LOCAL STEREO MATCHING USING MOTION CUE AND MODIFIED CENSUS IN VIDEO DISPARITY ESTIMATION

Zucheul Lee, Ramsin Khoshabeh, Jason Juang and Truong Q. Nguyen
UCSD, La Jolla, CA92093-0407
z1lee, ramsin, jajuang, tqn001@ucsd.edu

ABSTRACT

In the human visual system, proximity, similarity, and motion are fundamental attributes that group visual objects together locally. The objects grouped by these attributes are most likely to have the same depth. In previous works, proximity and similarity have been considered in the computation of image disparity maps. However, they are insufficient for video disparity estimation because motion cues are very important for accurate depth estimation near edges of moving objects. We incorporate motion flow to compute each pixel's support weight, a measure directly affecting the accuracy of disparity maps in local methods. For robustness to image noise in flat areas, we propose a modified census transform with a noise buffer. The experimental results show that the proposed method produces more accurate disparity maps than current state-of-the-art methods, both on edges and in flat areas according to subjective and objective measures.

Index Terms— stereo matching, disparity, motion flow, census transform

1. INTRODUCTION

Stereo depth information is a basic element of the 3D interpretation of a scene and disparity estimation is an important step in resolving that depth. The area of disparity estimation has been thoroughly studied over the past decade, and almost all research results have been focused strictly on images. The numerous advanced algorithms for image disparity estimation may be generally categorized as either local or global methods. The former are known to be fast and the latter tend to be more accurate.

While image disparity estimation is mature, video disparity estimation, on the other hand, is at an early stage. This is the consequence of two main factors: (1) lack of video datasets with ground-truth disparity maps and (2) temporal inconsistency problems, such as flickering resulting from simply applying current state-of-the-art image-based algorithms to video. To reduce this artifact, [1] uses median filtering along flow vectors computed by the method of Horn and Schunck [2]. However, the results are of moderate quality. Reference [3] shows impressive results by treating the video disparity as a spatio-temporal volume to improve spatial and temporal consistency and it presents the possibility for directly extending current image-based disparity algorithms to the video domain.

Local methods present themselves as more appropriate solutions for video disparity estimation because it often requires real-time processing capabilities. In most local methods, window-based matching is used to find corresponding pixels in a pair of left and right images. However, this results in the foreground smearing problem near depth discontinuities due to the assumption that all pixels in the window have the same disparity. To solve this problem, the adaptive-window method [4] finds an optimal window based on the local variation of intensity and disparity. This method uses a rectangular window, which is not suitable for arbitrarily shaped depth discontinuities. The multiple-window method [5] calculates the correlation with nine pre-defined windows and selects the disparity with the smallest matching cost. This method also has the limitation of window shape. To obtain more accurate results at depth discontinuities, the locally adaptive support weight approach (LASW) [6] adjusts the support weights of the pixels in the window by using the photometric and geometric distance with respect to the center pixel. This method deals with the pixels near depth discontinuities more effectively than the two methods mentioned above. Cost-filter [7] shows the best edge-preserving results by using the guided filter and it is a local method that outperforms all other local methods on the Middlebury benchmark. However, both LASW and Cost-filter do not provide a reliable solution for disparity estimation in textureless (flat) areas, which have different characteristics from edges.

In this paper, we propose a more accurate and noise tolerant stereo matching approach for video disparity estimation. Motion is a crucial factor in video processing and generally moving objects tend to have a higher degree of saliency. Every disparity algorithm tends to have difficulty dealing with moving edges and textureless areas in video scenes. We provide an advanced local method by using motion cues and a modified census transform with a noise buffer to obtain more accurate disparity information in the edges of moving objects and to be robust to image noise in the textureless areas, respectively. In addition, we enforce temporal consistency by refining our disparity estimates with the spatio-temporal consistency method described in [3].

