J. Shanghai Jiaotong Univ. (Sci.), 2016, 21(2): 184-191
DOI: 10.1007/s12204-016-1710-7

Consistent Depth Maps Estimation from Binocular Stereo Video Sequence

DUAN Fengfeng (段峰峰)
(Cultural Resources Research and Development Center of Hunan, Hunan Normal University, Changsha 410081, China;
School of Computer Science, Communication University of China, Beijing 100024, China)

Abstract: In the paper, an approach is proposed for the problem of consistency in depth maps estimation from binocular stereo video sequence. The consistent method includes temporal consistency and spatial consistency to eliminate the flickering artifacts and smooth inaccuracy in depth recovery. So the improved global stereo matching based on graph cut and energy optimization is implemented. In temporal domain, the penalty function with coherence factor is introduced for temporal consistency, and the factor is determined by Lucas-Kanade optical flow weighted histogram similarity constraint (LKWHSC). In spatial domain, the joint bilateral truncated absolute difference (JBTAD) is proposed for segmentation smoothing. The method can smooth naturally and uniformly in low-gradient region and avoid over-smoothing as well as keep edge sharpness in high-gradient discontinuities to realize spatial consistency. The experimental results show that the algorithm can obtain better spatial and temporal consistent depth maps compared with the existing algorithms.
Key words: consistent depth maps, binocular stereo video sequence, Lucas-Kanade optical flow weighted histogram similarity constraint (LKWHSC), joint bilateral truncated absolute difference (JBTAD)
CLC number: TN 911.73 **Document code:** A

0 Introduction

Currently, the industry and technologies of stereo video attract a great attention. The common format is binocular stereo video which is represented as side-by-side, top-and-down, interlaced synthesis or separated left-right parts, and can be displayed and viewed with complementary color or polarized lenses. The depth maps estimation is one of the most popular technologies in recent years. The main application for these depth maps is rendering new perspectives of the captured scene by means of depth image based rendering (DIBR), and they are also widely used in many other areas, such as feature extraction of stereo vision, 3D scene reconstruction, robot vision and tracking.

In binocular stereo video system, depth maps can be extracted by calculating the disparity and depth according to stereo matching of left view and right view. Currently, although the technologies of stereo matching are relatively perfect, the studies are few in temporal and spatial consistency of depth maps. From the related work and practice, it can be found that flickering artifacts occur without considering the temporal consistency of depth maps and result in visual distortion in synthesized or reconstructed views[1]. It is caused by miscalculation of disparity over time in depth maps sequence estimation. While the noise and blur of depth maps are caused without considering the spatial consistency, which has bad effect on the quality of synthetic views. So it is necessary to eliminate flickering artifacts and handle noise in order to achieve a spatial and temporal consistency. In this way, the high quality depth maps sequence can be obtained.

In related studies, the methods for handling temporal inconsistency of stereo video depth maps sequence are mainly on smooth filtering in temporal domain and temporal coherence constraint between frames. For smooth filtering in temporal domain, bilateral filtering, multilateral filtering and median filtering are mainly used. Cigla and Alatan[2] introduced the median filter to static or background pixels based on constancy of brightness to eliminate flickering artifacts. Garcia et al.[3] proposed an algorithm of multilateral filtering. Just as in these studies, only sparse depth maps sequence can be obtained by smooth filtering. In

Received date: 2014-05-18
Foundation item: the Science and Technology Innovation Project of Ministry of Culture of China (No. 2014KJCXXM08), the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (No. 2012BAH37F02), and the National High Technology Research and Development Program (863) of China (No. 2011AA01A107)
E-mail: dffeng2010@126.com

addition, blur distortion is caused due to the lack of treatment to time-domain motion. So the problem of temporal inconsistency cannot be solved efficiently. Temporal coherence constraint is to define temporal weighting function for color, brightness, motion object, scene or characteristics between adjacent frames or related frames. Several methods are mainly adopted, such as extending single frame to sequence frames, separation of dynamic and static pixels in scenes and motion tracking. Khoshabeh and Richardt et al.[1,4] studied the depth map extraction from a single pair of images and then extended to sequence frames. Although depth map can be well obtained from a single pair of stereo images for many stereo matching algorithms, it is not sufficient to simply apply them to temporal frames independently without considering the temporal consistency between adjacent frames. Lee and Pham et al.[5-6] proposed an algorithm for separation of dynamic and static pixels in scenes to introduce constraints, respectively. These studies usually assume that the scene is static or quasi-static, or the movement can be ignored compared with the sampling frequency. It is difficult to deal with temporal inconsistency in dynamic or constant brightness scenes. Meanwhile, the depth of static scene is not constant and always changing, so the problem of temporal inconsistency cannot be solved efficiently. Min et al.[7] studied the motion tracking based on optical flow method to realize temporal consistency. In these related studies, local related constraints are often used but there are many problems in object selection, construction of constraint functions, and optimal solution.

