ICIC Express Letters
Part B: Applications
Volume **8**, Number **9**, September **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 1315–1322

EDGE-BASED STEREO MATCHING AND DEPTH INFORMATION ACQUISITION OF HUMAN MOTION OBJECT FEATURE POINTS WITH BINOCULAR VISION

Qing Ye[1], Jun Deng[1], Yongmei Zhang[2] and Wenbo Huang[1]

[1]School of Electronic and Information Engineering
[2]School of Computer Science and Technology
North China University of Technology
No. 5, Jinyuanzhuang Road, Shijingshan District, Beijing 100144, P. R. China
yeqing@ncut.edu.cn

Received April 2017; accepted June 2017

Abstract. *Aimed at the impact of the occlusion problem and ambiguities problem from 3D-2D projection in human action recognition with monocular vision, an edge-based stereo matching and depth information acquisition of human motion object feature points with binocular vision is proposed. After using two-stage method based on model plane accomplishing camera calibration of binocular stereo-vision system, an algorithm of stereo matching is proposed for acquiring human depth information. This algorithm uses SURF features matching method based on human edge information to match feature points, and then it uses optimization matching algorithm based on limited constraints and region matching to accomplish feature block matching. After obtaining the coordinates of matching feature points, we acquire the accurate 3D spatial coordinate of human feature points through camera calibration results. Experimental results show that the algorithm can acquire accurate 3D spatial coordinate of human feature points with high accuracy, which effectively reduces the impact of the occlusion problem and ambiguities problem, and improves the robustness of human action recognition.*
Keywords: Binocular vision, Camera calibration, Stereo matching, Depth information acquisition, Human action recognition

1. **Introduction.** In recent years, human action recognition has attracted much attention in the field of computer vision. However, all the environmental information involved in the real world is three-dimensional, while monocular vision could only restore two-dimensional information and lose the deep information of the object. In order to acquire the three-dimensional spatial data of an object, three-dimensional information acquisition technology based on binocular vision is researched, which is non-contact and can avoid a large amount of hardware. And this technology is more convenient and precise. Those make three-dimensional information capture based on binocular vision more important and popular. Zhao et al. [1] show that depth information from monocular image, but the error is 35.77% in experimental result. Uchida et al. [2] propose passive stereo vision to capture 3D information and iterative closest point for matching. However, this paper is limited to human face. Wang and Zhu [3] propose that image-based 3D building reconstruction has become a hot topic in the fields of computer graphics and computer stereo vision. It shows the importance of 3D information acquisition for 3D reconstruction. Zhao and Sun [4] propose an algorithm which can locate the size and area of the template image fast and accelerate the speed of 3D tracking, but the number of feature points affects the efficiency of matching. Yang et al. [5] use self-adaptive partial matching weights and partial weight filtering algorithm to enhance the quality of the reconstruction of deep image. This method cannot effectively reduce the impact of occlusion.

From above, we can see that researches on three-dimensional space based on binocular vision have gained certain achievements and it still has some problems to be solved. On the other hand, the number of researches on human action recognition in the three-dimensional world through binocular vision is less, while the method based on binocular stereo vision can realize three-dimensional information acquisition and eliminate the influence of occlusion and ambiguities only by two-dimensional figure information under natural light source. In the paper, in order to reduce the impact of the problems in the above paper, an edge-based stereo matching and depth information acquisition of human motion objects feature points with binocular vision is proposed.

2. **Acquirement of Tree-Dimensional Information.** The research will be divided into five modules: camera calibration, pre-processing and human moving object detection, SURF optimized features matching based on edge information, optimization algorithm based on limited constraints and region matching and 3D spatial coordinate acquisition. After the camera calibration in the binocular vision by two-stage method based on model plane, the human moving object was extracted by pre-processing and background subtraction based on mixed Gaussian model. In the process of stereo matching, this algorithm uses SURF features matching method based on human edge information to match feature points, and then it uses optimization matching algorithm based on limited constraints and region matching to accomplish feature block matching. After obtaining the coordinates of matching feature points, we acquire the accurate 3D spatial coordinate of human feature points through camera calibration results. And the block diagram is shown in Figure 1.

