Improved MSER Pedestrian Detection Algorithm based on TOF Camera

ZiJin Wang, YiQin Zhao and CuiLian Zhao

School of Mechatronic Engineering and Automation, ShangHai University, Room 435, Bullding 9, 333 NanChen Road, ShangHai, China

Email:13167167040@163.com

Abstract. To solve the problems existing in the traditional pedestrian detection methods, such as the complexity of pedestrians wearing clothes and the change of light leading to the decrease of detection accuracy, this paper proposes an MSER pedestrian detection algorithm combined with D-NP (Double Nine-Grid). The algorithm utilizes the characteristics of high pixels of TOF (Time Of Flight) cameras and the adaptability of MSER (Maximally Stable Extreme Regions). Firstly, the depth image is subjected to hole patching, and median filtering and morphological operations are used for preprocessing. Then the MSER algorithm combined with D-NP is used to segment the image to extract candidate head regions. Finally, We perform experiments on our dataset, and the detection recognition rate of this algorithm reaches 97.26%, which can effectively reduce the impact of light, and has a good detection performance for pedestrians carrying backpacks, hats and other complex scenes.

1. Introduction

The number of people as an important reference data in the management of public places, its role and value can't be ignored [1]. As a key part of the population statistics, the accuracy of pedestrian detection determines the accuracy of the number of people obtained. However, under the influence of external light, complex background and other factors, the acquired image quality is poor, which further increases the difficulty of pedestrian detection. Therefore, it has become a research hotspot to find a method that can deal with complex external conditions and detect pedestrian targets quickly and accurately.

Zhang et al. [2] proposed the Waterfill algorithm based on the flow characteristics of water in reality. This algorithm simulates the flow direction of water droplets and the control of the stop point of water droplets in the image to obtain the local extremum point representing the head. Xu et al. [3] proposed a pedestrian segmentation method based on watershed algorithm, but it couldn't segment pedestrians well in the case of crowd crowding. Yin et al. [4] proposed a new method to detect pedestrians based on the three-dimensional information of human head and shoulder, and extracted the maximum gray value region of human head by using the nine-grid algorithm. Michael Rauter et al. [5] used local maximum search to determine the candidate region of the head, and used the gradient climbing algorithm to screen the candidate region again to determine the central region of the head. Wu et al. [6] proposed a new regional detection algorithm of maximum stable extremum combined with gradient stratification, and identified the heads concentrated at regional points through the relationship between area and gray mean. MSER algorithm is recognized as one of the better regional detection algorithms in the field of computer vision, which has the characteristics of fast computing speed, affine invariance and stability [7,8]. However, in the actual detection of pedestrians by this algorithm, there are many false positives in the...
extracted candidate head area.

In this paper, the depth images obtained from the top down of the TOF camera are used for detection. The high pixel of the TOF camera is less affected by light changes [9], which can effectively reduce the impact of light changes on pedestrian detection. Through experimental tests, the MSER algorithm combined with D-NP is used to extract the head region of human body, which not only has good real-time performance, but also has high detection accuracy and strong robustness, which has certain application value.

2. Algorithm description

2.1. Deep image preprocessing
In the depth image captured by the TOF camera, the pixel value represents the distance (depth) from the image collector to each point in the scene, which directly reflects the geometry of the visible surface of the scene [10]. According to this principle, a depth threshold is used to remove environmental noise outside human body. Let the original depth image be $D$, pixel value $D(i,j) \in D$, pixel value is:

$$D(i,j) = \begin{cases} D(i,j), & \alpha < D(i,j) < \beta \\ M, & D(i,j) \leq \alpha \\ M, & D(i,j) \geq \beta \end{cases}$$

where $M$ is the maximum depth value, $\alpha$ is the minimum standard depth value, and $\beta$ is the maximum standard depth value.

Through equation (1), all pixel points in the image are screened, and there are almost only pedestrians in the transformed depth image. Then, by extracting the connected domain, the connected domain of small area is removed, and the remaining area is the extracted foreground area. After extracting the foreground region, the image inpainting algorithm [11] was used to repair the empty region, and then the processed image was smoothed through median filtering and closed operation in morphological processing [12].