For spatial consistency, it is usually to set smooth term of energy optimization function. And the methods of smooth filtering in spatial domain and similarity based on threshold are mainly used. Bilateral filtering is mainly adopted for smoothing as it is efficient in noise reduction and edges preserving. Richardt et al.[4] proposed an improved bilateral filtering algorithm. Pham et al.[6] introduced the information permeability algorithm to implement smooth filtering. But these algorithms usually need to use linear interpolation and may lead to a decline in accuracy. Khoshabeh et al.[1] studied the truncated weighted function based on threshold for smooth processing and edge preserving. Usually, these methods could eliminate the noise effectively, but were still poor in smooth effect and had difficulties in non-texture regions.

This paper applies the method of graph cut and energy optimization to global matching and depth estimation. An algorithm in spatial and temporal domain is proposed in this paper to obtain high quality consistent depth maps. In temporal domain, the penalty function with constraint factor is introduced, and the factor is determined by Lucas-Kanade optical flow weighted histogram similarity constraint (LKWHSC) to associate adjacent frames for temporal consistency. In spatial domain, joint bilateral truncated absolute difference (JBTAD) is proposed to process the neighborhood pixels for spatial consistency.

1 Depth Estimation Based on Graph Cut and Energy Optimization

1.1 Preprocessing on Binocular Stereo Video Sequence

The method of quasi-euclidean uncalibrated epipolar rectification is used for correcting and constraining the left and right video sequence frames in order to improve the matching accuracy and handle the occlusion[8]. Similarly, color standardization is also applied to frames of dual-channel video sequence to accurately calculate the histogram as well as improve the matching accuracy. In this way, the color consistency can be obtained, too[9]. The brightness values of video sequence frames are $\beta(i,j)$, and the expressions in the image window of $m \times n$ are defined by

$$\bar{\beta}_{\mathrm{l}}(i,j) = [\beta_{\mathrm{l}}(i,j) - \mu_{\mathrm{l}}]/\delta_{\mathrm{l}}, \tag{1}$$

$$\bar{\beta}_{\mathrm{r}}(i,j) = [\beta_{\mathrm{r}}(i,j) - \mu_{\mathrm{r}}]/\delta_{\mathrm{r}}, \tag{2}$$

where the subscripts l and r correspond to the left and right video sequence frames, respectively; μ is the average brightness of the image window; δ is the parameter of light intensity distribution and is defined by

$$\delta^2 = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}[\beta(i,j) - \mu]^2. \tag{3}$$

1.2 Depth Estimation

Stereo matching of left and right views is the basis for calculating the disparity and depth. Stereo matching refers to finding the corresponding pixels in different images which are obtained from two perspectives when the same scene is observed. The essence is to find a matching path. The key is to construct an optimization model and find the optimal solution. The effectiveness of stereo matching algorithm lies in choosing the precise matching primitives, finding the appropriate matching criteria, and building a stable algorithm for solving process accurately.

The energy optimization based on graph cut is a common stereo matching algorithm. It is also an excellent method for calculating the disparity of stereo video. Based on the idea of minimum cut and maximum flow, the image blocks are selected as primitives to establish a matching path. The global energy function can be constructed according to the path cost. The matching path energy is minimized after the optimization calculation, and then the best matching can be achieved. The essence of the procedure is to convert the corresponding point matching into seeking a global

optimization solution of energy function. So the problem of disparity solving is converted into calculating the energy optimization[10]. First, the matching energy is calculated and the function is defined by

$$E(f) = E_{\text{smooth}}(f) + E_{\text{data}}(f), \tag{4}$$

where f is a labeling of image P, $E_{\text{smooth}}(f)$ measures the extent to which f is not piecewise smooth, and $E_{\text{data}}(f)$ measures the disagreement between f and the observed data. They are defined respectively by

$$E_{\text{smooth}}(f) = \sum_{\{p,q\}\in N} V_{\{p,q\}}(f_p, f_q), \tag{5}$$

$$E_{\text{data}}(f) = \sum_{p\in P} D_p(f_p), \tag{6}$$

where N is the set of pairs of adjacent pixels, $V_{\{p,q\}}(f_p, f_q) = |f_p - f_q|$ indicates the difference of adjacent pixels, and $D_p(f_p) = (f_p - I_p)^2$ indicates the intensity difference of the pixel p between measured data f_p and observed data I_p. Then a directed graph is built which includes nodes and non-negative weight edges. According to the principle of minimum cut and maximum flow, the iterative optimization scheme is used for solution and the directed graph is dynamically updated in the iteration[10].

