DISPARITY SEARCH RANGE ESTIMATION: ENFORCING TEMPORAL CONSISTENCY

Dongbo Min[1,2], *Sehoon Yea*[1], *Zafer Arican*[3] *and Anthony Vetro*[1]

Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA[1]
Yonsei University, Seoul, Korea[2]
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland[3]

ABSTRACT

This paper presents a new approach for estimating the disparity search range in stereo video that enforces temporal consistency. Reliable search range estimation is very important since an incorrect estimate causes most stereo matching methods to get trapped in local minima or produce unstable results over time. In this work, the search range is estimated based on a disparity histogram that is generated with sparse feature matching algorithms such as SURF. To achieve more stable results over time, we further propose to enforce temporal consistency by calculating a weighted sum of temporally-neighboring histograms, where the weights are determined by the similarity of depth distribution between frames. Experimental results show that this proposed method yields accurate disparity search ranges for several challenging stereo videos and is robust to various forms of noise, scene complexity and camera configurations.

Index Terms— Disparity search range, temporal consistency, disparity histogram, feature matching

1. INTRODUCTION

For decades, the correspondence problem has been one of the most important issues in the field of computer vision. Various stereo matching algorithms have been proposed to address this problem and recent years have witnessed significant improvements in those algorithms including the ones with real-time or near real-time performance [1][2].

Dense disparity maps acquired by stereo matching can be used in many applications, including image-based rendering, 3-D scene reconstruction, robot vision, tracking, etc. Such applications of the stereo matching method often presume the knowledge of appropriate search range for disparity or simply use a fixed range. In practice, an appropriate search range of a scene is needed to facilitate the use of any stereo matching algorithm. The lack of a search range often implies the need to search over a wider range of candidate disparity values, which generally requires more computation and memory. But more importantly, most stereo matching algorithms are more likely to get trapped in local minima when given an inappropriate search range, which might result in significantly compromised quality of the disparity map. However, the acquisition of appropriate disparity ranges is by no means a trivial issue especially when, for example, the scene or camera configuration changes over time.

Cyganek and Borgosz [3] proposed a way of estimating the maximum disparity range based on statistical analysis of the spatial correlation between stereo images, which they called image variograms. However, this method assumes that there are only positive disparities between stereo images. Another disparity search range estimation approach was proposed based on Confidently Stable Matching [4]. The disparity search range was obtained by setting an initial search range to the size of the whole image and then performing the proposed matching in a hierarchical manner [5]. This method appears to work well for several stereo images, but temporal aspects are not considered. There are also depth estimation techniques that directly impose temporal constraints as part of the estimation process, for example [6], but without an appropriate search range such techniques are still prone to false matches and incorrect estimation results.

In this paper, we propose a novel approach for estimating the disparity search range in the stereo video. The search range is estimated with a disparity histogram, which can be generated by a sparse or dense feature matching algorithm. We further propose a method to enforce temporal consistency in the estimation of the disparity search range. The proposed techniques produce a temporally-consistent estimate of the disparity range that is robust to various forms of noise, scene complexity and camera configurations.

The remainder of this paper is organized as follows. In Section 2, we discuss the disparity search range estimation based upon histograms. In Section 3, we describe a novel method to improve the temporal consistency of the histogram-based method. We present the experimental results in Section 4, and conclude the paper in Section 5.

2. DISPARITY SEARCH RANGE ESTIMATION

This work forms a disparity histogram to estimate the disparity search range for stereo video. The search range is com-

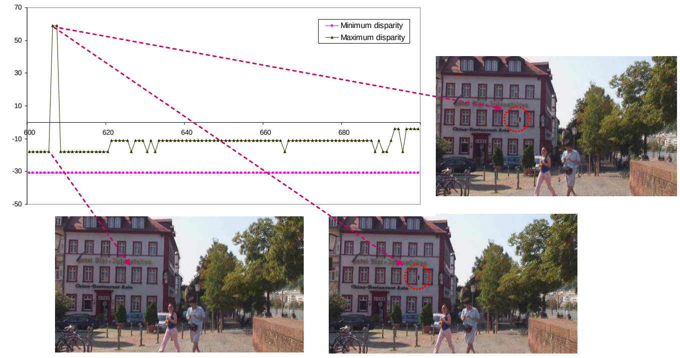


Fig. 1. Problem in disparity search range estimation without enforcing temporal consistency.

puted by thresholding the disparity histogram, which represents the distribution of depth information in a scene.