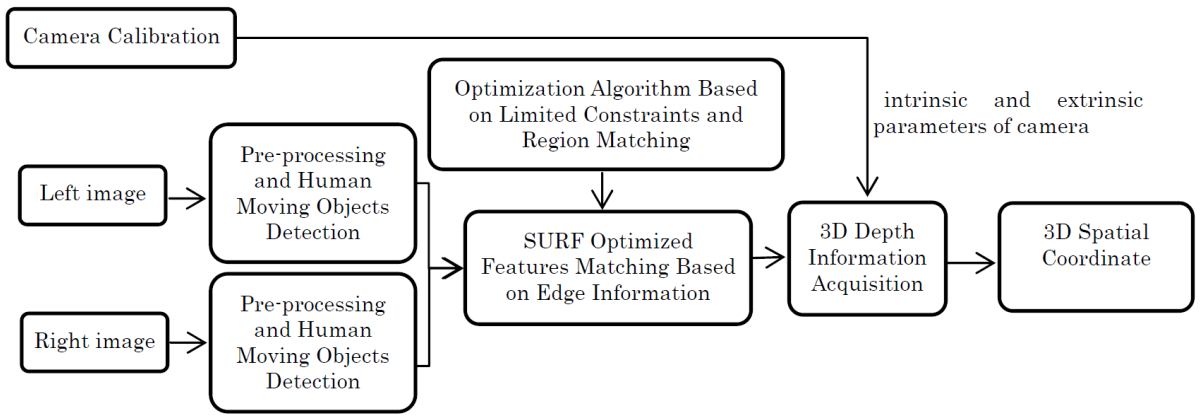


FIGURE 1. Stereo matching and depth information acquisition block diagram

2.1. **Camera calibration.** Camera calibration is mainly used to acquire the intrinsic and extrinsic parameters of two cameras to reconstruct 3D stereo space. The method based on binocular calibration [6] is used in this module.

Firstly, prepare a calibration board with 99 circular feature points, then take at least five pictures of the calibration board in different directions; detect these images with Sobel operator; denoise by filtering; finally use least square method for the fitting of elliptic boundary points and work out the coordinates of the elliptic center; use epipolar constraint rules to match the centers of characteristic circle in the left and right pictures; then use Leventberg-Marquardt algorithm to acquire the homography of the camera H, the intrinsic and extrinsic parameters and distortion factors of two cameras, and these parameters will be saved as xml files.

2.2. **Pre-processing and human moving objects detection.**

2.2.1. *Image pre-processing.* In order to reduce noise and improve image quality, image pre-processing is applied. Firstly, we choose original image as input and then to carry out smooth denoising by selecting a smooth window-neighboring domain S (such as a 3*3 square window) to handle a digital figure $f(x, y)$ with its center (x, y). Finally, median filtering method [7] is applied to image enhancement, to replace the value of one point in the digital image or numerical sequence with the value of points in a neighboring domain of this point, making the neighboring pixel values close to the true value, and eliminate the isolated noise points, which can improve high-frequency signals such as edge signals and make fuzzy images clear.

2.2.2. *Human moving objects detection.* Background subtraction is a frequently-used approach for detecting moving objects. The rationale in the approach is that of detecting the moving objects from the difference between current image and background image. We regard human action as an object of research in this module and then the background subtraction based on mixed Gaussian Model [8] is applied to detecting human moving objects.

2.3. **Edge-based SURF optimization stereo matching.** It is difficult to figure out how to describe the features of an object and precisely extract features in studying model identification. This algorithm uses SURF features matching method based on human edge information to match feature points, and then uses optimization matching algorithm based on limited constraints and region matching to carry out feature block matching.

2.3.1. *Edge detection.* Sobel operator is a gradient magnitude which can be folded with images on leveling and vertical templates to obtain two gradient matrixes G_x and G_y on human body images sequence. Then the gradient magnitude of each pixel in the figure can be obtained through Formula (1).

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * I \quad G = \sqrt{G_x^2 + G_y^2} \tag{1}$$

Use the method of threshold value to handle each pixel in the image to create an image of edge amplitude.

2.3.2. *SURF feature points matching.*

Step 1. Filtrate the figures with Gaussian filter after the obtained edge image establishes scale space. The size of the filter on the first floor of the first group is 9, 15 on the second and so on in the equal difference of 6. There are 3 groups of filers. The filters on first floor in the next group are the same as those on the second floor in the prior group.

