Displays 69 (2021) 102073

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Stereo disparity optimization with depth change constraint based on a continuous video

Baoli Lu [a,b,c,*], Yu He [d], Haining Wang [e,**]

[a] Institute of Semiconductors, Chinese Academy of Sciences, Beijing, 10083, China
[b] Center of Materials Science and Optoelectronics Engineering & School of Microelectronics, University of Chinese Academy of Sciences, Beijing, 100049, China
[c] Beijing Key Laboratory of Semiconductor Neural Network Intelligent Sensing and Computing Technology, Beijing, 100083, China
[d] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[e] School of Police Administration, People's Public Security University of China, Beijing, 100038, China

ARTICLE INFO

Keywords:
Disparity optimization
Three-dimensional reconstruction
Depth change constraint
Video images

ABSTRACT

Three-dimensional reconstruction based on stereo vision technology is an important research direction in the field of computer vision, and has a wide range of applications in industrial measurement, medical image reconstruction, cultural relic preservation, robot navigation, virtual reality and other fields. However, the three-dimensional reconstruction of moving objects usually has poor accuracy, low efficiency and poor visualization effect due to the image noise, motion blur, complex and time-consuming calculation etc. In this article, a disparity optimization method based on depth change constraint is proposed, which utilizes the correlation of the adjacent frames in the continuous video sequence to eliminate mismatches and correct the wrong disparity values by introducing a depth change constraint threshold. The experiments on the video images which are taken by a binocular stereo vision system demonstrate that our method of removing incorrect matches bears satisfactory results and it can greatly improve the effect of the three-dimensional reconstruction of the moving objects.

1. Introduction

Three-dimensional (3D) reconstruction is the process of obtaining the shape and appearance of real objects. With the help of relevant instruments, it can complete the digital reproduction of the 3D information of a specific target or scene. With the rapid development of reconstruction technology and computer software and hardware, the scale, quality and efficiency of 3D reconstruction have also been greatly improved. It has become an important means for humans to obtain spatial information, and has been widely applied in industrial measurement [1,2], medical image reconstruction [3], cultural relic preservation [4,5], robot navigation [6,7], augmented reality [8,9] and other fields [10–12].

Laser scanning is a common method of 3D reconstruction [13,14], which uses laser scanning equipment to obtain the point cloud and color information of the object surface. This kind of method is accurate and efficient, and can effectively obtain high-precision 3D model. However, laser scanning requires the use of professional measuring equipment, which is susceptible to various restrictions such as the use environment, the object to be measured, and the cost.

Vision-based 3D reconstruction [15,16] methods use computer vision technology to recover 3D scenes from 2D images, which hold extension applications of 2D image processing such as face recognition [17,18], face editing [19], person re-identification [20,21], gender recognition [22] and so on. This method has the characteristics of non-contact, strong flexibility and low cost, and does not require a large amount of hardware support. Therefore, vision-based 3D reconstruction is gradually being valued.

Using binocular or multi cameras for 3D reconstruction is a representative vision-based reconstruction method. By simulating the principle of human vision, this method observes the same scene from two (or more) viewpoints to obtain images under different perspectives, finds the corresponding points between the images through the stereo matching algorithm, and then uses the triangulation principle to calculate the position deviation (i.e. disparity) between the pixels of the corresponding points to restore the 3D information of the image scene, so as to achieve the 3D reconstruction of the target.

In recent decades, 3D reconstruction based on binocular stereo vision technology has received much attention and research of many

* Corresponding author at: Institute of Semiconductors, Chinese Academy of Sciences, Beijing, 10083, China.
** Corresponding author.
E-mail addresses: lubaoli@semi.ac.cn (B. Lu), manian_2002@163.com (H. Wang).

https://doi.org/10.1016/j.displa.2021.102073
Received 24 May 2021; Received in revised form 6 July 2021; Accepted 11 August 2021
Available online 26 August 2021
0141-9382/© 2021 Elsevier B.V. All rights reserved.

scholars, and has been continuously improved and perfected on the basis of predecessors, and has gradually been applied in real life [23–25]. However, there are still many difficulties and shortcomings that need to be solved urgently. For example, since this kind of method involves the matching of corresponding points between images in the process of point cloud restoration, the reconstructed object is often required to have certain texture information. How to better deal with the input image sequence containing weakly textured objects affects the robustness and applicability of the algorithm. In addition, the 3D reconstruction of moving objects usually has poor accuracy, low efficiency and poor visualization effect due to the image noise, motion blur, complex and time-consuming calculation etc.

This article is dedicated to studying how to optimize the disparity map obtained by stereo matching for the 3D reconstruction of moving objects. A method of using the inter-frame relationship of continuous video to eliminate the mismatches and correct the false disparity values by adding a depth change constraint is proposed. The experimental results indicate that our method effectively improves the effect of 3D reconstruction of moving objects.

The rest of this article is organized as follows. In Section 2, we discuss the previous work about 3D reconstruction. Section 3 introduces the principle and specific implementation steps of the proposed method. And Section 4 provides the experimental platform and the visualized experimental results. At last, the conclusions and future work are given in Section 5.

2. Related work

Recovering 3D structures from 2D images is a notoriously complex process that requires expertise with often limited results. In the literature, many methods have been proposed to solve the problem of 3D reconstruction, which can be divided into two classes: active methods and passive methods. Active methods [26–28] retrieve 3D point coordinates on the surface of objects/scenes by projecting a controlled light source (laser, structured light, etc.). While passive methods only use images or videos captured from multiple viewpoints by one or more digital cameras as sufficient input data to start the 3D reconstruction process.