2.2. MSER algorithm
The maximum stable extreme value region (MSER) can be used in the field of image segmentation. This algorithm was first proposed by J. Matas [13] in 2002 as an affine feature region extraction algorithm. Firstly, different threshold values are used for binarization of gray image. Pixel points with gray value greater than or equal to the threshold value are set as white, and those less than the threshold value are set as black. As the threshold value increases from 0 to 255, the image will form a closed region. If the area of the region does not change more than a certain threshold, the region is considered to be the maximum extreme value stable region. Its mathematical definition is:

$$q(i) = \frac{|Q_{i+\Delta} - Q_{i-\Delta}|}{|Q_i|}$$

where $Q_i$ refers to a connected region when the threshold value is $i$, $\Delta$ is the small change of gray threshold value, and $q(i)$ is the rate of change of region $Q_i$ when the threshold value is $i$. When $q(i)$ is a local minimum, $Q_i$ is the maximum stable extreme region.

2.3. MSER algorithm combined with D-NP
The MSER algorithm is used to detect the maximum stable extreme area of the extracted foreground area, forming a series of maximum extreme stable areas that change according to the threshold. Although the effective area of the box to select the man’s head, but if as in the process of testing $\Delta$ selection is too large, easy to produce residual target of pedestrians, which affect the accuracy of the pedestrian detection. In order to reduce missed detection as much as possible, $\Delta$ needs to be set smaller, but at the same time, more redundant information will be introduced to cause false detection. Based on this, this paper proposes an MSER algorithm combining D-NP, which can effectively solve the above problems.
The specific algorithm is as follows:

a) The gray-scale mapping is applied to the pre-processed depth image. The gray-scale mapping equation is as follows:

\[
DP(i,j) = a + \frac{b - a}{H_2 - H_1} \cdot [D(i,j) - H_1]
\]

where \(a\) is the lower limit of the grayscale range, \(b\) is the upper limit of the grayscale range, \(DP(i,j)\) is the processed depth image, \(D(i,j)\) is the original depth image, \(H_1\) is the standard depth limit, and \(H_2\) is the standard depth Ceiling.

b) Use equation (2) as the stability condition to determine the maximum stable extreme value region.

c) The maximum extremum stability region extracted in step b is fitted with the minimum external rectangle. According to the characteristics of the human head, the head is similar to the spherical shape with the middle high and the sides low [14]. The ideal image \(D_{head}(i,j)\) obtained through the depth camera should satisfy the following relationship:

\[
\begin{align*}
\text{Max}\{D_{head}(i,j)\} &= D_{head}\left(\frac{m}{2}, \frac{n}{2}\right), (0 \leq i \leq m, 0 \leq j \leq n) \\
D_{head}(x_1, y_1) &> D_{head}(x_2, y_2), ((x_1 - x_2)(x_1 + x_2 - m) < (y_2 - y_1)(y_2 + y_1 - n))
\end{align*}
\]

where \(m\) is the width of the image, and \(n\) is the height of the image. Because the actual image can't be completely consistent with the model, there are various errors. In order to reduce the interference caused by noise and the case of multiple extreme values, and at the same time to speed up the search speed, it needs to be normalized. The normalization method is as follows:

\[
D_{\text{mxn}} \xrightarrow{\text{Bicubic Interpolation}} D_{9x9}
\]

The normalization here adopts Bicubic interpolation [15], which utilizes the gray value of 16 points around the sampling point for cubic interpolation. A local maximum value search is performed on the normalized picture. If the maximum value is within the range of 7x7 pixels in the middle of the region, the region is considered as the candidate region of the head, otherwise it is excluded.

d) Divide the normalized area into 9 sub-areas of size 3x3, take the average value of the sub-areas in the direction of the four neighborhoods of the central area, and calculate the standard deviation according to the following equation:

\[
stddev = \sqrt{\frac{\sum_{i=1}^{4}(d_i - d)^2}{4}}
\]

where \(stddev\) is the standard deviation, \(d\) is the average of the four regions, and \(d_i\) is the average of each region. According to formula (7), the unqualified area is removed.

\[
g(i,j) = \begin{cases} f(i,j), stddev < A \\ 0, stddev \geq A \end{cases}
\]

where \(A\) is the set threshold, \(f(i,j)\) is the head candidate region collection before processing, and \(g(i,j)\) is the head candidate region collection after processing.

e) The candidate header area screened in step d has the problem of nesting and duplication. Use equation (8) to remove all the candidate head areas and only retain the innermost candidate head area.

\[
\{bBox\}' \& \{bBox\} = \{bBox\}'
\]

where \(bBox\) is all the candidate frames, \(bBox\)' is the final selected candidate frame, and \& is the overlapping area operation of two or two selection frames between the defined sets.

f) After filtering the candidate head area, a small middle area is left, which cannot contain all the head features. According to the characteristics of the depth map, as the distance between the head and the camera increases, the area it occupies in the image also becomes smaller and changes in a linear relationship. In combination with this feature, with the center of the candidate head region as the midpoint, an adaptive rectangular frame is selected according to the height of the human body to select.
the candidate head region. According to the national standard GB/T10000-1998 of Chinese adult body size statistics, the maximum head length of an adult male and female accounts for 9% to 13% of the height, and the maximum head width accounts for 7% to 11%. The middle proportion is taken as the edge length of the adaptive frame. The adaptive rectangle frame length and width equation are as follows:

\[
\begin{align*}
\text{Length} &= 0.11H; \\
\text{Width} &= 0.09H; \\
\end{align*}
\]

where \( \text{Length} \) is the length of the candidate head area, \( \text{Width} \) is the width of the candidate head area, and \( H \) is the depth of the center of the candidate head area.