In the system of binocular camera, disparity can be defined as vector difference of object points in each channel image associated with the focus. Binocular disparity is the difference of direction when a goal is observed from two points. The distance between the two points is called baseline. The relationship of disparity and depth is shown in Fig. 1, where M_{l} and M_{r} are the matching points, and O is the target point. The depth Z can be defined by

$$Z = BF/(x_{\text{l}} - x_{\text{r}}) = BF/d, \tag{7}$$

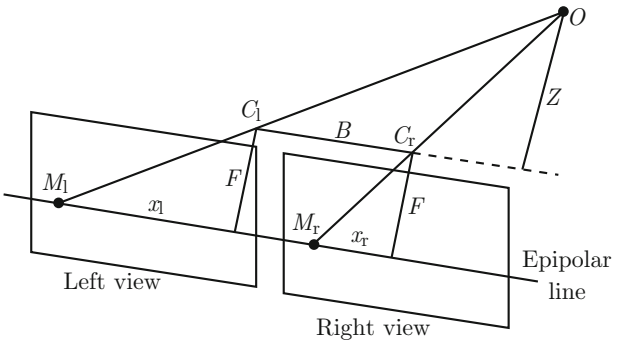


Fig. 1 Disparity-depth relationship

where B, F and d represent the camera baseline, focal length and disparity, respectively.

For stereo video, the depth maps can be represented by an 8-bit greyscale image to render new views. When the depth is represented by gray value from 0 to 255[5], the depth value can be defined as

$$\bar{Z} = \left\lfloor 255 - \frac{255(Z - Z_{\min})}{Z_{\max} - Z_{\min}} + 0.5 \right\rfloor, \tag{8}$$

where $Z_{\max}$ and $Z_{\min}$ represent the farthest and the nearest depth values, respectively.

2 Consistency in Depth Maps Estimation

The framework of proposed method in this paper for consistent depth maps is demonstrated in Fig. 2.

As described previously, the method of graph cut based on epipolar rectification and energy optimization is used for depth estimation. In temporal domain, the penalty function with constraint factor is introduced for temporal consistency; in spatial domain, smooth term in global optimization function is rebuilt for spatial

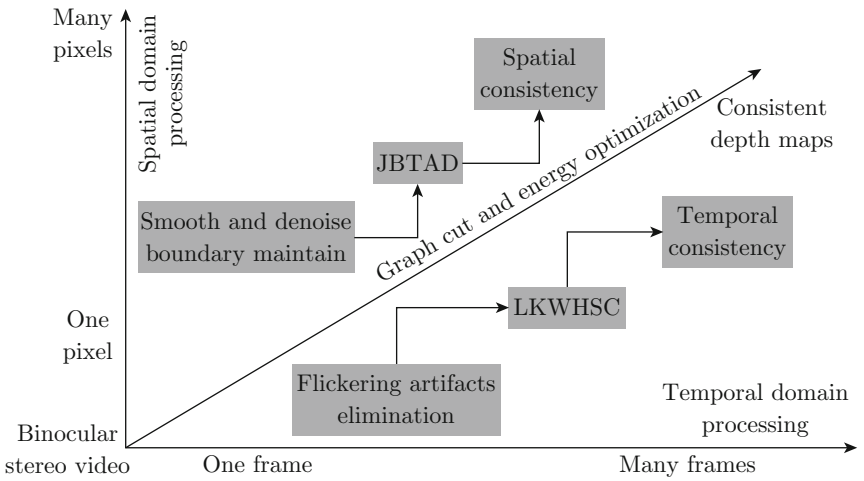


Fig. 2 The framework of proposed method

consistency. The global optimization function is redefined by

$$E(f) = E_{\text{data}}(f) + E_{\text{smooth}}(f) + E_{\text{occ}}(f), \quad (9)$$

where $E_{\text{occ}}(f)$ is a penalty function.