This paper is organized as follows. The details of our proposed method are presented in Section 2. Section 3 shows experimental results and discusses their significance. Section 4 concludes with some remarks.

2. PROPOSED METHOD

2.1. Gestalt Grouping

According to gestalt principles, human observers are able to group visual objects that share certain common characteristics [8]. The best-known grouping laws are proximity (objects that are close to each other are grouped together), similarity (objects that have similar color are grouped together), and common fate (objects that move at the same speed in the same direction are grouped together) [9]. Common fate is closely related to motion flow, which will be denoted as "motion" for simplicity. Whenever objects have characteristics in common, they get grouped and form a new, larger visual object, known as a gestalt [8].

From these observations, we can assume that human observers group pixels in a scene based on how close two pixels are spatially, how similar their colors are, and how similar their velocities are. Thus, we can use the strength of grouping when computing the support weight of a pixel, which should be proportional to the probability that the two pixels have the same disparity. The closer two pixels are in proximity and color, the larger their support weight. The same can be said about the motion flows of two pixels. These three observations may be treated in an integrated manner to obtain a reasonable grouping [6]. Each grouping law can compensate for the others when they fail in specific cases. For instance, the motion cue helps viewers distinguish figures and group them when the object color or outlines are not clear. Therefore, we can model the human visual system and segment objects by using support weights based on gestalt principles.

2.2. Benefits of a motion cue

Consider the example of Fig. 1 for illustration of the benefits of using motion cues. We use the LASW method, in which proximity and similarity are exploited, and extend it to evaluate how the three weighting terms (proximity, similarity, and motion) affect the quality of the disparity maps. As the local methods require pixel-based computation, we use classic optical flow with the weighted non-local term [10], which is one of state-of-the-art optical flow methods. We exploit the motion to compute the integrated support weight similarly as in [6]. The "car" video frames are processed at a resolution of 480 x 270 with a disparity range of 15 and the parameters used are fixed throughout the experiment. In Fig. 1, we show the selected left view and its optical flow. Fig. 1(c) is obtained by using only the proximity term for the support weight, Fig. 1(d) is obtained by adding the similarity term, and Fig 1(e) is obtained by adding the motion term. In Fig. 1(c), we observe many errors in the edges of the moving car (red circle in Fig. 1). In Fig. 1(d), some errors are recovered by using the color cue but edges are not preserved. In Fig. 1(e), incorporating the motion term preserves the edges even though they are visually ambiguous. We believe that this is due to the preserved background flow as shown in Fig. 1(b). Although there is ambiguity in the stereo correspondence, motion between a pair of successive video frames is much more consistent, especially in a localized window in background regions. Additionally, disparity is estimating spatial correspondences while motion estimates temporal correspondences, so the additional information promotes disambiguation. Consequently, the results in Fig. 1 imply that the support weight integrating the motion cue yields more accurate disparity estimates, especially near the edges of moving objects.

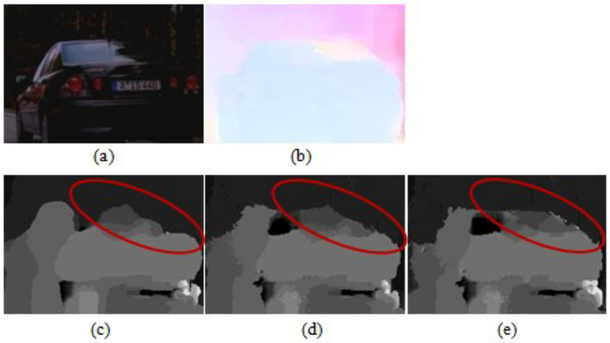


Fig. 1. Disparity maps for "Car"; (a) Left view; (b) Optical flow; (c) using only proximity; (d) using proximity and similarity; (e) using proximity, similarity, and motion