3. Experimental results and analysis

3.1. Experimental steps
The algorithm of this paper is based on the C language of the OpenCV library. The configuration of the computer platform is 8G memory and 2.5 GHz i5 CPU. The TOF camera is used to detect the depth image obtained by looking down. The head area of the human body is extracted using the MSER algorithm combined with D-NP. However, these candidate head areas include other feature areas that are caused by environmental changes and human changes. At this time, the candidate head area needs to be classified and identified. In this paper, the candidate head region is used as an input sample, and HOG (Histogram of Oriented Gridients) [16] features are extracted and the SVM (Support Vector Machine) is used to accurately identify the head region. The experimental process is as follows:

a) Foreground extraction: Since the pixel value of the depth image represents the distance from the image collector to each point in the scene, an inversion operation needs to be performed on the image in order to make the pixel value represent the real height of the pedestrian, that is, the camera mounting height minus the pixel value. According to equation (1), the foreground area was extracted, and \( M = 0, \alpha = 800, \beta = 2400 \) were taken in combination with the actual situation.

b) Preprocessing: In the process of image acquisition, the depth of image quality will inevitably be affected by factors such as external light and noise. As shown in the black area (Figure 1(a)), the depth map based on the TOF camera has holes in the edge and smooth areas of the target (the pixel value is 0). Such points are considered noise points. The loss of depth information in noise points will affect the image segmentation effect, so it needs to be repaired. Then, the median filter and closed operation are used to smooth the image. The image after smooth processing is shown in Figure 1(b).

c) Candidate head area detection: After the preprocessing is completed, the candidate head area is extracted according to the MSER algorithm steps combined with D-NP. In this study, the human head model was analyzed and the actual depth map data was used for repeated experiments. The \( \lambda \) value in equation (7) was finally determined to be 100. The results are shown in Figure 1(c), (d) and 1(e).

![Figure 1](image)

Figure 1. (a) Original depth image. (b) Effect after preprocessing. (c) Effect using a basic MSER detection algorithm. (d) Effect after being filtered by equation (7). (e) Effect after being equation (8).
d) Adaptive growth algorithm of candidate regions based on height: the head region of human body can't be completely included in the head region extracted by combining the MSER algorithm of D-NP. The linear relationship between height and head in the image is used to make reasonable adjustments to the head region. Figure 2(a) and 2(b) are the comparison diagrams before and after the adjustment.

![Figure 2](image)

**Figure 2.** (a) Effect before adjustment. (b) Effect after adjustment.

e) Extracting HOG features: The candidate region is uniformly normalized to a resolution range of 27×27, and their HOG features are extracted. The HOG parameters are set to: detection window:27×27, block size:15×15, cell size:5×5, block sliding step size:5, gradient direction quantization:18.

f) SVM classifier training: The HOG features of positive and negative samples are obtained by collecting a large amount of sample data. Here, 1000 positive samples are taken, the label is set to 1, the negative samples are 1000, and the label is set to -1. Training, where the SVM type is set to C_SVC, the kernel function is LINEAR, and a supported vector machine is obtained after training.

g) SVM classification: The obtained sample characteristics are input into the SVM classifier to obtain the classification results and realize the accurate recognition of the head region.

3.2. Performance test and comparison

In order to verify the effectiveness of the algorithm, this paper applies Zhang's [2] water injection algorithm, watershed algorithm [3] and the algorithm in this paper to test on our test set in the same environment. The experimental results are shown in Table 1. The relative error between the measured number and the real number is used as the error detection rate to obtain the accuracy. It can be seen from Table 1 that the detection algorithm in this paper is superior to the other two algorithms in both detection efficiency and running time, and the running frame rate of the detection algorithm in this paper is as high as 41.67 fps, which can completely meet the requirements of real-time detection. The number of individual false positives was mainly due to the fact that pedestrians carried objects similar to the outline of heads.