2.1 Temporal Consistency

In the global optimization function, the penalty function can be defined by

$$E_{\text{occ}}(f) = \sum_{p}(\lambda|d_p - d_i|), \quad (10)$$

where d_p represents the disparity of pixel p in current frame, d_i represents the disparity of the pixel which is most similar to p in the previous frame, and λ is a constraint factor. This paper proposes an algorithm of LKWHSC to obtain the factor and temporal consistency. In the algorithm, the weighted histograms of adjacent frames are calculated as the weights determined by pixel similarity measurement according to the principle of Lucas-Kanade optical flow. The number of pixels for similarity constraint between adjacent frames is determined by the quantified color histogram. The calculated values are compared with the color histogram of current frame according to Kullback-Leibler divergence. Based on the principle of brightness consistency, the comparative values are calculated for optimal solution, and then the temporal consistency of depth maps sequence can be achieved.

2.1.1 *Coherence Based on LKWHSC*

RGB is the most common color space in video, and most of the digital images are also expressed with the RGB color space. However, the spatial structure does not meet people in subjective judgment of color similarity. So it is necessary to convert it into HSV space which is closest to the subjective perception of human eyes[11]. Each component of HSV color space is quantified as non-equal interval for 8 steps, 3 steps and 3 steps. According to the steps of quantization above, these three components can be synthesized into one-dimensional feature vector. The HSV color space can be quantified for 72 steps[12]. Then the histogram distribution is calculated by

$$H(k) = \frac{n_k}{N_{\text{t}}} \quad (k = 0, 1, \cdots, K-1), \quad (11)$$

where n_k represents the number of pixels whose quantified color value is k, K represents the number of colors contained in the image, and N_{t} represents the total number of pixels within the image.

LKWHSC is proposed to associate adjacent frames for temporal consistency on the basis of optical flow method[7]. The color histogram value of current frame n_0 is set as $H_{n_0}(k)$, and the coherence constraint with the adjacent frame n is defined by

$$H_{n_0,n}(k) = w_{k_0}(p_{k_0}, p_k)H_n(k), \quad (12)$$

where $H_n(k)$ is the histogram value in each dimension of the quantized frame n, and $w_{k_0}(p_{k_0}, p_k)$ is similarity measure between the pixel in each histogram and the pixel in corresponding spatial of adjacent frame. The similarity measure is defined by

$$w_{k_0}(p_{k_0}, p_k) = \exp\Big[-\sum_{p_h \in N_{\text{n}}(p_k)} \rho|I_n(p_h) - I_{n_0}(p_{k_0})|\Big], \quad (13)$$

where $N_{\text{n}}(p_k)$ denotes the set of neighbors of pixel p_k, p_{k_0} is pixel of the current frame n_0, p_k is pixel of adjacent frame n, p_k is also the similar pixel of p_{k_0} and is denoted as $p_k \sim p_{k_0}$, p_h is the neighborhood pixel of p_k, $|I_n(p_h) - I_{n_0}(p_{k_0})|$ is the Euclidean distance of color brightness, and ρ is the weighting coefficient of color difference to adjust similar ratio between pixels[13]. In the proposed algorithm, K pixels of each frame are selected and they are combined with their corresponding histogram values for similar constraint. It can reduce computational complexity and improve constraint effect compared with the algorithm that a single pixel is selected[7] and the algorithm that all pixels are selected[14].

Shot segmentation is help for every scene epipolar rectification and can be applied in different matching disparity range. Moreover, the errors of scene transition, forward reference for start frame and backward reference for end frame can be avoided. Three-dimensional video shot boundary detection algorithm is introduced for video shot segmentation[15].

The reference constraint with forward and backward adjacent frames can be constructed based on shot segmentation, so Eq. (12) can be modified as

$$H_{n_0,n}(k) = w_{k_0}(p_{k_0}, p_k)H_n(k)|_{n \in T_{\text{prev}}(n_0)} + w_{k_0}(p_{k_0}, p_k)H_n(k)|_{n \in T_{\text{next}}(n_0)}, \quad (14)$$

where $T_{\text{prev}}(n_0)$ and $T_{\text{next}}(n_0)$ denote the forward and backward adjacent frames of n_0, respectively. There is no forward constraint reference frame if the current frame is the beginning of shot. Similarly, there is no backward constraint reference frame if the current frame is the ending.