Step 2. Figure out partial extreme points. Hessian matrix detection is adopted in SURF algorithm to figure out the extreme points. The speed rests with that the SURF algorithm defines the original Hessian matrix:

$$H(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix} \tag{2}$$

As:

$$H(x, \sigma) = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{3}$$

where, D_{xx}, D_{xy}, etc. are the values of box-filter. After the correction by filter approximate Gaussian kernel function and scale factor, the determinant of Hessian can be described as:

$$\det(H_{approx}) = D_{xx}D_{xy} - (\omega D_{xy})^2 \tag{4}$$

In the detection of extreme points, comparisons are made through 27 points in the 3*3*3 neighboring domain in the two floors up and down the point X. Feature points are selected through non-maximum restriction.

Step 3. Precise positioning of extreme points. As the alternative feature points, points with low contrast ratio should be eliminated from the partial extreme points being detected by box filter for the benefit of precise positioning.

Step 4. SURF Matching

① The Similarity Measurement Formula of SURF Algorithm is as follows:

$$dist = \sum_{i=0}^{i=63} (description_{real} - description_{base})^2 \tag{5}$$

It is the sum of squares of the differences between the descriptors of the two images.

② Find the two most matching points, mark them as $dist_f, dist_s$ respectively, if:

$$\frac{dist_f}{dist_s} < thresh \tag{6}$$

In the program, the value of thresh is 0.9, then the two points are recognized matched with each other in two images. When detecting feature points, we can figure out the feature value of Hessian matrix and its trajectory at the same time.

$$tarce(i) = \sum (d_x + d_y) \tag{7}$$

where, d_x, d_y are still the responses of integral figures on axis x and y of the filter. In measuring the similarity, the symbol of the trajectory of Hessian matrix is adopted to accelerate matching.

2.3.3. *Optimization algorithm based on limit constraints and domain matching.* In this part, we choose region matching algorithm [9,10] to realize optimization. This algorithm uses limit constraints and feature points of left and right images to locate results approximately. It is beneficial to compress search range and speed up. And its principle is as follows.

In Figure 2, left image I_l regards $p_l(u_l, v_l)$ as center point, then the method chooses one area W for $(2m+1)(2n+1)$ as window template T in the neighborhood of p_l. We need obtain basis matrix F by calibration parameters of camera and also polar l_{pr} of p_l in right plane image based on F and coordinates of p_l. Due to the fact that we have known the coordinates of p_l and polar l_{pr}, we can confirm a region of search Q where window template T that regards p_l as centre point can search. At the same time this method chooses T distribution of pixel gray to compare with other distribution of pixel gray that T searches from region of search Q in the window of the right image. Next we calculate its similarity and acquire maximal window of similarity as the window of the matching position, and its center is the corresponding matching point p'_r with the p_l.

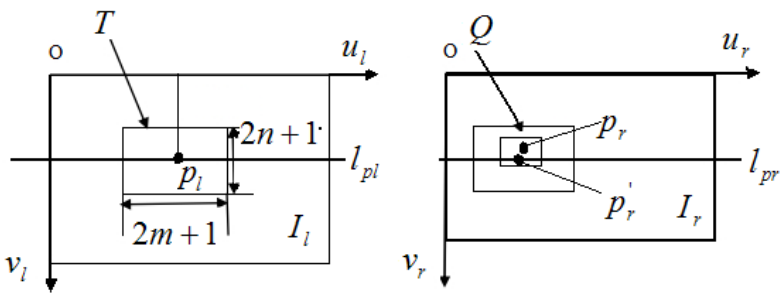


FIGURE 2. Region matching algorithm

In order to enhance the precision of the matching and reduce the calculated amount, limit constraints are usually applied within the search scope and similarity computation. In the previous stereo matching algorithm, we need to calculate the similarity of pixel gray that is covered with window template T at left and right images in the search process for searching best-match position of window. In order to avoid the influence of noise and brightness difference, this method uses the normalized cross correlation measurement (NCC) which will not be affected with white noise and scale factor errors as similarity measure. As for the normalized cross correlation measurement $C(p_l, p'_r)$ in Formula (8), p_l and p'_r are center points of left and right images, $(2m+1)(2n+1)$ is the window about matching, $\overline{f}(u_l, v_r)$ and $\overline{f}(u'_l.v'_r)$ are gray average, σ is variance. The value range of $C(p_l, p'_r)$ is $[-1, 1]$. Feature points are more matching in the left and right image areas, and the value of $C(p_l, p'_r)$ is much greater. Finally, the method of edge-based SURF optimization stereo matching obtains precise coordinates of human matching feature points.