To generate the disparity histogram, an initial set of matching points can be obtained by a sparse feature matching method such as KLT (Kanade-Lucas-Tomasi) [7, 8] and SURF (Speeded Up Robust Features) [9] trackers. These methods define a descriptor for interest points and track the points using gradient or nearest-neighborhood methods. Alternatively, any dense matching method, e.g., based on graph cuts [10] or belief propagation [11], can be used to build the histogram. To reduce computation, the disparity map is typically computed on a sub-sampled version of the original input images [5].

In this work, SURF was used to generate the disparity histogram. SURF is a scale and rotation invariant interest point detector and descriptor. It relies on integral images for image convolution and builds on the strengths of the existing detectors and descriptors with a Hessian matrix-based measure and a distribution-based descriptor [9].

Using the pairs of matched feature points, the disparity histogram is computed as follows:

$$h[i] = \sum_{j=1}^{N} f\left(D(i), \left\lfloor \frac{d_j}{B} \pm 0.5 \right\rfloor \cdot B\right) \qquad i = 1, 2, \cdots M \tag{1}$$

where $h[i]$ is the histogram count for the i^{th} bin (M: the total num. of bins) and $f(a, b)$ is set to 1 if $a = b$, 0 otherwise. By quantizing each disparity value d_j of a matching-points pair (out of the N total pairs) with the bin-size B, a histogram bin count with the closest representative value $D(\cdot)$ will be incremented by one. Fig. 2 shows sample histograms computed using the SURF matching points.

The disparity search range is then computed by thresholding the histogram and removing outliers. Generally speaking, points with positive disparity are more important than those with negative disparity since the human visual system (HVS) is more sensitive to near objects. Based on this assumption, the threshold we use is defined as follows:

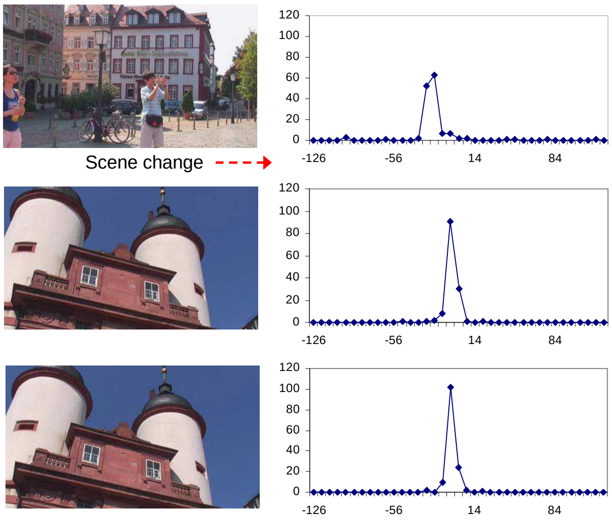


Fig. 2. Histogram similarity when scene changes.

$$T_h = \begin{cases} 2B & if \;\; d < 0 \\ [B/2] + 1 & otherwise \end{cases} \tag{2}$$

The disparity search range R is then determined by:

$$R = \{k | D(i) - B/2 \leq k \leq D(i) + B/2, \quad h[i] > T_h\} \tag{3}$$

Fig. 1 plots the minimum and maximum disparity for a sequence of stereo images using this scheme. The corresponding synthesized views based on an estimated disparity map at select instances of time are also shown. We can see that the estimated search range is unstable even though the consecutive frames have very similar depth distributions. The visual artifacts in the synthesized views are also apparent.

3. ENFORCING TEMPORAL CONSISTENCY

If the search range R has unnecessary disparities or misses significant ones, the stereo matching process can get trapped in local minima more easily. The threshold-based range detection method described in the previous section tends to be sensitive to false matching due to noise, color mismatches, repetitive patterns, etc. Even a few false matches can yield different disparity ranges in the temporally-neighboring frames with similar depth distributions, since consistency among consecutive frames is not considered.

In order to address this problem, one might attempt to build a temporally more consistent feature matching algorithm, but it is outside the scope of this paper. Instead, we consider enhancing the temporal consistency among the disparity histograms of consecutive frames so that we can leverage any decent feature-matching method in a temporally consistent manner. The new histogram is obtained by calculating weighted sums of temporally-neighboring histograms using the weights determined by the similarity of depth distributions between frames. The depth distribution of a scene depends on scene or camera configuration change, and it can be represented by the disparity histogram. Therefore, the similarity of disparity histograms can be used to identify scene or camera configuration change as well as to reduce the effects of outliers. The similarity of disparity histogram can be computed with the weighting factor w as follows:

$$w_{n,p} = \exp\left(-\sum_{i}^{M} \left|h_n^{nor}(i) - h_p^{nor}(i)\right| / \sigma_S\right) \qquad (4)$$

where $w_{n,p}$ represents the similarity measure between the n^{th} and p^{th} histograms. σ_S is the weighting constant for distance between histograms. Note the normalized histogram h_n^{nor} should be used in Eq. (4), since the total numbers of matching points vary among consecutive frames. Since the distance is calculated with the normalized histograms, it ranges from 0 to 2. Fig. 2 shows the histogram and its similarity measure when a scene change occurs. The similarity value between the 1^{st} and the 2^{nd} histogram was 0.149, while it was 1.719 between the 2^{nd} and the 3^{rd} ones. We empirically determined σ_s to be 0.4. The new histogram can be computed by calculating the weighted sum of temporally-neighboring histograms as follows:

