Stereo Matching Algorithm Based on Joint Matching Cost and Adaptive Window

Yu Chai
School of Electrization and Control Engineering
Xi`an University of Science and Technology
Xi`an, China
710054

Xiaojing Cao
School of Electrization and Control Engineering
Xi`an University of Science and Technology
Xi`an, China
710054

***Abstract*—Stereo matching is one of the key technologies in stereo vision. Due to its many problems, it has not been well resolved. Aiming at the problem that the traditional SAD similarity measure function easily causes amplitude distortion, a local stereo matching algorithm combining similarity measure functions is proposed. Firstly, in the traditional SAD similarity measure function, the Census transform is introduced to construct a linear weighted matching cost algorithm for adaptive weights; an adaptive window based on pilot filter is constructed using image structure and color information for cost aggregation; A detection strategy is used to detect matching anomalies, and sub-pixel enhancement and median filtering are performed on the obtained disparity map to obtain a final high-precision disparity map. The experimental results show that the algorithm is effective, the matching accuracy is high, and it has better robustness to the conditions of light distortion and edge information.**

Keywords—Stereo matching; Census transform; bootstrap filter; adaptive window; SAD

I. INTRODUCTION

Stereo matching technology as a basic problem in computer vision and research hotspots and difficulties, its essence is to find the corresponding point in the left and right images to calculate the parallax, and then get the depth information of the object. In the industrial monitoring, visual navigation, virtual reality and image-based rendering and other aspects of high importance and extensive research and application [1]. Many domestic researchers have conducted in-depth research on stereo matching algorithms. Scharstein et al. [2] studied some representative stereo matching algorithms in depth and summarized stereo matching into four steps: matching cost calculation, cost aggregation, disparity selection, and parallax fineness, and stereo matching is mainly divided into local algorithms and global There are two major categories of algorithms.

The global algorithm obtains matching results through the assumption of smoothness and minimization of the energy function. The global algorithms mainly include dynamic programming, image segmentation, and confidence expansion. The global algorithm can better deal with occlusion areas and weak texture areas. The accuracy is higher but the calculation is more complex and inefficient, and it can not meet the real-time application [3]. Compared to the global algorithm, the local algorithm uses the information in the neighborhood of the window. Single-pixel matching has the advantages of high efficiency and easy implementation, but supporting the selection of window size and the calculation of matching cost have always been a challenge [4]. Local algorithms are mainly divided into adaptive window method and adaptive weight method.

In this paper, the use of SAD as a similarity measure function for stereo matching band calculation, the amplitude distortion caused by it is very sensitive, leading to low accuracy. Combining with Census transformation, a linear weighted similarity measure method for adaptive weights is designed, which has higher accuracy and makes the algorithm more robust to lighting distortion. This paper proposes a filter-based cross-adaptive adaptive support window cost aggregation, which can achieve good matching results in low-texture areas and disparity discontinuous areas. In parallax optimization, left and right consistency detection strategies are used to detect matching outliers and modify them. Finally, the resulting disparity maps are subjected to sub-pixel enhancement and median filtering to obtain the final high-precision disparity map.

II. MATCHING COST ALGORITHM

A. Census Transformation

Census transformation is a kind of non-parametric image transformation, which can detect the local structure features in images, such as edge and corner features [5]. The basic idea of the Census transformation is to define a rectangular window in the image area and use this rectangular window to traverse the entire image. The center pixel is selected as a reference pixel, and the gray value of each pixel in the rectangular window is compared with the gray value of the reference pixel. The pixel whose gray value is less than or equal to the reference value is marked as 0, and the pixel larger than the reference value is marked as 1 Finally, they are connected in bits to obtain the transformed result. The transformed result is a binary code stream consisting of 0s and 1s. The transformation process can be expressed as:

$$T(p)=\underset{q\in N_p}{\otimes}\xi(I(p),I(q)) \tag{1}$$

$$\xi(I(p),I(q))=\begin{cases}1 & I(q)>I(p)\\ 0 & I(q)\le I(p)\end{cases} \tag{2}$$

Where p is the center pixel of the window, q is the pixel other than the center pixel of the window, and N_p represents the neighborhood of the center pixel p. $I(*)$ represents the gray value at the pixel*.