$$C(p_l, p'_r) = \frac{\sum_{i=-n}^{n}\sum_{j=-m}^{m}\left[f_l(u_l+i, v_l+j) - \overline{f_l}\,(u_l, v_l)\right]\left[f_r\,(u'_r+i, v'_r+j) - \overline{f_r}\,(u'_r, v'_r)\right]}{(2n+1)(2m+1)\sqrt{\sigma^2(f_l)\times\sigma^2(f_r)}} \tag{8}$$

2.4. **Three-dimensional information acquisition.** According to the camera imaging principle [11], the position relation between the coordinates of cameras on the left and right can be represented through spatial transformation matrix M_{1r} as follows:

$$\begin{gathered}\begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} = M_{1r}\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 & r_x \\ r_4 & r_5 & r_6 & r_y \\ r_7 & r_8 & r_9 & r_z \end{bmatrix}\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad M_{1r} = [R|T] \\ R = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix} \quad T = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}\end{gathered} \tag{9}$$

where, R and T are respectively the translational transform vectors between the coordinates of the left camera and that of the right camera. According to Formula (9), we can see that the relation between the figure planes of two cameras are as follows for a spatial point in a coordinate:

$$\rho_r\begin{bmatrix} X_r \\ Y_r \\ Z_r \end{bmatrix} = \begin{bmatrix} f_r r_1 & f_r r_2 & f_r r_3 & f_r t_x \\ f_r r_4 & f_r r_5 & f_r r_6 & f_r t_y \\ r_7 & r_8 & r_9 & r_z \end{bmatrix}\begin{bmatrix} zX_1/f_1 \\ zY_1/f_1 \\ z \\ 1 \end{bmatrix} \tag{10}$$

Then, the spatial three-dimensional coordinates can be represented as:

$$\begin{cases} x = zX_1/f_1 \\ y = zY_1/f_1 \\ z = \dfrac{f_1(f_r t_x - X_r t_x)}{X_r(r_7X_1 + r_8Y_1 + f_1r_9) - f_r(r_4X_1 + r_5Y_1 + f_1r_6)} \\ \quad = \dfrac{f_1(f_r t_y - Y_r t_x)}{Y_r(r_7X_1 + r_8Y_1 + f_1r_9) - f_r(r_4X_1 + r_5Y_1 + f_1r_6)} \end{cases} \tag{11}$$

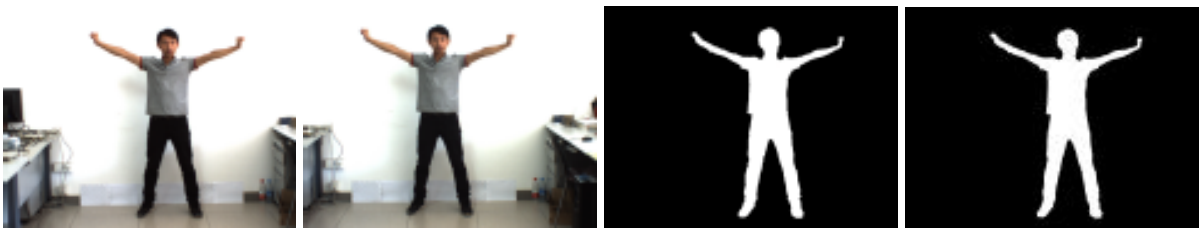
3. **Experimental Results.** The software platform in this paper is the development platform VS2010 with OpenCV2.4.2. And its hardware resource is a computer with high performance. The database of the VS-250D binocular camera and the binocular visual testing development software platform is self-established on visual human action recognition.

Figure 3 shows the original left and right images, and the result picture by the difference method. Results show that full use of the background subtraction based on Gaussian

Model can effectively eliminate the interference of various sophisticated backgrounds in the original images, which makes preparations for post features matching.