Structure from motion (SFM) is one of the typical passive methods. In [29], the authors proposed a novel global calibration approach based on the fusion of relative motions between image pairs. They defined an efficient Contrarious trifocal tensor estimation method to extract translation directions, and then used an efficient translation registration method to recover accurate camera positions. In [30], the authors proposed a Hybrid SFM for 3D reconstruction, which adopted Global SFM to compute the camera's parameters and utilized Incremental SFM to compute sparse point clouds. In [31], the authors presented a SFM system for multi-scale objects/scenes 3D reconstruction from uncalibrated images/video taken by a moving camera characterized by variable parameters. To reduce computation time, the authors [32] proposed an algorithm for the good choice of image pairs that would be used by the Modified Match Propagation (MMP) to improve the sparse 3D reconstruction. These image pairs would be selected on the basis of the result already achieved by SFM, and the MMP algorithm would be applied for each image pair to retrieve new matches and their 3D coordinates. The final 3D point cloud was achieved by fusion of results obtained from the image pairs selected.

Another typical passive method is multi-view stereo (MVS). An open-source MVS implementation named COLMAP proposed by Schonberger and Frahm [33] offers a wide range of features for the reconstruction of ordered and unordered image collections. And Open Multiple View Geometry (OpenMVG) [34] is a well-known open-source library that deals with multi-view solid geometry, providing feature extraction and matching methods and a complete toolchain for structure from motion. In [35], the authors presented a new method for large-scale MVS based on dense matching between very high-resolution images, which allowed to obtain a very dense 3D point cloud of high quality at a relatively low computational cost. However, this method required the use of rich texture images to avoid making use of costly optimization algorithms. In [36], the authors employed a Particle Swarm Optimization (PSO) method in the patched expansion process for the avoidance of possible local traps. To accelerate high-quality multi-view matching, the authors [37] presented a massively parallel method named Gipuma based on the patch-match idea, which can obtain more accurate depth maps and implement parallelization. The approaches based on MVS are often used to get high-quality dense 3D reconstruction results, but they require in input calibrated stereo images as well as a long computation time.

Recently, convolutional neural networks (CNNs) are increasingly introduced into these typical methods [38,39]. In [40], the authors presented a learning framework for surface reconstruction in passive multi-view scenarios. Their solution consisted in a N-view volume sweeping, trained on static scenes from a small scale dataset equipped with ground truth. In [41], the authors presented an end-to-end 3D reconstruction system that could produce high-quality 3D models from a set of unordered image collections. Their workflow was a typical 3D reconstruction architecture that consisted of SFM, MVS, and surface reconstruction, and could automatically recover desirable 3D models without any interactive operations. In [42], the authors presented a method for 3D face reconstruction from multi-view images with different expressions. They optimized the 3D face shape by explicitly enforcing multi-view appearance consistency and used a CNN network to regularize the non-rigid 3D face according to the input image and preliminary optimization results. In [43], the authors proposed a new MVS network which exploits the attention mechanism for the multi-scale feature pyramid to capture larger receptive fields and richer information.

3. Proposed method

In the real scene, since the position of the moving object in the three-dimensional space changes continuously, the change in the spatial position of the object is relatively small during the two consecutive frames of the video, and the change of the corresponding disparity value is also limited. Therefore, according to the measurement scene, we utilize some prior knowledge to restrict the change of the disparity value between the previous and next frames, and judge whether the matching is correct or not. Then the wrong disparity value would be corrected to achieve disparity optimization.

3.1. Depth change constraint threshold estimation

A depth change constraint threshold is introduced to represent the maximum change in the disparity value between the previous and next frames, which is denoted as $DCCT$.

First, suppose that the moving speed of the object is v and the frame rate of the camera is r, and the amount of change in the actual height of the object during the two consecutive frames of the shooting video can be obtained as:

$$\Delta h = \frac{v}{r} \tag{1}$$

Then, through analyzing the distance resolution of the used binocular stereo vision system in different positions, it can be known how much the distance changes in the 3D space would cause the change of imaging position on the image plane. It should be noted that here we only discuss the single-pixel accuracy of the camera itself, without considering the sub-pixels obtained by pixel interpolation. According to the principle of stereo disparity, binocular stereo vision system can achieve 3D space coordinates of an object from its two different view pictures captured at one time. The schematic diagram of the distance measurement model is illustrated in Fig. 1, O_L and O_R are the optical centers of the left and right camera. The two cameras have the same

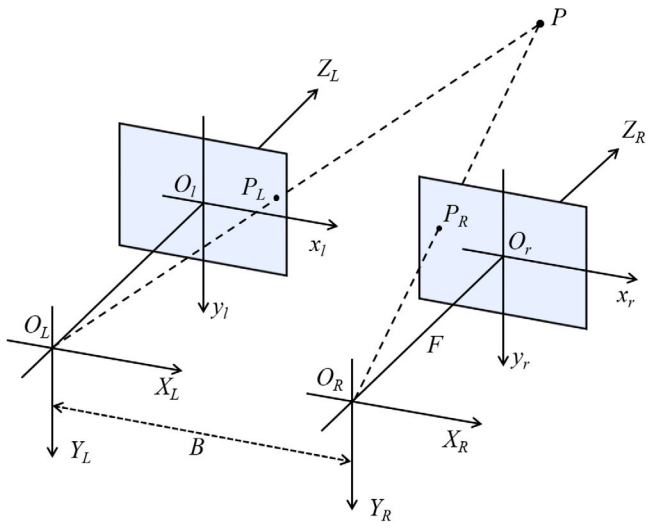

Fig. 1. Schematic diagram of the distance measurement model of binocular stereo vision system.