2.3. Support weight using correlated color and motion

Based on the main gestalt principles, the support weight using similarity and motion can be expressed as

$$w(c,q) = f(\Delta s_{cq}, \Delta m_{cq}) \tag{1}$$

where the function *f* represents the strength of grouping and Δs_{cq} and Δm_{cq} represent the color difference (a measure of similarity) and motion difference between the center pixel *c* and the neighbor pixel *q*, respectively. The color difference is computed by the Euclidean distance between pixel values in the CIELab color space, which gives a three dimensional representation for color perception. Let $s_c = (L_c, a_c, b_c)$ and $s_q = (L_q, a_q, b_q)$ be the color coordinates of pixel *c* and pixel *q* in the CIELab color space, respectively, as shown in Fig. 2. Then, Δs_{cq} is calculated by

$$\Delta s_{cq} = \sqrt{(L_c - L_q)^2 + (a_c - a_q)^2 + (b_c - b_q)^2}. \tag{2}$$

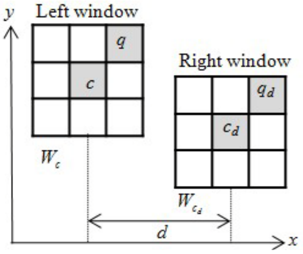


Fig. 2. Left support window and right support window

The second term is the motion difference, which is a measure of motion flow. There are two types of motion difference computation: absolute flow endpoint difference (ED) and angular difference (AD) [11]. We use ED because AD penalizes errors in larger flows less than errors in small ones [11], which is undesirable. Let $m_c = (u_c, v_c)$ and $m_q = (u_q, v_q)$ be the flow vectors of pixel c and pixel q, respectively. We use a truncated motion difference:

$$\Delta m_{cq} = \min\left[\sqrt{(u_c - u_q)^2 + (v_c - v_q)^2}, \tau\right] \tag{3}$$

where τ is a truncation value. Such a model reduces the influence of flow outliers just as the truncated matching cost limits the influence of wrong matches [12]. We must keep in mind that the optical flow is an estimated value and cannot be completely error free.

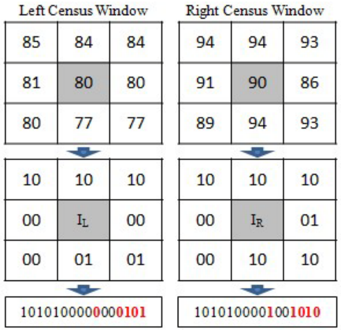
The strength of grouping by similarity is defined using Laplacian kernel as

$$f_p(\Delta s_{cq}) = \exp(-\frac{\Delta s_{cq}}{\gamma_s}) \tag{4}$$

with γ_s being an empirical similarity parameter. The strength of grouping by motion can be defined in the same manner. We suggest a correlated model for the integrated support weight as

$$w(c,q) = \exp(-\frac{\Delta m_{cq}}{\gamma_m})^{\frac{\Delta s_{cq}}{\gamma_s}} \exp(-\frac{\Delta s_{cq}}{\gamma_s}). \tag{5}$$

This model originates from the intuition that the gestalt principles tend to correlate with each other in general. For example, the center pixel and its neighboring pixel have a high likelihood of having different motion vectors if they also differ significantly in color, as expected near object edges. When this occurs, the correlated model decreases the overall support weight as compared with the independent model, as in [6], since the Laplacian is raised to a power based on the large color difference. Additionally, the two pixels are likely to have similar motion if they also have the same color, as in the flat areas of an object surface. In this case, we can also expect to find a positive correlation among the two metrics. Therefore, the support weight will increase in reference to the independent model. However, while color is an observed quantity, motion is an estimated value. Therefore, color should take precedence over motion when there is a discrepancy between the two of them and the correlation assumption fails. This is precisely what the model in (5) enforces. For example, if there is a large difference in color but a small one in motion, then the value for the correlated support weight is decreased. Therefore, the support weight depends more on the color cue than the motion cue. In contrast, the independent model always treats all of the gestalt principles equally. We verify through simulation that the correlated model generally improves the overall performance of video disparity estimation.