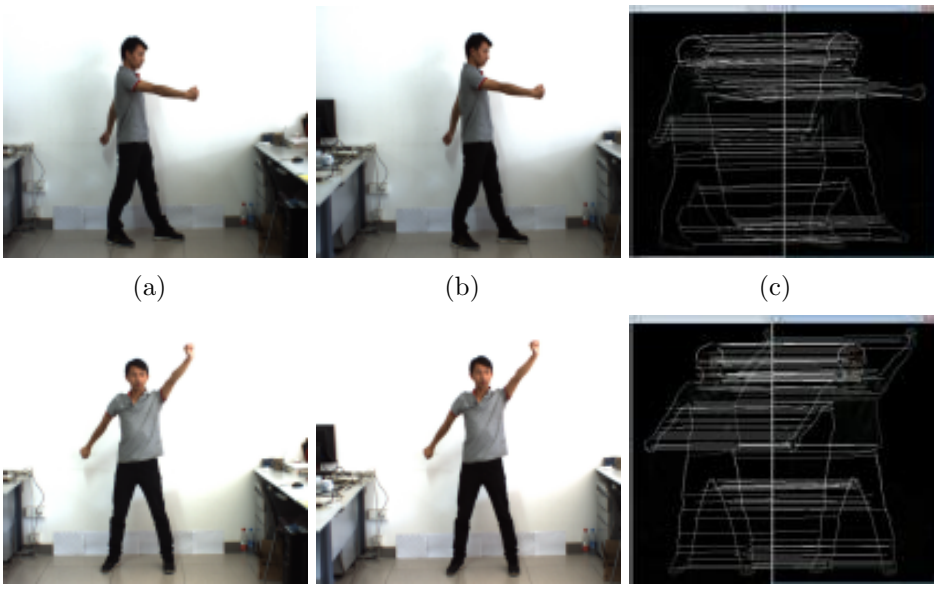
Figure 4 represents the original left and right images and the result of features matching. The result shows that SURF features matching method based on edge information can help realize the capture of the edge information concerning human body in a precise way, find the feature points of human body, and achieve fast and accurate matching. From Figure 4 we can see a very accurate and efficient matching, which can generally reach 91.6%.

Table 1 shows the depth information and errors, in which the measurement unit is cm. And the spatial three-dimensional coordinate is centered on the optic center of the left camera. Different matching windows are selected to work out the corresponding maximum matching coefficient and record it in the table. For more precise calculation, the results in this paper are calculated to six decimal places. Z coordinate value is worked out in Formula (9) through the principle of binocular vision, while the measured Z value is the linear distance between the left camera and human face by steel ruler. From Figure 5, what can be clearly seen is the comparison between the measured Z value and the real Z value. As the depth values of the mass center in the human face domain achieved by the principle of binocular stereo vision and by steel ruler are being compared, the relative error is within the range of allowance. From Table 1 we can see that the error ratio is small and the matching coefficient is above 90%.



(a) The original images (b) The foreground images

FIGURE 3. Result of pre-processing and human motion object



(a) (b) (c)

(d) (e) (f)

FIGURE 4. Result of features matching

TABLE 1. Depth information and error

Feature point sequence	X coordinate value	Y coordinate value	Z coordinate value	Z measured value	Error after optimization	Matching factor
1	10.06	93.14	−5.61	−5.45	2.70%	0.951
2	−12.29	85.89	−8.40	−8.04	4.36%	0.913
3	15.19	80.67	−8.86	−8.58	3.18%	0.933
4	−79.01	72.67	16.01	16.58	3.64%	0.928
5	40.17	65.22	6.12	6.64	5.58%	0.906
6	−34.05	43.07	1.08	1.15	6.61%	0.890
7	−9.73	20.32	6.92	7.20	4.09%	0.928
8	12.45	8.43	3.05	3.16	3.38%	0.931