Using the Hamming distance to calculate the difference of the binary stream after the Census transform, the smaller the Hamming distance, the higher the matching degree of the two points. The matching cost based on the Census transformation is:

$$C_{Census}(p,d) = \min(Ham\min g(T(p),T(p-d))) \quad (3)$$

Where $C(p,d)$ is the matching cost value when p's disparity value is d.

B. Absolute Error and Algorithm (SAD)

SAD (Sum of absolute differences) is a basic similarity measure function. The basic idea: sum the absolute value of the difference, build a small window, use the window to cover the left image, and select all the pixels in the window's covered area;Also use the window to cover the right image and select the pixels of the coverage area; the left cover area minus the right cover area, and find the sum of the absolute value of the grayscale difference of all pixels; move the window of the right image, repeat the previous processing Find the window with the lowest SAD value in this range, that is, find the best matching pixel region for the anchor point on the left.

Assume that the point $p(x,y)$ is a projection point of a point in the left image, and the point in the right image where the disparity is d with the point p is $(x\text{-}d,y)$, q represents the pixel within the neighborhood of the point p, N_p, $I(x,y)$ represents the gray value of the point p, then the above-mentioned matching cost relation can be expressed as:

$$C_{SAD}(x,y,d) = \sum_{q\in N_P} \min\left\{\left|I_L(x,y) - I_R(x-d,y)\right|, T_1\right\} \quad (4)$$

An upper threshold T_1 is introduced in the above equation, and the gray value exceeding this threshold is replaced with T_1, which is a basic M-estimation and can increase robustness.

C. Joint Matching Cost

Based on the advantages and disadvantages of the above two similarity measure functions, this paper combines the two forms of joint matching cost, and the joint matching cost formula is as follows:

$$C(p,d) = w_1 C_{Census}(p,d) + w_2 C_{SAD}(x,y,d) \quad (5)$$

Among them, w_1 is the matching weight of the Census transformation, and w_2 is the matching weight of the SAD. They reflect the degree of importance in the current support domain. Through the joint matching cost, the noise problem is well suppressed, the mis-match is reduced, and the matching accuracy is improved.

III. Matching Algorithm Based on Filter Adaptive Window

A. Guided Filtering

Guided filtering is a filtering method with edge-preserving properties [6], assuming that q is the output image, I is the leading image, and a_k and b_k are the invariant coefficients of the linear function when the window center is at k. The assumption of this method is that q and I have a local linear relationship in the window centered on pixel k. The output result of a pixel is:

$$q_i = a_k I_i + b_k, \quad \forall i \in w_k \quad (6)$$

Where: w_k is a square window with pixel k as the center and r as the radius. In order to solve the invariant coefficient in the above equation, suppose that p is the result of q-filtering and satisfies the minimization of the difference between q and p. According to the method of unconstrained image restoration, it can be converted into an optimization problem whose value function is:

$$q_i = p_i - n_i \quad (7)$$

$$E(a_k,b_k) = \sum_{i\in w_k}\left(\left(a_k I_i + b_k - p_i\right)^2 + \varepsilon a_k^2\right) \quad (8)$$

(7)where n is noise and p is a degraded image of q contaminated by noise n; (8) where the first term is a fidelity term, minimizing the difference between q and p on the basis of a linear model; The second term is a smooth term that prevents the invariant coefficient ak from being too large; ε is a regularization parameter and usually requires ε>0. Equation (8) is similar to the least-squares method and its solution is:

$$a_k = \frac{\frac{1}{|w|}\sum_{i\in w_k} I_i p_i - \mu_k \overline{p}_k}{\sigma_k^2 + \varepsilon} \quad (9)$$

$$b_k = \overline{p}_k - a_k \mu_k \quad (10)$$

Where μ and $\sigma^{\wedge 2}$ represent the mean and variance of I in the local window w; $|w|$ is the number of pixels in the window. Then, take the window operation in the whole image, and finally take the average value to get the result of equation (8):

$$q_i = \frac{1}{|w|}\sum_{k:i\in w_k}(a_k I_i + b_k) = \overline{a}_k I_i + \overline{b}_k \quad (11)$$

B. Adaptive Window Generation

The traditional fixed-window-based stereo matching algorithm is limited by the fixed window size, and it is difficult to obtain a good matching accuracy. Therefore, the key to the algorithm is to select an appropriate window. In low-texture areas, larger windows are needed to improve matching accuracy; in high-texture areas, smaller windows are needed to protect information such as object edges. Zhang et al. [7] proposed a cross-adaptive window constructed based on the

color and positional relationship of adjacent pixels. This paper proposes a filter-based cross-adaptive adaptive window construction, uses bootstrap filter to preprocess the image, and then adaptively selects any size and shape in the preprocessed image according to the color information of neighboring pixels and the spatial relationship of the image space. Support window.