85	84	84	94	94	93
81	80	80	91	90	86
80	77	77	89	94	93

10	10	10	10	10	10
00	I_L	00	00	I_R	01
00	01	01	00	10	10

1010100000000101   1010100001001010

Fig. 3. Example of modified census transform with α=1

2.4. Modified census transform

The census transform is robust to radiometric distortions. In addition, from the evaluation results of [13], the census transform applied to raw matching cost computation shows the best overall performance in both local and global methods. However, it experiences difficulties in finding the correct correspondences in flat areas, as most methods do. This difficulty is due to the fact that the census matching cost is extremely sensitive to image noise since all pixels in flat areas have a similar intensity. To solve this problem, we proposed a three moded census transform with a noise buffer. The original census has two modes where a bit is set to 1 if the neighboring pixel in the census window has a higher intensity than the center pixel and 0 otherwise. On the other hand, our modified census uses two bits to implement three modes, where the two bits are set to 10 if the neighboring pixel has an intensity value higher than the center pixel by noise buffer threshold (α), 01 if the neighboring pixel's intensity is lower than the center pixel by α, and 00 otherwise. Recognizing that noise levels are not linearly related to image intensity values, we set a different noise buffer value for each intensity band. For instance, α is set to 0 if the intensity value is between 0 to 50 and it is set to 1 if the intensity value is between 50 and 100 (2, 3 and 4 for 100~150, 150~200 and 200~255, respectively). The Hamming distance is then calculated by the bitwise XOR operation upon the left and right census transformed bit strings. To further improve the matching accuracy, we incorporate the intensity difference (|I_L-I_R|) between two center pixels as shown in Fig. 3. In other words, we use the census transform to compare the spatial structure of two census windows, while we use the intensity difference to compare two center pixel values. For the integrated raw matching cost, we consider two distances (Hamming distance and intensity difference) in the same way as with color and motion discussed above. At a true correspondence when both the spatial structure and the center pixel intensity of the left window match those of the right window, the matching cost has the lowest value. The raw matching cost is expressed as

$$C_0(q,q_d) = 1 - \exp(-\frac{\Delta I_{qq_d}}{\gamma_I})^{\frac{\Delta H_{qq_d}}{\gamma_H}} \exp(-\frac{\Delta H_{qq_d}}{\gamma_H}) \qquad (6)$$

where ΔI_{qq_d} and ΔH_{qq_d} are intensity difference and Hamming distance, respectively, between pixel q and pixel q_d as shown in Fig. 2. γ_I and γ_H are empirical parameters. The support weight and the raw matching cost are inversely proportional but possess a similar formulation. We use the example of Fig. 4 to evaluate how the modified census improves the disparity map. In Fig. 4, we show the left view and three disparity maps; Fig. 4(b) is computed by the original census, Fig. 4(c) is computed by the modified census without incorporating intensity, and Fig. 4(d) is computed by the full modified census. In Fig. 4(a), there is a flat area highlighted with the red box. The original census exhibits some errors in that area as shown in Fig. 4(b) but the modified census without intensity recovers them. Finally, the modified census, incorporating intensity, shows the best quality of disparity map as shown in Fig. 4(d).

2.5. Aggregation and disparity computation

Once the support weights are calculated, the aggregated matching cost between pixels is computed by aggregating the raw costs, scaled by the support weights in the window. If we consider only the left support window, the cost computation may be erroneous since the right support window may have pixels from different depth levels. To minimize such errors, the matching cost is computed by combining the support weights of both support windows as in [6]. The aggregated matching cost between pixel c and pixel c_d in Fig. 2 is given in the weighted mean form:

$$A(c,c_d) = \frac{\sum_{q \in W_c, q_d \in W_{c_d}} w(c,q) \cdot w(c_d,q_d) \cdot C_0(q,q_d)}{\sum_{q \in W_c, q_d \in W_{c_d}} w(c,q) \cdot w(c_d,q_d)} \qquad (7)$$