2.1.2 *The Optimal Solution*

According to the principle of Kullback-Leibler divergence, the similarity can be calculated between the histogram value $H_{n_0}(k)$ of current frame n_0 and the constraint value $H_{n_0,n}(k)$, which is defined by

$$D_{\text{KL}}(H_{n_0,n}(k)||H_{n_0}(k)) = \sum_{k=0}^{K-1} H_{n_0,n}(k)\ln\frac{H_{n_0,n}(k)}{H_{n_0}(k)}. \quad (15)$$

If

$$H_{n_0,n}(k) = H_{n_0}(k),$$

then

$$D_{\mathrm{KL}}(H_{n_0,n}(k)\|H_{n_0}(k))=0.$$

Symmetry calculation is performed and the expression is defined by

$$D_{\mathrm{KL}}=\frac{1}{2}\Big[\sum_{k=0}^{K-1}H_{n_0,n}(k)\ln\frac{H_{n_0,n}(k)}{H_{n_0}(k)}+\sum_{k=0}^{K-1}H_{n_0}(k)\ln\frac{H_{n_0}(k)}{H_{n_0}^{n}(k)}\Big]. \tag{16}$$

In video sequence, pixel intensity should keep consistency between current frame and adjacent frames in instantaneous motion. The energy function in temporal domain constraints the movement of pixels between adjacent frames, and the variation of luminance value should be minimized[13]. According to the method of Lucas-Kanade optical flow, the pixels can be considered constant when the time interval between adjacent frames is very short and the change of image gray is as little as possible. The frame n_0 is in time of t_0, and the adjacent frame n is in time of t. If $\Delta t=t-t_0\to 0$ then $p_h\to p_{k_0}$, there is

$$\lim_{\Delta t\to 0}D_{\mathrm{KL}}\to 0. \tag{17}$$

The minimum value of $|I_n(p_h)-I_{n_0}(p_{k_0})|$ can be obtained by iteration and limitation solving. So the neighboring pixels of p_h can be normalized to p_k. Because p_k is the most similar pixel of p_{k_0}, it can be regarded as an equivalent pixel, which is denoted by $p_k\approx p_{k_0}$.

According to the solution of equivalent pixels, the value of weighted histogram can be calculated and it is also the corresponding constraint factor λ. Therefore, the penalty function can be represented as

$$E_{\mathrm{occ}}(f)=\sum_{k=0}^{K-1}H_{n_0,n}(k)\sum_{p}(|d_p-d_i|). \tag{18}$$

2.2 Spatial Consistency

It is necessary to smooth and reduce noise in order to obtain high quality and spatial consistent depth maps of stereo video. Based on the local consistency of pixels, smooth energy function can be defined for each pixel in spatial domain so that the function can smooth uniformly in low gradient region and maintain the properties in high gradient region[13]. Segmentation smoothing is necessary for depth maps sequence estimation in spatial domain. It needs not only to smooth, reduce noise and realize consistency, but also to keep the disparity information and object edge sharpness as much as possible[16].

This paper proposes the algorithm of JBTAD in spatial domain to smooth naturally in low-gradient region, avoid over-smoothing and preserve edges of discontinuities in high-gradient discontinuities[17]. In the algorithm, smooth term in global optimization function is rebuilt, and it includes the similarity term as well as the smooth and boundary maintaining term. For the similarity term, the color similarity truncated absolute difference is adopted; while the smooth and boundary maintaining term is the bilateral filtering function which contains pixel space Gauss kernel function and depth Gauss kernel function. As it is known, the depth for the neighboring pixels has strong correlation in the same plane, and so does the disparity. The smooth term is defined as

$$E_{\mathrm{smooth}}(f)=\sum_{p}\sum_{q\in N_{\mathrm{n}}(p)}\phi(p,q)\rho(T_p,T_q). \tag{19}$$

In Eq. (19),

$$\phi(p,q)=\frac{1}{\omega_p}Q_{\mathrm{s}}(||p-q||)Q_{\mathrm{d}}(||d_p-d_q||)d_q, \tag{20}$$

where ω_p is the regularization factor, Q_{s} is the spatial weighted Gaussian kernel function centered by pixel p, and Q_{d} is the weighted Gaussian kernel of depth difference. The expression is defined to process noise, preserve edges and encourage the disparity discontinuity to be coincident with abrupt intensity/color change for depth maps sequence. The two kernel functions are defined respectively by

$$Q_{\mathrm{s}}(\|p-q\|)=\exp\Big(-\frac{1}{2}\frac{|p-q|^2}{\sigma_{\mathrm{s}}^2}\Big), \tag{21}$$