For the current pixel to be matched p, the horizontal and vertical directions are respectively extended to form a cross-intersection area, which is represented by $H(p)$ and $V(p)$, respectively. As shown in Fig. 1, the size of the area is determined by the arm length $\{h_p^+, h_p^-, v_p^+, v_p^-\}$ in four directions, and can be adaptively changed according to the structure and color information of the image.

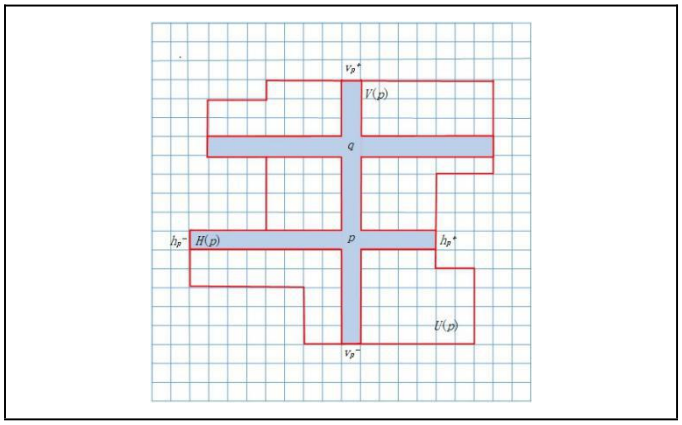

Fig. 1. Adaptive window

Then, the criterion for determining the length of the four arms based on the color difference and the positional relationship is:

$$D_c(p_i, p) < \tau, \quad D_s(p_i, p) < L \tag{12}$$

Where: $D_c(p_i,p)$ is the color difference between pixels pi and p, $D_s(p_i,p)$ is the spatial distance difference between pixels p_i and p, and τ, L are preset color thresholds and distance thresholds. The guidelines define the difference in color and set the maximum arm length. This article uses a variable threshold arm length criterion. Starting from the image texture features, it is believed that the color threshold τ should change with the growth of the arm length l_p, so a linear relationship is expressed as follows:

$$\tau(l_p) = \tau_{\max}(1 - \frac{l_p}{L_{\max}}) \tag{13}$$

Among them: τ_{max} and L_{max} respectively preset the maximum color and distance threshold, l_p is the current arm length. Using the above criteria, cross-sections $H(p)$ and $V(p)$ can be obtained:

$$\begin{cases} H(p) = \{(x,y) \mid x \in [x_p - h_p^-, x_p + h_p^+], y = y_p\} \\ V(p) = \{(x,y) \mid x = x_p, y \in [y_p - v_p^-, y_p + v_p^+]\} \end{cases} \tag{14}$$

Repeat the above process for each pixel q in $V(p)$, find the horizontal support area $H(q)$ for q, and finally combine all $H(q)$ to obtain the adaptive region for any pixel p in the image:

$$U(p) = \bigcup_{q \in V(p)} H(q) \tag{15}$$

C. Parallax Optimization

For the two corresponding matching points $p=(x,y)$ and $p_d=(x_d,y)$ in the left and right images, the adaptive regions $U(p)$ and $U(p_d)$ can be separately generated by the above method, and the common regions are Set as the final support area window:

$$U_d(p) = \{(x,y) \mid (x,y) \in U(p), (x-d,y) \in U(p_d)\} \tag{15}$$

In the final support domain, the cost matching is aggregated, and the total cost in the area is obtained as:

$$C'(p,d) = \frac{1}{N} \sum_{k \in U_d(p)} C(k,d) \tag{16}$$

Where N is the total number of pixels that support the window. Finally, Winner-Takes-All strategy is adopted to select the disparity value corresponding to the minimum matching cost as the initial disparity:

$$d_p = \arg\min_{d \in R} C'(p,d) \tag{17}$$

Where: R is all possible disparity values.

