TEMPORAL AND COLOR CONSISTENT DISPARITY ESTIMATION IN STEREO VIDEOS

Nicholas Vretos, Member, IEEE and Petros Daras, Senior Member, IEEE

Centre for Research and Technology Hellas
Information Technologies Institute
6th Km Charilaou-Thermi Road, Thessaloniki, Greece
e-mail:{vretos, daras}@iti.gr

ABSTRACT

In this paper, a novel method for stereo videos disparity estimation is proposed. Temporal consistency in the disparity and color spaces of the video are combined to enhance the disparity estimation algorithm. Outlier detection along the temporal dimension in the color space is used to find areas where disparity can be temporally enhanced from previous and future frames. Moreover, a forward-backward approach is utilized for consistency checking and error correction in cases of fast motion in the video, which usually introduce disparity artifacts. The method can work in real-time even in the hard case of high definition (HD) videos. Experimental results prove that the proposed approach outperforms state of the art methods in a publicly available dataset.

Index Terms— Spatiotemporal disparity, stereo matching, disparity temporal consistency

1. INTRODUCTION

Disparity estimation from stereo pair cameras is a very old research field dating back to 1838 when Sir Charles Wheatstone first described the principles of stereopsis [1]. Since then, a large amount of research effort has been invested in diverse research fields to better understand the notion of stereo disparity and the potential applications of stereopsis. Nowadays, with the advent of 3DTV and 3D movie technologies, stereo disparity estimation has become a hot topic in the computer vision research community with the main goal of achiving better stereo disparity estimation from stereo camera pairs. There has been a significant research effort in the last years towards stereo disparity estimation from image pairs with more than 150 published works competing in the Middlebury stereo evaluation site [2]. However, in the case of stereo video disparity estimation in a non frame-by-frame approach the work presented so far is significant less.

In the case of stereo image pairs we can categorize the existing methods in local and global ones. While local [3]-[5] methods try to calculate a dense disparity mapping by assuming that within a certain spatial window the disparity is constant, global ones [6]-[9] are minimizing a global function with respect to a disparity assignment. Both approaches have their merits with global based ones being more accurate in details and more robust to textureless areas, while local based methods being less accurate but much more faster and suitable for real-time applications. A recent survey in disparity estimation methods can be found in [10].

Although a vast amount of research effort has been invested in calculating the disparity between image pairs, only few works have reported results in video disparity estimation, while making use of the temporal dimension of the video. The term spatiotemporal or spacetime stereo matching was first introduced in [11] and [12] and account for the stereo matching algorithms that use the temporal dimension of a stereo video to estimate the disparity of each frame. In these papers, the authors proposed a spatiotemporal support window to calculate the disparities for one frame. In [13], the difference between two consecutive disparity frames was taken into account to enforce temporal consistency. These rather old works, were mostly used either in the case of still videos or in ones with very little motion and consequently do not work well in the case of greater amount of motion within a scene. To cope with this problem, [14] and [15] used the optical flow to enhance the disparity estimation. However, these methods by incorporating the optical flow to the disparity estimation process have to deal with a much more complicated problem that has implications to speed and accuracy of the final output. Some recent works have successfully implemented local methods and incorporated the temporal dimension to enforce temporal consistency of the disparity. In [16], Richardt *et al.* used the method in [5] and also a fast approximation of the bilateral filter to aggregate the costs in the spatiotempral domain. Finally, in [17] the authors proposed a three-pass aggregation process and the use of information permeability for stereo matching, where multiple images are used to calculate the disparity of each frame. By doing so, they do not need to estimate the motion within the frames.

In this paper, we propose a novel approach that takes into account a temporal window to calculate the disparity and enforce its temporal consistency for those pixels where the color remains within a certain distribution. The novelty of the approach, is that by using an outlier detection method along the temporal dimension in the color space we can correct motion based errors that are induced in simple disparity post processing techniques. By doing so, we provide a real-time spatiotemporal stereo matching technique that can also be applied to high definition images. More specifically, we first calculate the outliers, based on the color information, within a temporal frame for each pixel and therefore a median filter is applied to the disparity values of the remaining pixels. That way, disparity temporal consistency is enforced in those areas where the color do not have hard/adrupt changes (as is the case of edges).