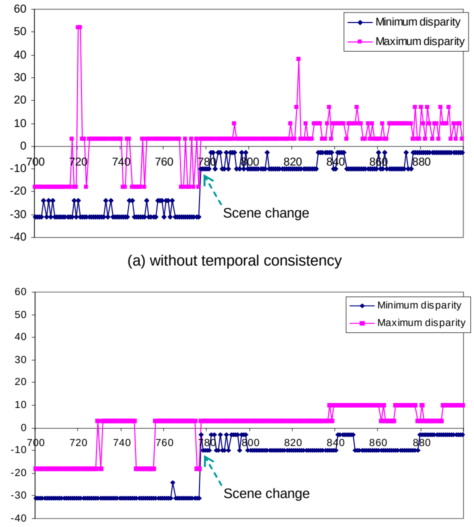
$$h_p^w(i) = \sum_{n \in N(p)} w_{n,p} h_n(i) \qquad (5)$$

where h_p^w is the p^{th} weighted histogram, and $N(p)$ represents the set of neighboring frames for the p^{th} frame. In this work, only the temporally causal neighboring frames were included in $N(p)$. Also each weighted histogram h_p^w is used only to estimate the disparity search range, but is not used to replace the original histogram since it may cause error propagation.

4. EXPERIMENTAL RESULTS

To validate the performance of the proposed method, we performed experiments with the stereo video sequences 'Heidelberg' and 'RhineValley', available at [12], whose sizes are 1280 × 720 and 720 × 576, respectively. For both sequences, the bin size B was set to 7 and the number of temporally-neighboring frames was 12. The input images were sub-sampled by a factor of 2 prior to applying the sparse feature matching by SURF; this tended to improve the performance.

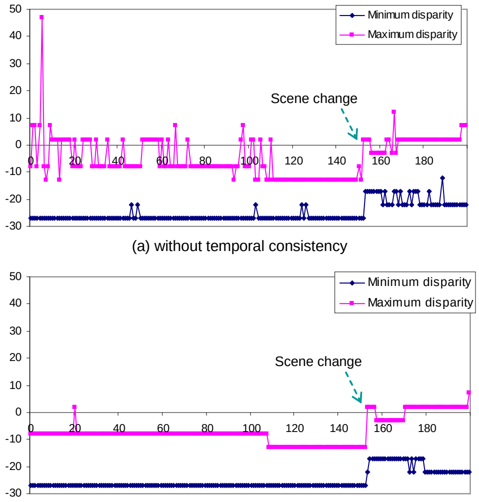
Fig. 3 and 4 show the estimated disparity search ranges for 'Heidelberg' and 'Rhine Valley', respectively. Both stereo sequences include two segments with a scene change between them. Our results confirm that the proposed method provides more reliable disparity search ranges and is robust to the scene change. Next, we demonstrate the effect of a more accurate search range on the estimated disparity maps and synthesis results. Figs. 5 and 6 show sample disparity maps obtained with and without temporal consistency in the search range estimation for both sequences. The dense disparity maps were computed based on the estimated search ranges and a hierarchical cost aggregation method [13]. The corresponding view synthesis results are also shown, where a novel view is synthesized at 30% of the original baseline distance. As shown by these results, incorrect disparity search ranges without temporal consistency often lead to incorrect disparity maps and visual artifacts in the synthesized views. The temporal consistency ensures a more reliable search range, which improves the estimated disparity maps and view synthesis.