WTA strategy is relatively simple and easy to implement, but it will introduce the wrong choice of disparity. Different disparity values may have the least post-aggregate match cost, but WTA will only select one of them as the time difference of the point to be matched. At this time, the disparity value is not reliable. Therefore, postprocessing is used to refine the initial disparity map and obtain a precise disparity map.

If the left image is the reference image, the matching point of p in the right image is p_d. When the right image is the reference image, the corresponding point of p_d in the left image should be point p. If this point is not satisfied, it is occlusion. Point or mis-matching point, can be detected through the left and right consistency detection (LRC) to match the abnormal point and modify. Points that do not satisfy the following formula are occlusion points or mismatched points.

$$d_L(x,y) = d_R(x - \max(d_L, 0), y) \tag{18}$$

Among them: dL, dR is the left and right initial disparity map. For the detected occlusion point or mismatched point, the first and left effective points in the horizontal direction are scanned, and the original parallax of the mismatched point is replaced by a value with smaller parallax in both, so as to obtain a parallax map.The sub-pixel enhancement of the disparity map obtained above can make the image disparity value fine from the pixel level to the sub-pixel level, effectively improve the resolution, make the error smaller, and optimize the disparity map. Then use the median filter for smoothing

operation to filter out possible noise points, and finally get a fine disparity map.

IV. Experimental Results and Analysis

After In order to verify the accuracy of this algorithm, the algorithm is implemented in windows10 operating system, intel Core i5-7300HQ CPU, 2.50GHz computer, and through the VS2010 compiler and opencv2.4.9 as a software platform, C language as an editing language to experiment and use At present, the Middlebury Stereo Dataset 2 released by the Middlebury website [8] accepted by the academics conducts algorithm experiments and evaluations. The website provides four sets of standard color stereoscopic image pairs and corresponding real disparity maps. By comparing the experimental disparity map with the real disparity map, the quantitative matching error can be obtained, thereby objectively evaluating whether the algorithm is accurate.

Experiments are performed on the proposed algorithm using standard stereo image pairs. (a), (b), (c), (d) are the experimental results of the Tsukuba, Cones, Teddy, and Venus image pairs in sequence, as shown in Fig. 2. The first column is the original left image to be matched; the second column is the real disparity map; the third column is the disparity map obtained by the algorithm in this paper.

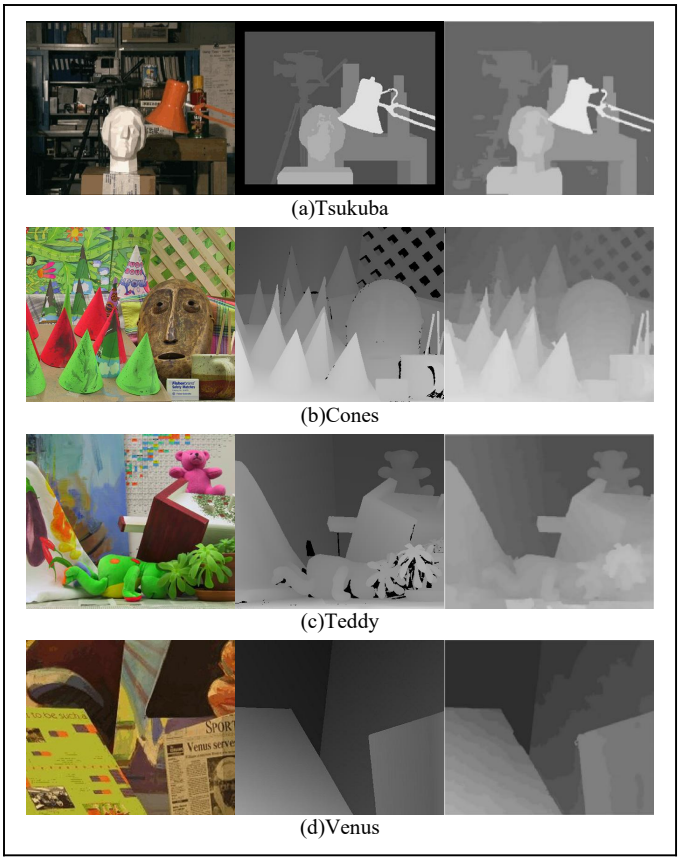

Fig. 2. Experimental results using the proposed method for Middlebury stereo pairs

Upload the results of the algorithm experiment to the Middlebury website, select the error threshold as 1 pixel, and evaluate the performance of the algorithm. As shown in Table 1, the algorithm is compared with the results of other local stereo matching algorithms, where Nonocc(N) indicates the ratio of mismatched pixels in non-occluded regions. All(A) indicates the total false-match pixel ratio, and Disc(D) indicates the depth-discontinuous region mismatch pixel ratio. As can be seen from Fig. 2, this algorithm has obtained a high-precision dense disparity map. From the data in Table 1, it can be seen that the proposed algorithm has a low false matching rate in both non-occlusion and depth discontinuity regions. The average false matching percentage is lower than that of other local stereo algorithms, indicating that the algorithm has achieved the purpose of initial research.