The rest of the paper is organized as follows: the proposed method is outlined in Section 2. In Section 3, the spatiotemporal disparity estimation framework is detailed and in Section 4 experiments are illustrated. Finally, in Section 5 conclusions are drawn.

2. PROPOSED METHOD

Our method aims at finding temporal outliers in the color space, thus it calculates the median of the disparity values for the inliers that are found using the Grubbs test [18]. We apply this method to each color channel and a pixel is considered an inlier, if and only if it is in every color channel that we apply the Grubbs test. In case where the pixel is an outlier in any of the color channels it is considered an outlier. More specifically, let us suppose that $I(x, y, c, t)$ is the value of a pixel in position (x, y), in the c color channel at time t. Moreover, let $\mathbf{r}$,$\mathbf{g}$ and $\mathbf{b}$ are the vectors which contain the values of $I(x, y, c, t)$ for all t in a certain predefined temporal window of size N for each color channel c, respectively. To decide if a pixel is an outlier we do the following:

- we calculate the vector mean as $\bar{r}$, $\bar{g}$ and $\bar{b}$ as well as their standard deviations σ_r, σ_g and σ_b,
- for each element of these vectors we calculate the absolute deviation from the mean as $\delta_{r_i} = |r_i - \bar{r}|$, $\delta_{g_i} = |g_i - \bar{g}|$ and $\delta_{b_i} = |b_i - \bar{b}|$ (where r_i (resp. g_i and b_i) is the i-th element of vector $\mathbf{r}$ (resp. $\mathbf{g}$ and $\mathbf{b}$)),
- finally, we decide whether to keep an element if $\delta_{r_i} \leq \tau\sigma_r$, $\delta_{g_i} \leq \tau\sigma_b$ and $\delta_{b_i} \leq \tau\sigma_b$ are all valid for the same i, where τ the Grubbs statistic calculated as:

$$\tau = \frac{N - 1 \cdot \sqrt{\left(\tau^2_{\frac{\alpha}{2N},(N-2)}\right)}}{\sqrt{N \cdot \left(N - 2 + \tau^2_{\frac{\alpha}{2N},(N-2)}\right)}}, \tag{1}$$

where $\tau_{\frac{\alpha}{2N},(N-2)}$ denote the upper critical value of the t-distribution with N-2 degrees of freedom and significance level of $\frac{\alpha}{2N}$ and α the significant level for the rejection of the hypothesis of no outlier.

Once we have calculated the set of inlier pixels in the color space throughout the temporal dimension, we create a vector $\mathbf{d}$ containing the respective disparities for the specific temporal window as calculated using the Semi-Global Block Matching (SGBM) algorithm in [19]. Finally, to enhance robustness towards motion, a forwards-backwards approach is used. That is, in order to calculate the disparity at pixel (x, y) for a frame in time t we use a range of frames within $\left[t - \lfloor\frac{N}{2}\rfloor, t + \lfloor\frac{N}{2}\rfloor\right]$, where $\lfloor\cdot\rfloor$ is the largest previous integer. Moreover, since in cases of very fast motion (i.e., low frame rates) pixels may be very distant in the color domain, we calculate the median value of the disparity backward (i.e, in the range $\left[t - \lfloor\frac{N}{2}\rfloor, t - 1\right]$) and forward (i.e, in the range $\left[t + 1, t + \lfloor\frac{N}{2}\rfloor\right]$) and test if the difference between the two median values is close enough with respect to a threshold. In the opposite case the algorithm keeps the value of the pre calculated disparity without applying the correction step. Finally, as it is common in most of the disparity estimation methods [10] we apply a simple 3×3 spatial median filter on the resulted disparity image.

3. SPATIOTEMPORAL DISPARITY ESTIMATION

In this section the framework for video stereo disparity estimation is detailed. The proposed framework consists of 3 distinct steps: a) SGBM disparity calculation per frame, b) temporal correction of disparity values and c) post processing of the resulted disparity. At the first step the SGBM algorithm is used to calculate the disparity of each frame. Therefore, in the second step, the resulted disparity mapping is processed in a per pixel bases and corrects values that are not outliers, as defined in the previous section, based on the already calculated disparity of the previous frames as well as in future frames. In the cases of real-time applications, a buffering process allows for future frames to be present in the calculation process with the drawback of a half temporal window latency. Finally, post processing of the disparity mapping can be applied in the last step once the disparity estimation process is concluded. Such post processing tasks can be spatial median filtering, bilateral filtering, erosion and dilation, as well as other methods for disparity enhancement [10]. In our case, we have used a spatial median filter with a 3 spatial window and a bilateral filtering of the disparity mapping. By applying these post processing filters we can smooth the disparity mapping (median filter) and in the same time preserve the edge information (bilateral filtering). The proposed framework is illustrated in Figure 1.