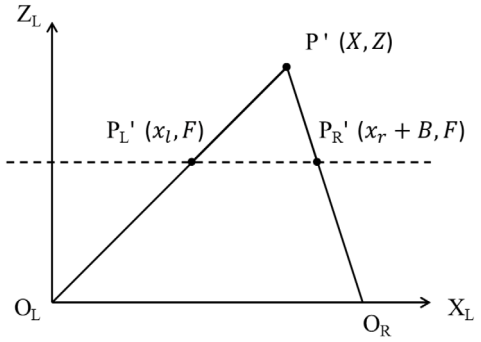

Fig. 2. Projection model of the binocular stereo vision system on the $O_L - X_L Z_L$ plane.

focal length F. The distance between optical center O_L and O_R is baseline denoted by B. $O_L O_l$ and $O_R O_r$ individually express the optical axes of the left and right camera which are parallel to each other. $P(X, Y, Z)$ is an object point in the world coordinate system. It is projected through the projection center of the lens to the points $P_L(x_l, y_l)$ and $P_R(x_r, y_r)$ in the image plane, where two image coordinates of the cameras are denoted by $O_l - x_l y_l$ and $O_r - x_r y_r$. Set the original point of the world coordinates in the optical center O_L. Then through this simple mathematical model we can get the expression of 3D world coordinates: $O_L(0, 0, 0)$, $O_R(B, 0, 0)$, $P_L(x_l, y_l, F)$, $P_R(x_r + B, y_r, F)$. If the binocular system is projected onto the $O_L - X_L Z_L$ plane as illustrated in Fig. 2, the coordinates of the points P', P'_L, P'_R are (X, Z), (x_l, F), $(x_r + B, F)$, which P', P'_L, P'_R represent the projection points of P, P_L, P_R on the $O_L - X_L Z_L$ plane respectively. Therefore, It can be obtained from the triangle similarity theorem as follows.

$$\frac{x_l}{X} = \frac{F}{Z} \tag{2}$$

$$\frac{-x_r}{B - X} = \frac{F}{Z} \tag{3}$$

From Eq. (2), we can get:

$$Z = \frac{FX}{x_l} \tag{4}$$

Z is the coordinate value of the object point P along Z_L axis in the world coordinates, expressing the distance between object and camera. After the object moves, when its projection point on the image plane

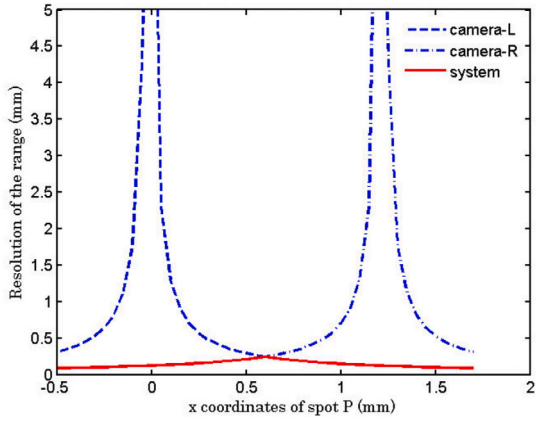

Fig. 3. The change of the range resolution in the horizontal direction.

differs by one pixel, the distance between object and camera becomes Z':

$$Z' = \frac{FX}{x_l + a} \tag{5}$$

here, a represents the actual physical size of a pixel. So for the left camera, the change of the distance in the 3D space is ΔZ_L, which represents the spatial resolution of the left camera.

$$\begin{aligned} \Delta Z_L &= | Z' - Z | \\ &= | \frac{FX}{x_l + a} - \frac{FX}{x_l} | \\ &= | \frac{aZ^2}{FX + Za} | \end{aligned} \tag{6}$$

Similarly, according to Eq. (3), the spatial resolution of the right camera ΔZ_R can be obtained as follows.

$$Z = \frac{F(B - X)}{-x_r} \tag{7}$$

$$\begin{aligned} \Delta Z_R &= | Z' - Z | \\ &= | \frac{F(B - X)}{-x_r + a} - \frac{F(B - X)}{-x_r} | \\ &= | \frac{aZ^2}{F(B - X) + Za} | \end{aligned} \tag{8}$$

Fig. 3 shows the relationship between the range resolution and the X-coordinate of the spatial point P under the same measurement distance when setting $a = 0.003$ mm, $B = 1.2$ mm, $F = 2$ mm, $Z = 10$ mm. The dotted line and the dash-dotted line respectively represent the curves of the distance resolution ΔZ_L and ΔZ_R of the left and right cameras changing with the X-coordinate of the point P in space. We believe that as long as one of the two cameras can distinguish a certain amount of change in distance, the entire system can distinguish the amount of change. So the distance resolution ΔZ of the entire system should take the smaller value of ΔZ_L and ΔZ_R. The red solid line in the figure represents the curve of the distance resolution ΔZ of the entire system varying with the X-coordinate of the spatial point P. It can be seen that when $\Delta Z_L = \Delta Z_R$, i.e. $X = B/2$, the distance resolution of the system takes the largest value, which means the distance resolution is lowest at this location. So taking the distance resolution value of this position for subsequent estimation as Eq. (9), the result will be relatively more reliable.

$$\Delta Z = | \frac{aZ^2}{F \cdot \frac{B}{2} + Za} | = \frac{2aZ^2}{FB + 2Za} \tag{9}$$

After that, according to Eqs. (1) and (9), we can further estimate the real range of the variation in the disparity value caused by the

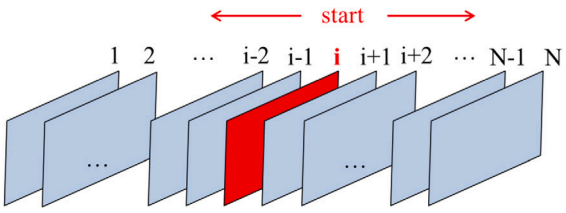

Fig. 4. Starting from the reference frame of the disparity map, perform a two-way comparison with adjacent frames in positive and reverse order respectively.