| Algorithm          | Real number of people | Measured number of people | False number of people | Missed number of people | Accuracy (%) | Frame rate (FPS) |
|--------------------|-----------------------|---------------------------|------------------------|-------------------------|--------------|------------------|
| Reference [2]      | 324                   | 270                       | 61                     | 54                      | 70.13        | 7.89             |
| Reference [3]      | 324                   | 308                       | 13                     | 17                      | 91.39        | 21.73            |
| This paper         | 324                   | 320                       | 5                      | 4                       | 97.26        | 41.67            |

No matter it is multiple pedestrians walking side by side or pedestrians carrying backpacks, hats and other complex scenarios, the algorithm in this paper can better detect pedestrians, and some experimental results are shown in Figure 3. It can be known from the above tests that the algorithm in this paper has high recognition rate, strong robustness, good real-time performance, and excellent comprehensive performance.

![Figure 3](image)

**Figure 3.** (a) Detection effect of the algorithm for pedestrians carrying backpacks. (b) Detection effect of the algorithm for pedestrians carrying hats.
4. Conclusion

In this paper, we propose an improved MSER pedestrian detection algorithm based on a TOF camera, which solves the problems that traditional pedestrian detection methods reduce the accuracy of detection due to the complexity of pedestrian wearing clothing and lighting changes. This algorithm utilizes the D-NP algorithm for the maximum stable region extracted by the MSER algorithm to complete the screening of the candidate head region. Then, the adaptive rectangular frame is used to make the extracted area contain more head features, which is convenient for the accurate detection of pedestrians. Under various lighting conditions, the algorithm detection recognition rate reaches 97.26%, which effectively reduces the impact of lighting changes on pedestrian detection, and can better detect pedestrians in complex scenes where pedestrians carry backpacks and wear hats. Experimental results show that this method has significantly improved accuracy, speed and robustness compared with traditional methods.

References
[1] Kim S W, Jung J Y, Lee S J, Morales A and Ko S J 2013 Proc. Int. Conf. on Consumer Electronics (Las Vegas, NV, America) pp 65-66
[2] Zhang X C, Yan J J, Feng S K, Lei Z, Dong Y and Stan Z L 2012 Proc. Int. Conf. on Advanced Video and Signal-Based Surveillance (Beijing, China) pp 215-20
[3] Xu S S.Statistical Study of Pedestrians Based on Kinect.[D].Tohoku University:Zhang Y Z, 2014,pp 29-38
[4] Yin Z Q, Gu G H, Chen Q and Qian W X 2014 Application of 3D Depth Image in Automatic Passenger Flow Counting System China Laser 6 198-204
[5] Rauter M 2013 Proc. Conf. on Computer Vision and Pattern Recognition Workshops (Portland, OR, America) pp 529-34
[6] Xia L, Chen C C and Aggarwal J K 2011 Proc. Conf. on Computer Vision and Pattern Recognition (Colorado Springs, CO, America) pp 15-22
[7] Ge Y and Chen C 2018 Research on Multiple License Plate Location Based on Text Location and Color Detection Computer and Digital Engineering 3 575-79
[8] Zhang X N, Gao X B and Tian C N 2018 Text detection in natural scene images based on color prior guided MSER Neurocomputing 307 61-71
[9] Marco J, Hernandez Q, Muñoz A, Dong Y, Jarabo A,Kim M H, Tong X and Gutiérrez D 2018 DeepToF: Off-the-Shelf Real-Time Correction of Multipath Interference in Time-of-Flight Imaging Acm Transactions on Graphics 6 1-12
[10] Shi W W and Pan A 2018 Proc. Int. Conf. on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (Prague, Czech Republic) pp 277-80
[11] Bertalmio M, Bertozzi A L and Sapiro G 2011 Proc. Conf. on Computer Vision and Pattern Recognition (Kauai, HI, America) p 1
[12] Yang B and Yang Q 2014 Two-way license plate positioning method combining Lab color space and morphological processing Science Technology and Engineering 26 108-110
[13] Matas J, Chum O,Urban M and Pajdla T 2004 Robust wide-baseline stereo from maximally stable extremal regions Image and Vision Computing 22 761-67
[14] Liu L, Chen Z H, Zhang Y and Zhao D N 2017 Multi-feature head counting algorithm for depth images Journal of Shenzhen University (Science and Technology) 6 38-44
[15] Keys R G 1981 Cubic convolution interpolation for digital image processing Transactions on Acoustics, Speech, and Signal Processing 29 1153-60
[16] Dalal N and Triggs B 2005 Histograms of Oriented Gradients for Human Detection Proc. Conf. on Computer Vision and Pattern Recognition (San Diego, CA, America) pp 886-93