4. EXPERIMENTAL RESULTS

We have tested the proposed approach in two cases. One with the publicly available database [16] and one with high definition stereo video from the commercial stereo camera SONY-HDR-TD30Ve. The first set of experiments was selected for

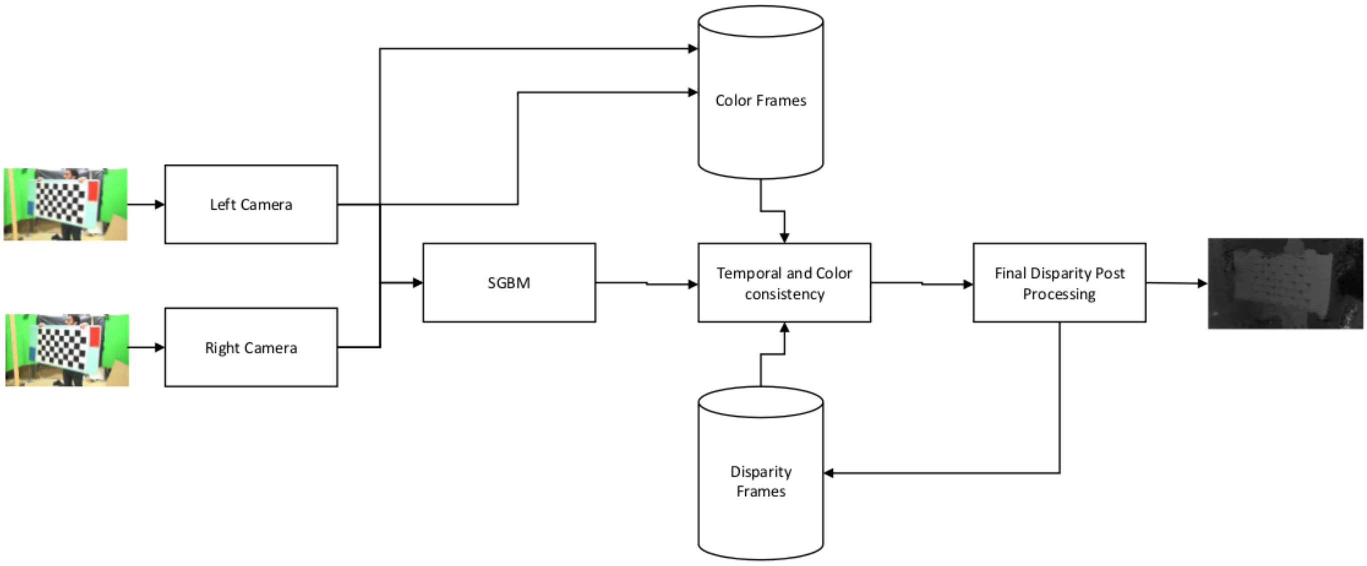
Color Frames
Left Camera
SGBM
Temporal and Color consistency
Final Disparity Post Processing
Right Camera
Disparity Frames

Fig. 1: Disparity estimation framework

comparison with other state of the art methods in spatiotemporal disparity estimation and the latter in order to establish evidence of the algorithm efficiency in working with real data in real time. The first dataset consists of 5 stereo sequence named Book, Street, Tanks, Temple and Tunnel with a resolution of 400 × 300 pixels and variable frame lengths ranging from 41 to 100 frames. Moreover, in order to be consistent with the results reported in [17] we have added noise in the same manner as in their case (i.e., Gaussian zero-mean noise with $\sigma = 20$ and therefore bilateral filtering of the images with $r = 3$, $s_s = 2$ and $s_c = 0.3$ where r, s_s and s_c the radius, spatial and color variance parameters of the filter, respectively).