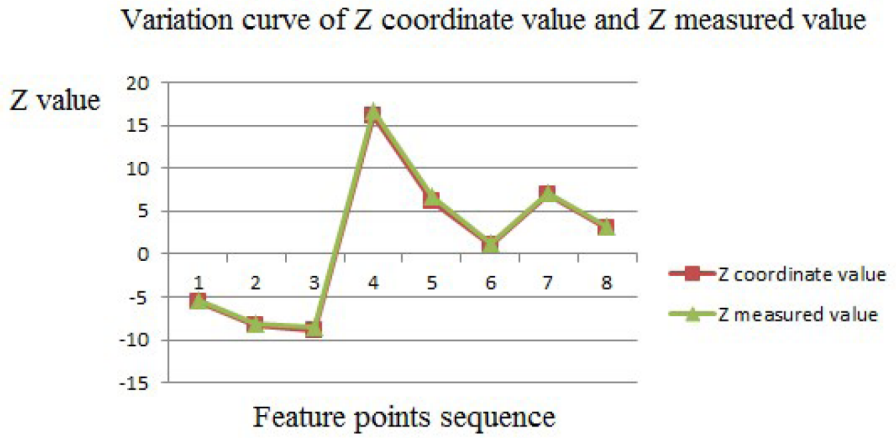


FIGURE 5. Variation curve of Z coordinate value and Z measured value

Comparing the features extraction method mentioned in this paper with those in other papers, Tang and Cai [12] use a novel SIFT descriptor based on a color quantization matrix and can achieve a matching ratio more than 90%, which is limited to color images. Wirayuda [13] uses weighted Euclidean score to feature matching and the accuracy rises to 87.83%. In this paper, we use the SURF features matching method based on edge information and optimize it based on domain matching and limit constrains which can effectively eliminate the impact of the occlusion problems and ambiguities problems. The algorithm leads to a high matching coefficient and accomplishes stereo matching and depth information acquisition.

4. **Conclusion.** This paper carries out an algorithm of stereo matching and depth information acquisition, which includes two key points: SURF features matching method based on edge information optimization algorithm, limited constraints and region matching. This method fully utilizes the edge information of human body. The algorithm can obtain the three-dimensional spatial coordinates of the human feature points accurately, and overcome the interference of occlusion and ambiguity effectively. It provides reliable guarantee to subsequent human action recognition.

Acknowledgment. This paper is supported by National Natural Science Foundation of China (Grant No. 61371143), Excellent Talent Training Project of Beijing (2013D0050020-00002), Advantage Disciplines Project by North China University of Technology (XN078), Initial Funding Project and Excellent Young Teacher Training Project by North China University of Technology.

REFERENCES

[1] X. Zhao, X. Zhu and J. Yu, Learning depth information from monocular image, *Manufacturing Automation*, vol.32, no.3, pp.15-17, 2010.

[2] N. Uchida, T. Shibahara, T. Aoki, H. Nakajima and K. Kobayashi, 3D face recognition using passive stereo vision, *International Conference on Information Processing*, Genova, Italy, pp.950-953, 2005.

[3] J. Wang and L. Zhu, 3D building facade reconstruction based on image matching-point cloud fusing, *Chinese Journal of Computers*, vol.35, no.10, pp.2072-2075, 2012.

[4] Q. K. Zhao and Y. K. Sun, Three dimensional tracking with fast locating of image scale and area, *Journal of Image and Graphics*, vol.21, no.1, pp.0114-0121, 2016.

[5] T. Yang, M. Gao, K. Yin and Z. Wu, High-quality depth map reconstruction combining stereo pair, *Journal of Image and Graphics*, vol.20, no.1, pp.0001-0010, 2015.

[6] Z. Y. Zhang, A flexible new technique for camera calibration, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, no.11, pp.1330-1334, 2000.

[7] L. Zhong and Y. Zhang, Vector median filter of ranked thresholds for color images, *Journal of Image and Graphics*, vol.16, no.3, pp.330-335, 2011.

[8] G. H. Tian, J. Q. Yin, Y. Z. Yan and G. D. Li, Gaussian mixture models and principal component analysis based human trajectory behavior recognition, *Acta Electronica Sinica*, vol.44, no.1, pp.143-148, 2016.

[9] H. Shi and H. Zhu, Stereo matching based on adaptive matching windows and multi-feature fusion, *Pattern Recognition and Artificial Intelligence*, vol.29, no.3, pp.193-200, 2016.

[10] Y. P. Yang, Dynamic programming stereo matching algorithm based on region segmentation, *Journal of Graphics*, vol.36, no.1, pp.90-94, 2015.

[11] M. L. Andreas, V. G. Peter and N. Sebastian, Efficient nonlinear Markov models for human motion, *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, pp.220-228, 2014.

[12] B. C. Tang and N. Cai, A novel SIFT descriptor based on a color quantization matrix, *Journal of Shandong University (Engineering Science)*, vol.41, no.2, pp.46-50, 2011.

[13] T. A. B. Wirayuda, Palm vein recognition based-on minutiae feature and feature matching, *International Conference on Electrical Engineering and Informatics*, Legian-Bali, Indonesia, pp.350-355, 2015.