TABLE I. Comparison Between the Proposed Algorithm and Other Matching Algorithms

Image \ Algorithm		Adapt Weight[9]	Segment [10]	Variable Cross[11]	Algorithm
Tsukuba	***N***	1.38	1.25	1.99	1.29
	A	1.85	1.62	2.65	1.98
	D	6.90	6.68	6.77	6.53
Venus	***N***	0.71	0.25	0.62	0.74
	A	1.19	0.64	0.96	0.53
	D	6.13	2.59	3.20	3.01
Teddy	***N***	7.88	8.43	9.75	6.02
	A	13.3	14.2	15.1	12.6
	D	18.6	18.2	18.2	16.3
Cones	***N***	3.97	3.77	6.28	4.09
	A	9.79	9.87	12.7	7.65
	D	8.26	9.77	12.9	6.85
Average error		6.67	6.44	7.60	5.63

V. In Conclusion

This paper proposes a joint cost calculation of joint SAD and Census transforms and constructs a linear weighted similarity measure method for adaptive weights. In the cost aggregation stage, the algorithm uses a filter based cross-correlation algorithm based on the image structure and color information. Adapting to the window generation method, the introduction of guided filtering can well preserve the edge characteristics to meet the different requirements of the window size of the low texture region and disparity discontinuous region, and achieve a more accurate cost aggregation; using left and right consistency detection strategies to detect matches Abnormal points, sub-pixel enhancement and median filtering are performed on the obtained disparity map to obtain the final high-precision disparity map. The standard stereoscopic image pair was tested on the opencv software platform. The experimental results show that compared with the existing local stereo matching algorithm, the proposed algorithm has a low false matching rate in both the non-occlusion area and the depth discontinuity area. Higher-precision disparity maps can be obtained.

References

[1] ZITNICK C L, KANG S B. Stereo for Image-Based Rendering Using Image Over-Segmentation. International Journal of Computer Vision, 2007, 75(1) : 49-65.

[2] SCHARSTEIN D, SZELISKI R. A Taxonomy and Evaluation of DenseTwo-Frame Stereo Correspondence Algorithms . International Journal of Computer Vision, 2002, 47(1) : 7-42.

[3] Zhu Shiping, Yang Liu. Stereo matching algorithm with graph cuts based on adaptive watershed [J]. Acta Optical Sinica, 2013, 33(3): 0315004.

[4] Q X Yang. A non-local cost aggregation method for stereo matching [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2012. 1402-1409.

[5] Zabih R, Woodfill J. Non-parametric local transforms for computing visual correspondence[C]. Stockholm, Sweden: [s.n.], 1994: 151-158.

[6] Meng Hao, Cheng Kang. Binocular vision location based on SIFT feature points[J]. Hei long jiang: Journal of Harbin Engineering University, 2009, 30(6). 649-652.

[7] K Zhang, J B Lu, C Lafruit. Cross-based local stereo matching using orthogonal integral images [J]. IEEE Transactions on Circuits and systems for Video Technology, 2009, 19(7): 1073-1079.

[8] D Scharstein, R Szeliski. The middlebury stereo vision page [OL]. http//vision.Middlebury.edu/stereo/, 2014.

[9] Yoon K J, Kweon I S . Adaptive support weight approach forcorrespondence search[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(4) : 650-656.

[10] TOMBARI F, MATTOCCIA S, DI STEFANO L. SegmentationBased Adaptive Support for Accurate Stereo Correspondence / / Procof the 2nd Pacific Rim Symposium on Advances in Image and VideoTechnology. Santiago, Chile, 2007 : 427-438.

[11] Zhang K, Lu J, Lafruit G. Cross-based local stereo matching using orthogonal integral images[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2009, 19(7) : 1073-1079.