$$Q_{\mathrm{d}}(\|d_p-d_q\|)=\exp\Big(-\frac{1}{2}\frac{|d_p-d_q|^2}{\sigma_{\mathrm{d}}^2}\Big), \tag{22}$$

where σ_{s} and σ_{d} denote the spatial and depth width parameters, respectively[4]. They have effects on the spread range of functions and smoothness. In Eq. (19),

$$\rho(T_p,T_q)=\min(|T_p-T_q|,\eta) \tag{23}$$

is defined to realize smooth uniformly between neighboring pixels of objects in space-time as we enforce that the disparity smoothness assumption values should vary smoothly except object boundaries[1], where η is the threshold and the value is set from 0.05 to 0.1 according to the changing of smoothing gradient, and $|T_p-T_q|$ is the color similarity measurement of pixel and can be represented by Euclidean distance of R, G, B color components, which is defined by

$$|T_p-T_q|=|R_p-R_q|+|G_p-G_q|+|B_p-B_q|. \tag{24}$$

Many of the existing algorithms usually can obtain sparse or low-resolution depth maps by stereo matching and depth calculation. In the algorithm, the spatial depth super-resolution algorithm of iterative bilateral

filtering for two-view is introduced as a post-processing stage[18]. It executes up-sampling iteratively to enforce the full resolution of input depth maps. In our implementation, the number of iterations is 3. The better effects of depth maps and running efficiency are achieved.

3 Experimental Results

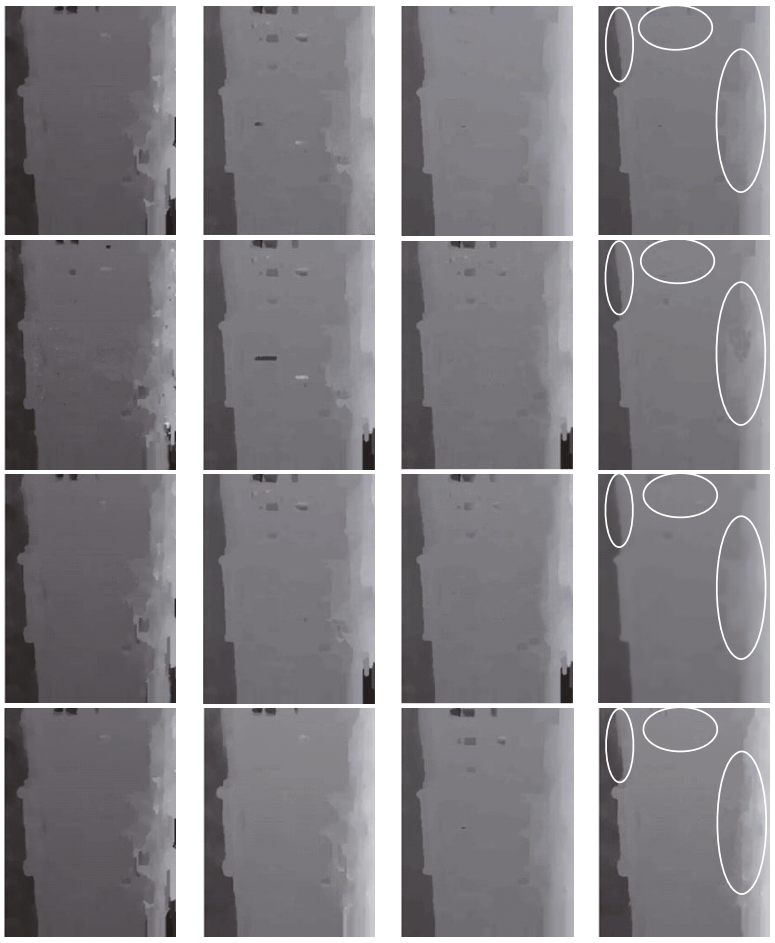
In our experiment, we use the test sequences "Street", "Tanks", "Temple" and "Tunnel" as well as the ground truth disparities provided by the University of Cambridge Computer Laboratory for experimental implementation and evaluation of results[4]. The length of each sequence is 100 frames, and each frame is 400 pixel × 300 pixel in resolution with a disparity range of 64 pixels. For experimental environment, we use Windows 7 of 32-bit dual-core processor whose frequency is 3.3 GHz. MATLAB 7.10.0 is used for algorithm simulation. Experimental results are compared with the DCBGrid, TDCBGrid[4] and IP algorithms[6]. Average percentages of bad pixels (α) and mean squared errors of pixels (MSE) for all frames of each sequence in different algorithms are demonstrated in Table 1 and Table 2, respectively. The subjective comparison of depth maps is shown in Fig. 3.