To assess the performance of the proposed method, we use two metrics defined in [2], the bad pixel percentage and the root mean square error. These two metrics are defined as:

$$B = \frac{1}{K} \sum_{(x,y)} |d_c(x,y) - d_g(x,y)| < \delta_d, \tag{2}$$

and

$$R = \sqrt{\frac{1}{K} \sum_{(x,y)} |d_c(x,y) - d_g(x,y)|^2}, \tag{3}$$

respectively, where K is the total number of pixels, $d_c(x,y)$ the calculated disparity for pixel (x,y), d_g the ground truth disparity for pixel (x,y) and δ_d the disparity error tolerance (in our case it is set to 1). R is calculated in disparity units while B is a percentage. In our results we average the values across the frames as in [17]. Finally, to provide hard evidence of the additional value that our methods brings into the framework of spatiotemporal disparity estimation, we have also run experiments with the core disparity estimation technique that we have used (i.e., the SGBM algorithm). By doing so, we can assess the individual added value of the proposed framework towards the frame-by-frame approach using the same technique. Results are depicted in Table 1 and Table 2.

Table 1: Average percentages of bad pixels across frames

Method	Book	Street	Tanks	Temple	Tunnel
SGBM [19]	16.07	24.16	16.85	30.65	3.25
Pham[17]	19.39	**13.79**	16.43	**10.77**	13.99
Proposed	**14.41**	17.01	**15.26**	24.17	**2.21**

Table 2: Average percentages of root mean squared error across frames

Method	Book	Street	Tanks	Temple	Tunnel
SGBM [19]	10.24	3.70	7.65	25.41	1.47
Pham[17]	3.82	4.11	4.57	**3.53**	4.30
Proposed	**1.47**	**1.74**	**3.62**	19.37	**0.51**

In all the cases demonstrated, the proposed framework gives better results than the simple application of the SGBM algorithm in a frame-by-frame basis. Nevertheless, in some cases our algorithm does not outperform the results reported in [17]. This is due to the big errors of the SGBM algorithm towards occlusion handling and therefore even if our method ameliorates its results the overall accuracy is lower than state of the art methods. However, applying simple occlusion handling techniques like the ones mentioned in [10] one may reach to better results. In order to assure fair comparison between our techniques and others we keep the framework as simple as possible. Unfortunately, in [17], there is no discussion on occlusion handling so we preferred to keep the

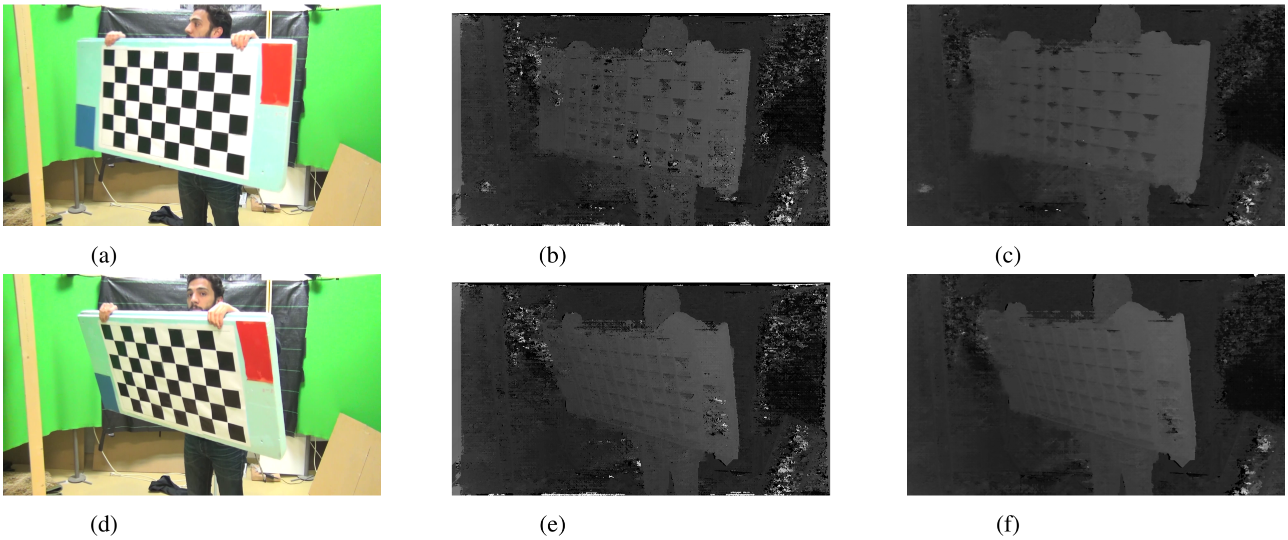

Fig. 2: (a),(d): Two different left frames from the video sequence. (b),(e): Disparity Estimation with SGBM per frame. (c),(f): disparity estimation with the proposed method

Table 3: Average percentages of bad pixels (B) and root mean squared error(R) across frames with occlusion handling

Method	Book	Street	Tanks	Temple	Tunnel
B	13.10	11.10	15.03	9.74	2.20
R	1.07	1.07	3.40	3.26	0.50

framework as is for comparison. For the sake of completeness though we provide results after applying simple occlusion handling of left-right consistency in Table 3.