movement of the object during the time of shooting two consecutive frames. And in this way, $DCCT$ can be set as:

$$DCCT = \lceil \frac{\Delta h}{\Delta Z} \rceil = \lceil \frac{v(FB + 2Za)}{2aZ^2 r} \rceil \tag{10}$$

3.2. Mismatch elimination

For each frame of stereo images in the continuous video, a corresponding disparity map can be obtained. After determining the depth change constraint threshold, perform differential processing on the disparity maps corresponding to two consecutive frames of images in the video. And for each disparity value in the disparity map, we can get a corresponding change value which will be compared with the depth change constraint threshold. If the change value of the disparity is less than the depth change constraint threshold, it means that the corresponding disparity value is reasonable. Otherwise, if the variation of a disparity value exceeds the depth change constraint threshold, it indicates that the corresponding disparity value is incorrect, and accordingly it can be judged that the corresponding matching is a mismatch and needs to be eliminated.

Suppose there are N frames of disparity map in a video in total, and $D[i]$ is used to represent the disparity value in the ith frame disparity map, $i \in (1, N]$. Similarly, the disparity value at the same image coordinate position in the disparity map of the adjacent frame, i.e. $(i-1)$th frame is expressed as $D[i-1]$. Then the disparity difference between two adjacent frames which is denoted as $change[i]$ can be obtained.

$$change[i] = D[i] - D[i-1] \tag{11}$$

The specific implementation of mismatch elimination is divided into three steps.

The first step is to determine the starting reference disparity map which is used for comparison. It needs to make the starting reference frame of the disparity map meet the following condition as shown in Eq. (12), with the purpose of ensuring that the starting reference frame is reliable. If we just start from the first frame of the video and compare the disparity maps sequentially, we cannot guarantee that there are no mismatches in the disparity map of the first frame. The starting reference disparity map is labeled as $start$.

$$D[i-2] = D[i-1] = D[i] = D[i+1] = D[i+2] \tag{12}$$

The second step is that starting from the reference frame, perform a two-way comparison with adjacent frames in positive and reverse order respectively. As shown in Fig. 4, for the disparity maps corresponding to each frame from $start$ to N in the video sequence, compare the adjacent frames one by one in positive order.

$$change[i+1] = D[i] - D[i+1] \tag{13}$$

And at the same time, for the disparity maps corresponding to each frame from 0 to $start$ in the video sequence, compare the adjacent frames one by one in reverse order.

$$change[i-1] = D[i] - D[i-1] \tag{14}$$

Algorithm 1 Disparity value correction

Input: $DCCT, D, start$

$DCCT \leftarrow$ the depth change constraint threshold
$D \leftarrow$ the original disparity value obtained by stereo matching
$start \leftarrow$ the order of the starting reference frame in continuous video

for $i = start; i < N; i++$ **do**
  $change[i+1] = D[i] - D[i+1]$
  if $|change[i+1]| \le DCCT$ **then**
    $D[i+1] = D[i+1]$
  else
    if $|change[i+1] + change[i+2]| \le 2DCCT$ **then**
      $D[i+1] = (D[i] + D[i+2])/2$
    else
      $D[i+1] = D[i] + (D[i+3] - D[i])/3$
    end if
  end if
end for
for $i = start; i > 0; i--$ **do**
  $change[i-1] = D[i] - D[i-1]$
  if $|change[i-1]| \le DCCT$ **then**
    $D[i-1] = D[i-1]$
  else
    if $|change[i-1] + change[i-2]| \le 2DCCT$ **then**
      $D[i-1] = (D[i] + D[i-2])/2$
    else
      $D[i-1] = D[i] + (D[i-3] - D[i])/3$
    end if
  end if
end for
return the disparity value after correction

The third step is to compare all the disparity difference values with the depth change constraint threshold $DCCT$ individually to judge whether they are mismatches. For the disparity maps from $start$ to N,

$$P_i = \begin{cases} 1, & |change[i+1]| \le DCCT \\ 0, & |change[i+1]| > DCCT \end{cases} \tag{15}$$

where P_i represents the probability that the disparity map of the ith frame is correct, here $i \in [start, N)$. And for the disparity maps from 0 to $start$,

$$P_i = \begin{cases} 1, & |change[i-1]| \le DCCT \\ 0, & |change[i-1]| > DCCT \end{cases} \tag{16}$$

here $i \in [0, start)$.

By repeating the above three steps for each element of all disparity maps, we can realize the judgment of whether all the image points in the continuous video are mismatched.

3.3. Disparity value correction

In order to obtain a complete disparity map, the wrong disparity values need to be corrected.

The specific correction method is summarized in Algorithm 1. For the disparity maps corresponding to each frame from $start$ to N in the video sequence, the adjacent frames are compared in positive order starting from the reference frame. If the variation of the disparity value of the next frame satisfies the depth change constraint condition which means the stereo matching of the next frame is correct, the disparity value of the next frame remains unchanged. Otherwise, it is necessary to continue to judge whether the variation of the disparity value of the frame after next is satisfied. And if it is satisfied, the disparity value of the frame after next is used for correction, otherwise, the disparity value of the second frame after next is used for correction. In addition,

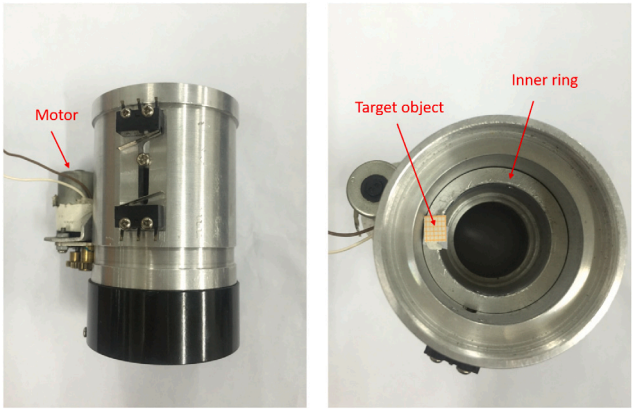

Fig. 5. Motor assembly and target object.

similar processing is performed on the disparity maps corresponding to each frame from 0 to *start* in the video sequence, while the main difference is that the comparison is performed in reverse order starting from the reference frame.