Table 1 Average percentages of bad pixels for all frames

Algorithm	α/%			
	Street	Tanks	Temple	Tunnel
DCBGrid	20.70	17.83	24.62	14.83
TDCBGrid	16.23	17.18	19.21	23.14
IP	13.79	16.43	10.77	13.99
The proposed	10.52	12.75	9.03	11.84

Table 2 Mean squared errors for all frames

Algorithm	MSE			
	Street	Tanks	Temple	Tunnel
DCBGrid	27.04	23.91	41.34	19.71
TDCBGrid	18.58	25.40	33.06	26.42
IP	16.89	22.56	12.46	18.49
The proposed	4.77	8.53	3.66	8.12



(a) DCBGrid (b) TDCBGrid (c) IP (d) The proposed

Fig. 3 Comparison of consistent depth maps in subjective effects

From Table 1, the proposed algorithm always has lower average percentage of bad pixels in spatial-temporal stereo matching for each sequence. Similarly, in Table 2 compared to the other algorithms, the proposed method always has smaller average mean squared errors for the test sequences.

Experiments are carried out on the test sequences. Figure 3 only lists four corresponding depth maps of the "Street" sequence in different algorithms. According to the experimental results, as demonstrated in Fig. 3, the proposed method can reduce the flickering artifacts and obtain more consistent depth maps sequence frames compared with the other algorithms. At the same time, it can be also seen that the proposed method can obtain smoother depth maps and more efficiently and clearly maintain the boundary of objects in discontinuities.

The Gaussian noise is added in depth maps to validate the robustness of the proposed method. The range of standard deviation σ is from 0 to 100. The average percentage of bad pixels gradually increases with the addition of noise for each algorithm. Compared with the DCBGrid, TDCBGrid and IP algorithms, the average percentage of bad pixels of the proposed method is nearly always smaller in all cases. Experimental results show that the proposed algorithm has better robustness. The comparison results are demonstrated in Fig. 4.