In the case of occlusion handling all our results outperform state of the art methods. Finally, the time performance is in the order of 10 frames per second (fps), that is 0.1s in average. All tests has been performed on an Intel®Core$^{\text{TM}}$i7 with 3.50GHz CPU and 16GB RAM.

The second series of experiments provide evidence of the functionality of the proposed method towards HD video from real scene data in real time. In Figure 2 the left frame and the estimated disparity are illustrated before and after the application of the proposed method. Moreover, the time consumed for the case of HD video is in the order of 3-4 fps (0.30 s) per video in average. It can be clearly shown on these images the amelioration of the final disparity mapping due the the proposed temporal disparity enhancement.

5. CONCLUSIONS AND FUTURE WORK

In this work, a new spatiotemporal disparity estimation algorithm is proposed making use of the color information to correct motion based errors. Moreover, a forward-backward approach is utilized to better handle fast motion artifacts that can be apparent in cases of fast motion or low frame rates in capturing. The algorithm can work in real time in cases of videos of standard resolution (up to 640 × 480) and in near real time (3-4 fps) in cases of HD videos (up to 1920 × 1080).

The intuition behind the proposed method is that in the temporal domain color and disparity not only vary smoothly but also the smoothness of color and disparity should vary in the same manner. This assumption allowed us to work in the color space so as find outliers end better estimate the disparity by aggregating disparity values in the temporal domain. Experimental results proved that our approach outperforms state of the art methods in spatiotemporal disparity estimation based and also that the proposed algorithm's time consumption may be maintained under the real-time constraint.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement no 287966 (EXPERIMEDIA)

6. REFERENCES

[1] Charles Wheatstone, *The Scientific Papers of Sir Charles Wheatstone*, Cambridge University Press, 2011.

[2] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[3] Geoffrey Egnal, "Mutual information as a stereo correspondence measure," *Technical Reports (CIS)*, p. 113, 2000.

[4] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 229–246, 2002.

[5] Kuk-Jin Yoon and In So Kweon, "Adaptive support-weight approach for correspondence search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 650–656, 2006.

[6] Michael Bleyer and Margrit Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 59, no. 3, pp. 128–150, 2005.

[7] Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewénius, and David Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 492–504, 2009.

[8] Cheng Lei, Jason Selzer, and Yee-Hong Yang, "Region-tree based stereo using dynamic programming optimization," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2378–2385.

[9] Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum, "Symmetric stereo matching for occlusion handling," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 399–406.

[10] Béatrice Pesquet-Popescu, "Disparity estimation techniques," *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*, p. 81, 2013.

[11] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 2, pp. II–359.

[12] Li Zhang, Brian Curless, and Steven M Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 2, pp. II–367.

[13] Carlos Leung, Ben Appleton, Brian C Lovell, and Changming Sun, "An energy minimisation approach to stereo-temporal dense reconstruction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 4, pp. 72–75.

[14] Minglun Gong, "Enforcing temporal consistency in real-time stereo estimation," in *Computer Vision–ECCV 2006*, pp. 564–577. Springer, 2006.

[15] Michael Bleyer and Margrit Gelautz, "Temporally consistent disparity maps from uncalibrated stereo videos," in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*. IEEE, 2009, pp. 383–387.

[16] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, and Neil A Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Computer Vision–ECCV 2010*. 2010, pp. 510–523, Springer.

[17] Cuong Cao Pham, Vinh Dinh Nguyen, and Jae Wook Jeon, "Efficient spatio-temporal local stereo matching using information permeability filtering," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 2965–2968.

[18] Peter J Rousseeuw and Annick M Leroy, *Robust regression and outlier detection*, vol. 589, Wiley. com, 2005.

[19] Heiko Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 328–341, 2008.