By correcting the wrong disparity value, the optimization of the disparity is realized. And if the 3D reconstruction is performed based on the optimized disparity maps, the effect of the 3D reconstruction of the moving object can be improved.

4. Experiment

In order to verify the effectiveness of the proposed method of mismatch elimination based on the depth change constraint, some experiments have been done in this section. The experimental process and results are introduced as below.

4.1. Experimental platform

In our experiment, the binocular stereo vision system proposed in Ref. [44] is used to capture the video. The system utilizes a single CCD camera as an image sensor and combines the CCD with a biprism, as a result that stereo image pairs of the object can be obtained in a single frame of the CCD from different views. After calibration by the Bouguet camera calibration toolbox, the intrinsic and extrinsic parameters of the stereo system can be obtained as $B = 1.253$ mm, $F = 2.071$ mm, rotation vector $R = [-0.01137 - 0.37638 - 0.00810]$, and translation vector $T = [-1.09614 - 0.00663 - 0.32264]$.

Furthermore, the experiment employs the motor assembly as shown in Fig. 5 as the motion carrier of the target object. As long as the motor assembly is powered on, the motor can drive the inner ring to move and the speed is adjustable. The grid image is used as the target object to simulate the repeated texture area and the weak texture area, and it is fixed on the inner ring of the motor assembly.

4.2. Experiment on 3D reconstruction of moving surface

We conducted a 3D reconstruction experiment on the moving surface, the experimental workflow is shown in Fig. 6.

Firstly, use the binocular stereo vision system mentioned above to shoot the target object, and control the speed of the target object to be 1 mm/s. The distance between the target object and the system is about 6 mm.

After recording a video, decompose the video into a sequence of images and take 100 consecutive frames as experimental images. Fig. 7 shows the original image of the 97th frames. As we can see, two views of the target object are obtained in a single frame of the camera.

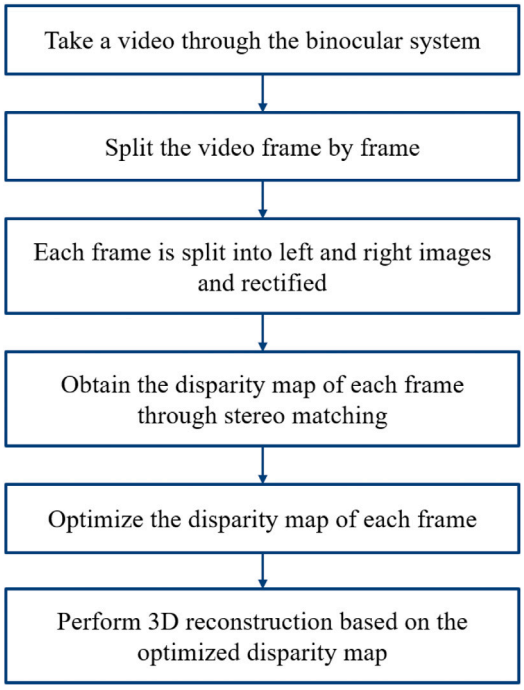

Fig. 6. Experimental workflow of the 3D reconstruction.

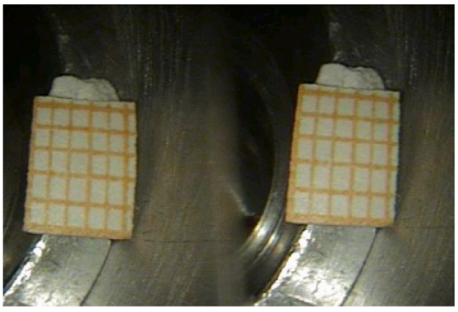

Fig. 7. The original image of the 97th frame.

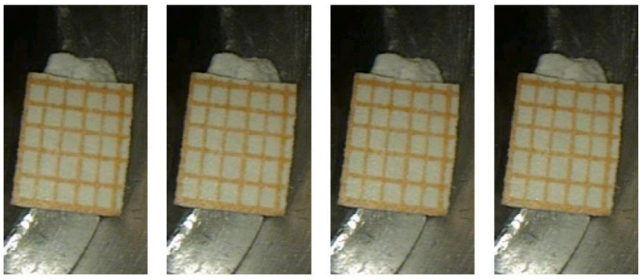

Fig. 8. The left images after rectification.

Then divide all the original images into left and right images and rectify the stereo image pairs using the result of the camera calibration by applying the perspective matrix. The effective region of the left images of the 97th, 98th, 99th, 100th frames after rectification are exhibited in Fig. 8.

