

mis-detection of false-positive areas and to achieve a higher signal to noise ratio in the detection. To detect pedestrians entering the street, we introduced a watching window with interior lattice points. The number of lattice points on which an object area was extracted was correlated to the existence of a pedestrian. Omitting the processing area was effective in decreasing the processing time. Future areas of study include improving the accuracy of pedestrian detection by tuning the parameters in the similarity evaluation, designating a prototype system, and evaluating the total performance with regards to pedestrians entering the street.

References

- (1) S. Tokoro, K. Moriizumi, T. Kawasaki, T. Nagao, K. Abe, and K. Fujita: "Sensor Fusion System for Pre-crash Safety System", IEEE Intelligent Vehicles Symposium, pp. 945-950, 2004
- (2) T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka: "Development of Night-Vision System", IEEE Transactions on Intelligent Transportation Systems, Vol. 3, No. 3, pp. 203-209, 2002
- (3) M. Kasahara and H. Hiroshi: "A Dynamic Background Subtraction Method For Detecting Walkers Using Mobile Stereo-Camera", The 2013 International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV, 2013
- (4) L. Zhao and C. E. Thorpe: "Stereo- And Neural Network-Based Pedestrian Detection", IEEE Transactions on Intelligent Transportation Systems, Vol. 1, No. 3, pp. 148-154, 2000
- (5) I. Giosan and S. Nedeveschi: "Shape Improvement Of Traffic Pedestrian Hypotheses By Means Of Stereo-Vision And Superpixels", IEEE Int. Conf. ICCP 2015, pp. 217-222, 2015
- (6) A. Abbas, S. Hoefler, B. Fardi, and G. Wanielik: "Stereo Vision Based Pedestrian Detection Using B-Spline Modeling", IEEE Int. Conf. ICVES 2008, pp. 63-68, 2008
- (7) C. D. Pantilie and S. Nedeveschi: "SORT-SGM: Subpixel Optimized Real-Time Semiglobal Matching for Intelligent Vehicles", IEEE Transactions on Vehicular Technology, Vol. 61, pp. 1032-1042, 2012
- (8) K. Yamagishi, S. S. Fujimura, H. Toyoda, and M. Inamura: "Nonparametric Classification Method for Multispectral Images Based on Smooth Test", Trans. SICE, 20-1, pp. 49-55, 1984