(a) without temporal consistency

(b) with temporal consistency

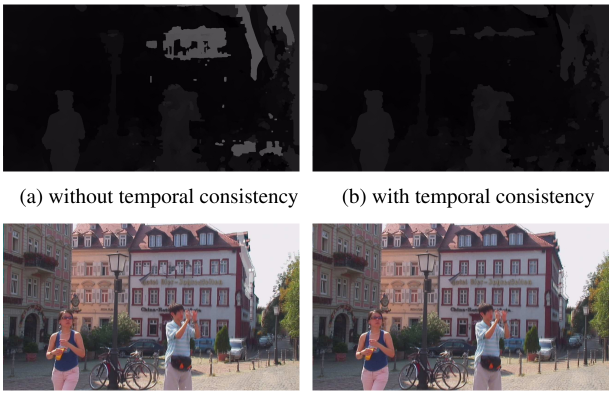
Fig. 3. Range estimation results for 'Heidelberg' video



(a) without temporal consistency

(b) with temporal consistency

Fig. 4. Range estimation results for 'Rhine Valley' video



(a) without temporal consistency (b) with temporal consistency

(c) synthesized frame using (a) (d) synthesized frame using (b)

Fig. 5. Disparity maps by search ranges without/with temporal consistency, 720^{th} frame, Heidelberg

5. CONCLUSION

In this paper, we proposed a temporally-consistent method for estimating the disparity search range for a stereo video. The weighted sum of temporally-neighboring histograms is formed taking the histogram similarity into account in order to make the disparity search range estimation more consistent over time and robust to scene or camera configuration changes. The experimental results show that the proposed method works well for challenging stereo video sequences. In the future, Kullback-Leibler divergence [14, 15], which represents the similarity of two probability distributions, might be used to compute the histogram similarity. Since the histogram is normalized to calculate the weighting factor w, it could be also regarded as a propability distribution function. Furthermore, an extension of the proposed method to multiview video might be considered. In contrast to stereo video, multiple histograms exist in the multiview video case and it would be necessary to estimate a consistent search range from the multiple histograms.

6. REFERENCES

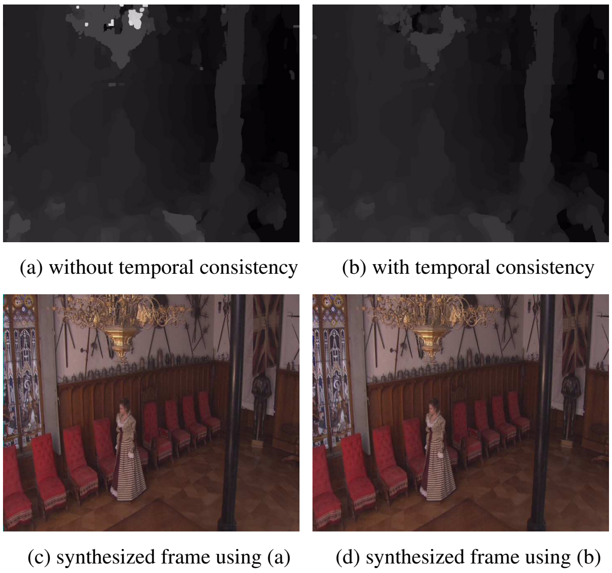
[1] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *IJCV*, vol. 47, no. 1-3, pp. 7-42, Apr. 2002.

[2] http://vision.middlebury.edu/stereo

[3] B. Cyganek and J. Borgosz, “An improved variogram analysis of the maximum expected disparity in stereo images,” *Proc. SCIA, LNCS*, pp. 640-645, 2003.

[4] R. Sara, “Finding the largest unambiguous component of stereo matching,” *Proc. ECCV*, pp. 900-914, 2002.

[5] J. Kostkova and R. Sara, “Automatic Disparity Search Range Estimation for Stereo Pairs of Unknown Scenes,” *Proc. CVWW*, 2004.



(a) without temporal consistency (b) with temporal consistency

(c) synthesized frame using (a) (d) synthesized frame using (b)

Fig. 6. Disparity maps by search ranges without/with temporal consistency, 5^{th} frame, RhineValley

[6] M. Gong, “Enforcing Temporal Consistency in Real-Time Stereo Estimation,” *Proc. ECCV*, p. 564-577, 2006.

[7] B. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” *International Joint Conference on Artificial Intelligence*, pp. 674-679, 1981.

[8] C. Tomasi and T. Kanade, “Detection and Tracking of Point Features,” *Carnegie Mellon University Technical Report CMU-CS-91-132*, April 1991.

[9] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, “SURF: Speeded Up Robust Features,” *Computer Vision and Image Understanding* (CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

[10] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” *Proc. IEEE Int. Conf. Computer Vision*, pp. 508-515, 2001.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient belief propagation for early vision,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 261-268, 2004.

[12] http://www.3dtv.at/

[13] D. Min and K. Sohn, “Cost aggregation and occlusion handling with WLS in stereo matching,” *IEEE Trans. Image Processing*, vol. 17, no. 8, pp. 1431-1442, Aug. 2008.

[14] S. Kullback, R.A. Leibler, “On Information and Sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79-86, 1951.

[15] S. Kullback, “The Kullback-Leibler distance,” *The American Statistician*, vol. 41, pp. 340-341, 1987.