Next, the popular graph cut algorithm is chosen for stereo matching of each rectified stereo image pair to obtain the corresponding disparity map. And according to the method and implementation steps stated in Section 3, the mismatches in the disparity map of each frame are

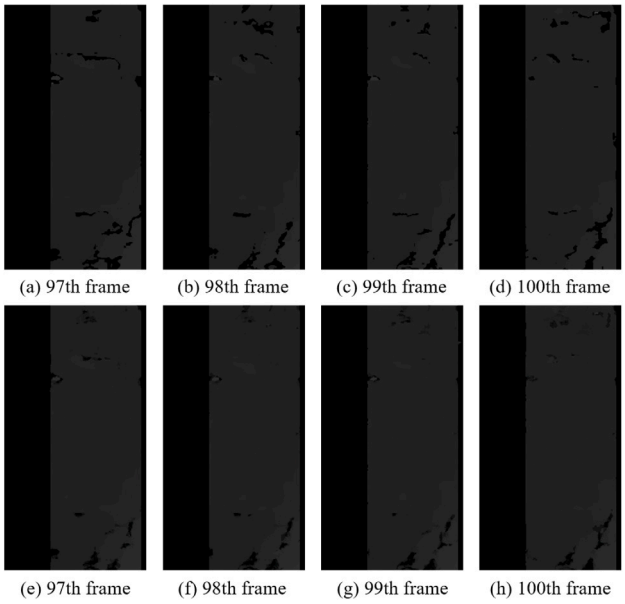

Fig. 9. The disparity maps. (a), (b), (c), (d) are the initial disparity maps for frames 97 to 100, and (e), (f), (g), (h) are the disparity maps after optimization for frames 97 to 100.

eliminated. Fig. 9 lists the disparity maps before and after optimization for frames 97 to 100.

Finally, 3D reconstruction is carried out based on the disparity maps before and after optimization for intuitive comparison. As displayed in Fig. 10, (a), (b), (c), (d) are the reconstruction results based on the initial disparity maps for frames 97 to 100, while (e), (f), (g), (h) are the reconstruction results based on the disparity maps after optimization for frames 97 to 100. It can be clearly seen that the effect of 3D reconstruction based on the optimized disparity maps has been significantly improved, which proves that the method proposed in this paper is effective to eliminate mismatches by controlling the depth change based on the inter frame relationship of continuous video.

4.3. Experiment on detection and matching of feature points

In order to more intuitively show the effect of using the proposed method to eliminate mismatches, we also conduct a feature matching experiment. The SURF algorithm is adopted to perform feature points detection and matching on the rectified stereo image pairs obtained in Section 4.2, and the feature matching point pairs are marked as shown in Fig. 11. It can be seen from the figure that there are some obvious mismatches in the matching results.

For the optimization, we carry out the following operations: Firstly, SURF feature matching is performed on all the stereo image pairs to obtain the disparity map, in which the disparity values obtained by feature matching are some discrete points. Then traverse all the points in the disparity map of the previous frame, and judge one by one whether there is a corresponding point in the searching window on the disparity map of the current frame. If there is, calculate the disparity variation of the two corresponding points on the two adjacent frames and determine whether it exceeds the depth change constraint threshold. If it does not exceed, the disparity value represented by the corresponding point will be retained as a correct match, otherwise the disparity value will be removed as a false match. The experimental result shown in Fig. 12 indicates that through our approach the mismatches in the stereo matching have been effectively eliminated.

5. Conclusion

Aiming at the 3D reconstruction of moving objects, this article proposes a method to eliminate mismatches by taking advantage of the inter-frame relationship of continuous video and taking the depth change as the constraint condition. The method first decomposes the captured video frame by frame, and obtains the corresponding disparity map of each frame of image through stereo matching. Then utilize prior knowledge such as the frame rate of the camera, the movement speed of the object, and the approximate measurement distance to estimate the change range of the real disparity value caused by the movement of the object within the time of shooting two consecutive frames of images, so as to set the depth change constraint threshold between adjacent frames of the disparity map. After the disparity map of a certain frame is determined as the starting reference disparity map, the disparity maps of the adjacent frames are sequentially differentiated in the positive order and the reverse order respectively. For the relative

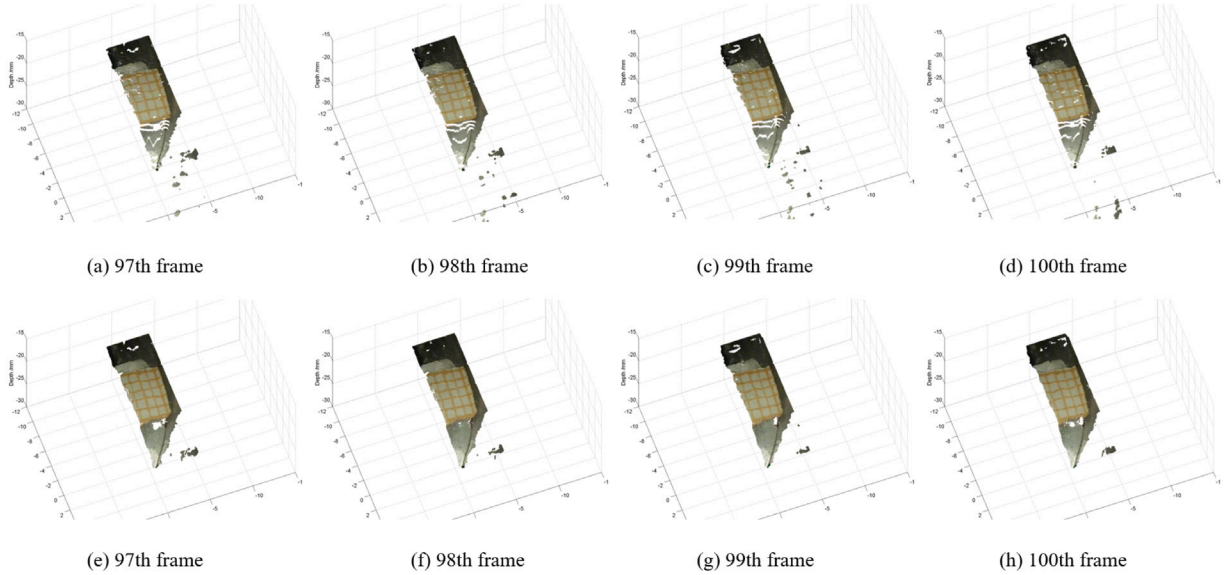

Fig. 10. 3D reconstruction results. (a), (b), (c), (d) are based on the initial disparity maps for frames 97 to 100, and (e), (f), (g), (h) are based on the disparity maps after optimization for frames 97 to 100.