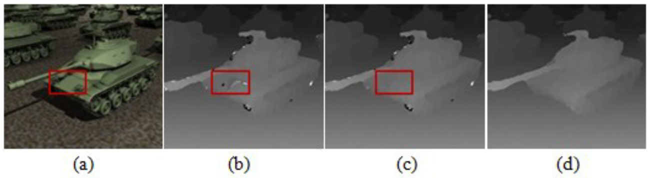


Fig. 4. Disparity map for "Tanks"; (a) Left view; (b) Original census; (c) Modified cenus without incorporating intensity; (d) Modified census.

where W_c and W_{c_d} represent the left and right support windows, respectively, and the function $w(c_d,q_d)$ is the support weight of pixel q_d in the right window.

After the aggregated matching costs have been computed within the disparity range, the disparity map is obtained by determining the disparity d_p of each pixel p through the Winner-Takes-All (WTA) algorithm:

$$d_p = \arg\min_{d \in D} A(c,c_d) \qquad (8)$$

where D represents the set of all possible disparities.

3. EXPERIMENTS AND RESULTS

To assess the performance of our proposed method quantitatively, we use 5 synthetic stereo videos (400 x 300, 64 disparity range) with ground truth disparity [14]. We compare three methods (LASW, Cost-filter, and our method) without post-processing to compare their pure performance. The LASW method ranks 53rd and the Cost-filter, the best performing local method ranks 16th on the Middlebury benchmark. Both methods do not perform any iterative process, just as in our method. The parameters are set to constant values: $\gamma_s = 17$, $\gamma_m = 1$, $\gamma_I = 3$, $\gamma_H = 20$, and $\tau = 1$. The size of the support and census windows are set to 11 x 11 and 7 x 7, respectively. Table 1 shows the average percentage of bad pixels (threshold of 1) over all frames. We ignore borders when computing statistics since they lack correspondences. Table 1 illustrates that the proposed method, using the motion cue and modified census has the best performance on all datasets except for "street."

To assess the performance of the proposed method subjectively, we perform experiments on a real-world video, "Jamie1," a scene from the Microsoft i2i database (320 x 240, 64 disparity range). These video frames contain large flat areas and repetitive patterns, as shown in Fig. 5. Fig. 5(b) shows the disparity maps produced by LASW, Fig. 5(c) shows the disparity maps produced by Cost-filter, and Fig. 5(d) shows the disparity maps produced by our proposed method. Fig. 5 illustrates that the proposed method

Video/ # of Frames	LASW	Cost-filter	Our method
Tunnel/99	1.382 %	2.157 %	**1.032 %**
Temple/99	12.530 %	10.700 %	**10.164 %**
Book/40	6.102 %	4.919 %	**4.758 %**
Street/99	9.907 %	**7.305 %**	7.619 %
Tanks/99	5.591 %	4.826 %	**4.803 %**

Table 1. Performance comparison of methods

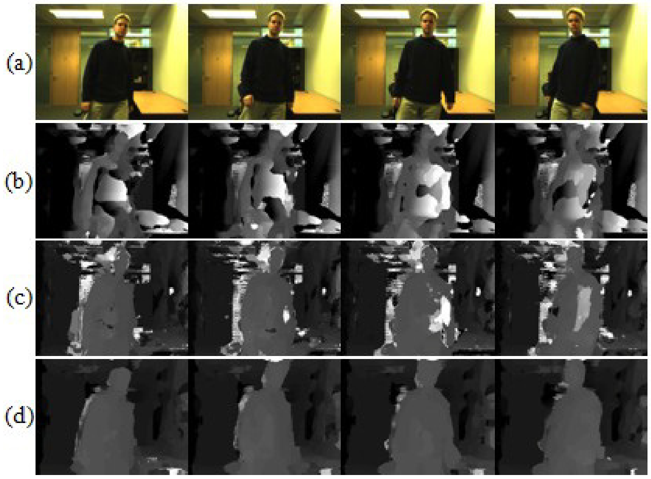


Fig. 5. Disparity map for "Jamie1"; (a) Left frames; (b) LASW; (c) Cost-filter; (d) Our method.