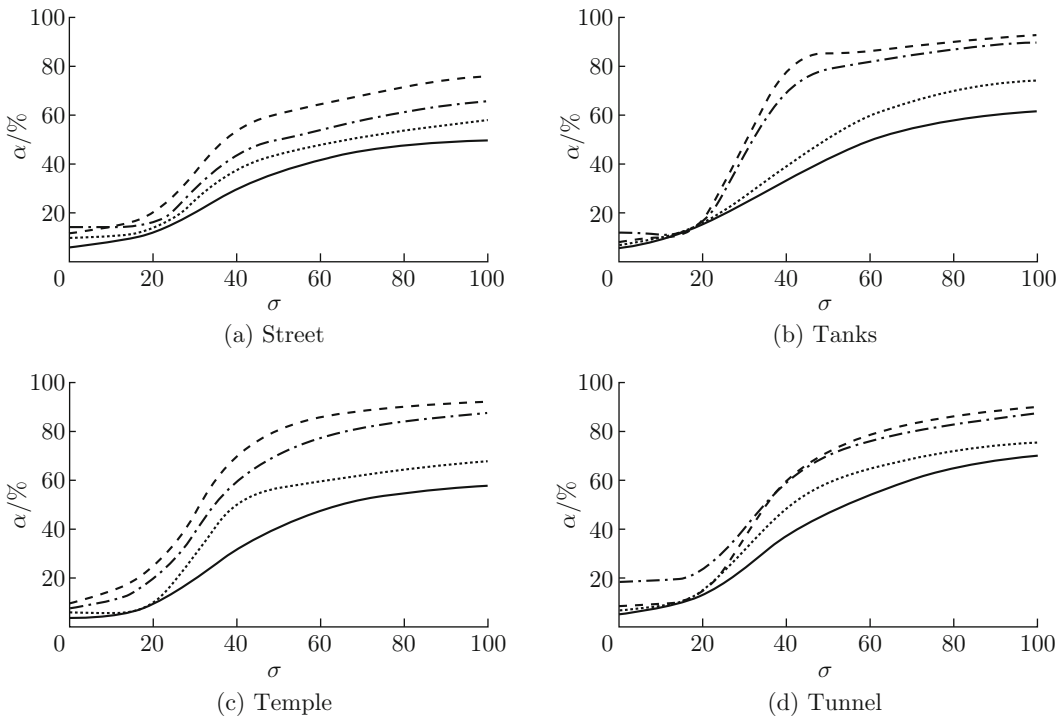


Fig. 4 Comparison of robustness with Gaussian noise

4 Conclusion

This paper proposes a method of LKWHSC in temporal domain and JBTAD in spatial domain for spatial-temporal consistency to eliminate the flickering artifacts and smooth inaccuracy in binocular video depth maps sequence estimation. In temporal domain, shot segmentation and appropriate selection number of the pixels can efficiently improve the matching accuracy and reduce the flickering artifacts. In spatial domain, smooth and noise reduction as well as object edge sharpening can be achieved. Future work will focus on the optimization for time complexity of correlation constraint and spatial smoothing.

References

[1] KHOSHABEH R, CHAN S H, NGUYEN T Q. Spatio-temporal consistency in video disparity estimation [C]//*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Prague, Czech Republic: IEEE, 2011: 885-888.

[2] CIGLA C, ALATAN A A. Temporally consistent dense depth map estimation via belief propagation [C]//*IEEE 3DTV Conference on the True Vision—Capture, Transmission and Display of 3D Video*. Potsdam, Germany: IEEE, 2009: 1-4.

[3] GARCIA F, AOUADA D, MIRBACH B, et al. A new multi-lateral filter for real-time depth enhancement [C]//*IEEE International Conference on*

Advanced Video and Signal-Based Surveillance. Klagenfurt, Austria: IEEE, 2011: 42-47.

[4] RICHARDT C, ORR D, DAVIES I, et al. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid [C]//*Proceedings of the European Conference on Computer Vision*. Hersonissos, Greece: Springer-Verlag, 2010, **6313**: 510-523.

[5] LEE S B, HO Y S. Temporally consistent depth map estimation for 3D video generation and coding [J]. *China Communications*, 2013, **10**(5): 39-49.

[6] PHAM C C, NGUYEN V D, JEON J W. Efficient spatio-temporal local stereo matching using information permeability filtering [C]//*IEEE International Conference on Image Processing*. Orlando, USA: IEEE, 2012: 2965-2968.

[7] MIN D B, LU J B, DO M N. Depth video enhancement based on weighted mode filtering [J]. *IEEE Transactions on Image Processing*, 2012, **21**(3): 1176-1190.

[8] FUSIELLO A, IRSARA L. Quasi-Euclidean uncalibrated epipolar rectification [C]//*19th International Conference on Pattern Recognition*. Tampa, USA: IEEE, 2008: 1-4.

[9] HEO Y S, LEE K M, LEE S U. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(5): 1094-1106.

[10] BOYKOV Y, VEKSLER O, ZABIH R. Fast approximate energy minimization via graph cuts [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(11): 1222-1239.

[11] QIN S, XIE G. LW-PGD method and fusion feature based real-time pedestrian detection in video [J]. *Journal of Computational Information Systems*, 2014, **10**(6): 2273-2281.

[12] JIANG L C, SHEN G Q, ZHANG G X. An image retrieval algorithm based on HSV color segment histograms [J]. *Mechanical & Electrical Engineering Magazine*, 2009, **26**(11): 54-57 (in Chinese).

[13] ZHANG Y, ZHANG J W, YANG G Q, et al. Video de-hazing using spatial-temporal coherence optimization [J]. *Application Research of Computers*, 2011, **28**(10): 3983-3985 (in Chinese).

[14] BUADES A, COLL B, MOREL J M. Nonlocal image and movie denoising [J]. *International Journal of Computer Vision*, 2008, **76**(2): 123-139.

[15] FERREIRA L, ASSUNCAO P, DA SILVA CRUZ L A. 3D video shot boundary detection based on clustering of depth-temporal features [C]//*2013 11th International Workshop on Content-based Multimedia Indexing*. Veszprem, Hungary: IEEE, 2013: 1-6.

[16] MA G H, WANG C, LIU P, et al. Sequential similarity detection algorithm based on image edge feature [J]. *Journal of Shanghai Jiaotong University* (*Science*), 2014, **19**(1): 79-83.

[17] ZHANG G F, JIA J Y, WONG T T, et al. Consistent depth maps recovery from a video sequence [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(6): 974-988.

[18] YANG Q X, YANG R G, DAVIS J, et al. Spatial-depth super resolution for range images [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, USA: IEEE, 2007: 1-8.