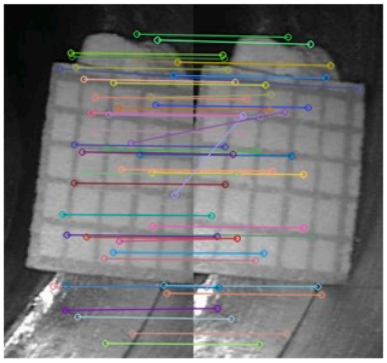

Fig. 11. Result of feature matching for the first frame.

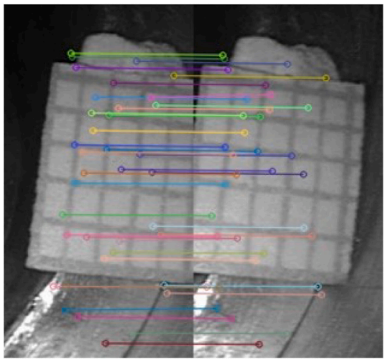

Fig. 12. Result of feature matching after eliminating mismatches.

change of each disparity value in the disparity maps, compare it with the depth change constraint threshold one by one. If the amount of change exceeds the depth change constraint threshold, it is judged that the stereo matching corresponding to the disparity value is a mismatch which needs to be eliminated. Finally, the disparity maps are optimized by correcting the wrong disparity values. In order to verify the effectiveness of the proposed method, experiment on 3D reconstruction of moving surface and experiment on detection and matching of feature points were carried out respectively. And it is proved that the proposed method has superior performance in mismatch elimination and disparity optimization.

In the future, we will explore the possibility of applying this method to real-time 3D reconstruction of moving objects, and promote it to the systems that require high real-time performance.

CRediT authorship contribution statement

Baoli Lu: Writing - original draft, Conceptualization, Methodology, Software, Writing - review & editing. **Yu He:** Data curation. **Haining Wang:** Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (No. 61901436) and the Key Research Program of the Chinese Academy of Sciences, Grant NO. XDPB22.

References

[1] M. Ma, Y. Wang, X. Ling, H. Deng, P. Yao, J. Zhang, X. Zhong, A multidistance constraint method for three-dimensional reconstruction with coaxial fringe projection measurement system, Opt. Lasers Eng. 132 (2020) 106103.
[2] Y. Zhang, W. Liu, Y. Lu, X. Cheng, W. Luo, H. Di, F. Wang, Accurate profile measurement method for industrial stereo-vision systems, Sensor Rev. (2020).
[3] M. Klodt, R. Hauser, 3D image reconstruction from X-ray measurements with overlap, in: Proc. ECCV Conf., Springer, 2016, pp. 19–33.
[4] G. Du, M. Zhou, P. Ren, W. Shui, P. Zhou, Z. Wu, A 3D modeling and measurement system for cultural heritage preservation, in: Proc.IcOPEN, Vol. 9524, International Society for Optics and Photonics, 2015, 952420.
[5] X. Wu, W. Feng, K. Wang, Application of multi-baseline digital close-range photogrametry technique in 3D reconstruction of underground tomb, in: Advanced Materials Research, Vol. 346, Trans Tech Publ, 2012, pp. 847–851.
[6] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, T. Pajdla, Are large-scale 3d models really necessary for accurate visual localization? in: Proc. CVPR Conf., 2017, pp. 1637–1646.
[7] J. Hou, L. Yu, S. Fei, A highly robust automatic 3D reconstruction system based on integrated optimization by point line features, Eng. Appl. Artif. Intell. 95 (2020) 103879.
[8] R. Wang, Z. Geng, Z. Zhang, R. Pei, X. Meng, Autostereoscopic augmented reality visualization for depth perception in endoscopic surgery, Displays 48 (2017) 50–60.
[9] M. Cao, L. Zheng, W. Jia, H. Lu, X. Liu, Accurate 3-D reconstruction under IoT environments and its applications to augmented reality, IEEE Trans. Industr. Inform. 17 (3) (2020) 2090–2100.
[10] X. Ning, F. Nan, S. Xu, L. Yu, L. Zhang, Multi-view frontal face image generation: A survey, Concurr Comput.-Pract. E (2020) e6147.
[11] S. Cheng, X. Li, W. Zhu, W. Li, J. Wang, J. Yang, J. Wu, H. Wang, L. Zhang, X. Li, et al., Real-time navigation by three-dimensional virtual reconstruction models in robot-assisted laparoscopic pyeloplasty for ureteropelvic junction obstruction: our initial experience, Transl. Androl. Urol. 10 (1) (2021) 125.
[12] X. Ning, P. Duan, W. Li, S. Zhang, Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer, IEEE Signal Process Lett. 27 (2020) 1944–1948.
[13] K. Chen, K. Zhan, X.-C. Yang, D. Zhang, 3D reconstruction method for laser spiral scanning point cloud, in: Seventh International Conference on Optical and Photonic Engineering (IcOPEN 2019), Vol. 11205, International Society for Optics and Photonics, 2019, 1120529.
[14] J. Liu, Y. Wang, 3D surface reconstruction of small height object based on thin structured light scanning, Micron 143 (2021) 103022.
[15] X. Bai, C. Yan, H. Yang, L. Bai, J. Zhou, E.R. Hancock, Adaptive hash retrieval with kernel based similarity, Pattern Recognit. 75 (2018) 136–148.
[16] Z. Gao, G. Zhai, H. Deng, X. Yang, Extended geometric models for stereoscopic 3D with vertical screen disparity, Displays 65 (2020) 101972.
[17] L. Sun, W. Li, X. Ning, L. Zhang, X. Dong, W. He, Gradient-enhanced softmax for face recognition, IEICE T. Inf. Syst. 103 (5) (2020) 1185–1189.
[18] L. Zhang, W. Li, L. Yu, L. Sun, X. Dong, X. Ning, GmFace: An explicit function for face image representation, Displays 68 (2021) 102022.
[19] P. Chen, Q. Xiao, J. Xu, X. Dong, L. Sun, W. Li, X. Ning, G. Wang, Z. Chen, Harnessing semantic segmentation masks for accurate facial attribute editing, Concurr. Comput.-Pract. E (2020) e5798.
[20] X. Ning, K. Gong, W. Li, L. Zhang, Jwsaa: joint weak saliency and attention aware for person re-identification, Neurocomputing 453 (2021) 801–811.
[21] C. Yan, G. Pang, X. Bai, C. Liu, N. Xin, L. Gu, J. Zhou, Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss, IEEE Trans. Multimed. (2021).
[22] Y. Lu, W. Li, X. Ning, X. Dong, L. Zhang, L. Sun, C. Cheng, Blind image quality assessment based on the multiscale and dual-domains features fusion, Concurr Comput.-Pract. E (2021) e6177.
[23] G.-S. Hong, B.-G. Kim, A local stereo matching algorithm based on weighted guided image filtering for improving the generation of depth range images, Displays 49 (2017) 80–87.
[24] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, K. Yang, Adaptive unimodal cost volume filtering for deep stereo matching, in: Proc. AAAI Conf., Vol. 34, 2020, pp. 12926–12934.
[25] L. Zhong, J. Qin, X. Yang, X. Zhang, Y. Shang, H. Zhang, Q. Yu, An accurate linear method for 3D line reconstruction for binocular or multiple view stereo vision, Sensors 21 (2) (2021) 658.
[26] Y. Li, R. Lu, Uncalibrated euclidean 3-D reconstruction using an active vision system, IEEE Trans. Robot. Autom. 20 (1) (2004) 15–25.
[27] Y. Wang, K. Liu, Q. Hao, X. Wang, D.L. Lau, L.G. Hassebrook, Robust active stereo vision using Kullback-Leibler divergence, IEEE Trans. Pattern. Anal. Mach. Intell. 34 (3) (2012) 548–563.
[28] A.A. Al-Temeemy, S.A. Al-Saqal, Laser-based structured light technique for 3D reconstruction using extreme laser stripes extraction method with global information extraction, Opt. Laser Technol. 138 (2021) 106897.