exhibits the best quality of disparity map. On the other hand, LASW yields the worst quality. Cost-filter produces many errors in flat and repetitive areas.

In our method, it takes about 19s to compute the disparity map for a stereo pair with 400 x 300 resolution, as used in Table 1. Our method has a similar framework to [6] and it has been shown in [15] that [6] can be adopted into a real-time application by using a Graphics Processing Unit (GPU). Thus, the same could be done with our work.

Although not tabulated due to limited space, refinement using the TV method [3] reduces errors such as spatial noise and temporal inconsistencies in the background significantly. For more results, please refer to our website: http://videoprocessing.ucsd.edu/~zucheul/laswm.

4. CONCLUSION

An accurate local stereo matching method using motion cue and modified census transform for video disparity estimation is proposed in this paper. In the local window methods, the accuracy of the disparity map depends on the support weight and the raw matching cost. To compute more accurate support weights, we consider object motion and suggest a correlated model. To obtain more reliable raw matching costs in flat areas, modified census with a noise buffer incorporating intensity is used. Simulation results verify that the proposed method outperforms previous works.

REFERENCES

[1] M. Bleyer and M. Gelautz, "Temporally Consistent Disaprity Maps from Uncalibrated Stereo Videos," ISPA, pp. 383-387, 2009.

[2] B. Horn and B. Schunck, "Determining Optical Flow," Artificial Intelligence, 17, pp. 185-203, 1981.

[3] R. Khoshabeh, S. H. Chan, and T. Q. Nguyen, "Spatio-Temporal Consistency in Video Disparity Estimation," ICASSP, pp. 885-888, 2011.

[4] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiments," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 9, pp. 920-932, 1994.

[5] A. Fusiello, V. Roberto and E. Trucco, "Efficient Stereo with Multiple Windowing," CVPR, pp. 858-863, 1997.

[6] K. J. Yoon and I. S. Kwon, "Locally Adaptive Support-Weight Approach for Visual Correspondence Search," CVPR, vol. 2, pp. 924-931, 2005.

[7] C. Rhemann, A. Bleyer, C. Rother, M. Gelautz, "Fast Cost-volume Filtering for Visual Correspondence and Beyond," CVPR, pp. 3017-3024, 2011

[8] D. Angens, From Gestalt theory to image analysis : a Probabilistic approach, NY : Springer, 2008.

[9] G. Papari and N. Petkov, "Adaptive Pseudo Dilation for Gestalt Edge Grouping and Contour Detection," IEEE trans. Image Processing, vol. 17, pp. 1950-1962, 2008.

[10] D. Sun, S. Roth, M.J. Black, "Secrets of Optical Flow Estimation and Their Principles," CVPR, pp. 2432-2439, 2010.

[11] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, and R. Szeliski, "A database and Evaluation Methodology for Optical Flow," ICCV, pp. 1-8, 2007.

[12] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann, "Local Stereo Matching using Geodesic Support Weights," ICIP, pp. 2093-2096, 2009.

[13] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. IEEE TPAMI, 31(9):1582–1599, 2009.

[14] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time Spatiotemporal Stereo Matching Using the Dual-Cross-Bilateral Grid," ECCV, 2010.

[15] M. Gong, R. Yang, L. Wang, and M. Gong, "A Performace Study on Different Cost Aggregation Approaches used in Real-time Stereo Matching," IJCV, vol. 75, no. 2, pp. 283-296, 2007.