[29] P. Moulon, P. Monasse, R. Marlet, Global fusion of relative motions for robust, accurate and scalable structure from motion, in: Proc. IEEE Int. Conf. Comput. Vis.(ICCV), 2013, pp. 3248–3255.
[30] H. Cui, X. Gao, S. Shen, Z. Hu, HSfM: Hybrid structure-from-motion, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1212–1221.
[31] S. El Hazzat, M. Merras, N. El Akkad, A. Saaidi, K. Satori, 3D reconstruction system based on incremental structure from motion using a camera with varying parameters, Vis. Comput. 34 (10) (2018) 1443–1460.
[32] S. El Hazzat, N. El Akkad, M. Merras, A. Saaidi, K. Satori, Fast 3D reconstruction and modeling method based on the good choice of image pairs for modified match propagation, Multimedia Tools Appl. (2019) 1–15.
[33] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proc. CVPR Conf., 2016, pp. 4104–4113.
[34] P. Moulon, P. Monasse, R. Marlet, La bibliothèque openMVG: open source multiple view geometry, in: Orasis, Congrès Des Jeunes Chercheurs En Vision Par Ordinateur, 2013.
[35] E. Tola, C. Strecha, P. Fua, Efficient large-scale multi-view stereo for ultra high-resolution image sets, Mach. Vis. Appl. 23 (5) (2012) 903–920.
[36] W.-C. Chen, Z. Chen, P.-Y. Sung, Stochastic optimization based 3D dense reconstruction from multiple views with high accuracy and completeness., J. Inf. Sci. Eng. 31 (1) (2015).
[37] S. Galliani, K. Lasinger, K. Schindler, Massively parallel multiview stereopsis by surface normal diffusion, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 873–881.
[38] C. Wang, X. Wang, X. Bai, Y. Liu, J. Zhou, Self-supervised deep homography estimation with invertibility constraints, Pattern Recognit. Lett. 128 (2019) 355–360.
[39] C. Wang, X. Bai, X. Wang, X. Liu, J. Zhou, X. Wu, H. Li, D. Tao, Self-supervised multiscale adversarial regression network for stereo disparity estimation, IEEE Trans. Cybern. (2020).
[40] V. Leroy, J.-S. Franco, E. Boyer, Volume sweeping: Learning photoconsistency for multi-view shape reconstruction, Int. J. Comput. Vision 129 (2) (2021) 284–299.
[41] Y. Cai, M. Cao, L. Li, X. Liu, An end-to-end approach to reconstructing 3D model from image set, IEEE Access 8 (2020) 193268–193284.
[42] Z. Bai, Z. Cui, J.A. Rahim, X. Liu, P. Tan, Deep Facial Non-Rigid Multi-View Stereo, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 5850–5860.
[43] K. Zhang, M. Liu, J. Zhang, Z. Dong, PA-MVSNet: Sparse-to-dense multi-view stereo with pyramid attention, IEEE Access 9 (2021) 27908–27915.
[44] B. Lu, Y. Liu, X. Wang, Compact three-dimensional fingerprint acquisition system based on a single camera with a biprism, Acta Photonica Sin. 45 (7) (2016) 710004.