SPATIO-TEMPORAL CONSISTENCY IN VIDEO DISPARITY ESTIMATION

Ramsin Khoshabeh * *, Stanley H. Chan, Truong Q. Nguyen*

University of California, San Diego
Department of Electrical and Computer Engineering

ABSTRACT

We present a novel stereo *video* disparity estimation method. The proposed method is a two-stage algorithm. During the first stage, initial disparity maps are computed in a frame-by-frame basis. In the second stage, the initial estimates are treated as a space-time volume. By setting up an l_1-normed minimization problem with a novel three-dimensional total variation regularization, spatial smoothness and temporal consistency are handled simultaneously. Due to our unique formulation, any existing image disparity estimation technique may utilize our method as a post-processing step to refine noisy estimates or to be extended to videos. The proposed method shows superior speed, accuracy, and consistency compared to state-of-the-art algorithms.

***Index Terms*—** stereo vision, video disparity, signal denoising, augmented Lagrangian, total variation minimization

1. INTRODUCTION

Stereo disparity estimation is an integral problem associated with 3D content delivery. Disparity is an important element for an accurate 3D visual representation. In a two-camera imaging system, disparity is defined as the vector difference between the imaged object point in each image relative to the focal point [1]. It is this disparity that allows for depth estimation of objects in the scene via triangulation of the point in each image. In rectified stereo, where both camera images are in the same plane, only horizontal disparity exists. In this case, multiview geometry shows that disparity is inversely proportional to actual depth in the scene.

The problem of estimating disparity has been well-studied for images. Numerous methods have presented impressive results. However, simply applying even the best of these methods to individual frames of stereo sequences yields temporally inconsistent disparity maps. This is perceived as a high-frequency flickering effect when the sequence is visualized.

Our goal in this paper is to present a systematic method by which we generate accurate and spatio-temporally consistent disparity maps from complex *stereo video sequences*. We leverage the strengths of current state-of-the-art image-based techniques, but, in addition, we explicitly enforce the consistency of estimates in both space and time by treating the video as a space-time volume corrupted by noise. In so doing, we provide an algorithm that has the capability of refining arbitrary image-based disparity estimation techniques and, at the same time, extending them to the video domain.

In Section 2 we briefly discuss existing image- and video-based disparity estimation techniques. Section 3 presents the proposed algorithm. We evaluate the method and compare results with one of today's most advanced techniques in Section 4. In Section 5 we discuss the implications of using our algorithm and finally conclude with some closing remarks.

2. RELATED WORK

For *static images*, the problem of disparity estimation has been thoroughly studied, with standardized databases, such as Middlebury [2], aiding in the fast evolution of the myriad techniques. The existing algorithms may be categorized into one of two groups: local or global methods. Local methods treat each pixel (or an aggregated region of pixels) in the reference image independently and try to infer the optimal horizontal displacement to match it with the corresponding pixel/region in the other image. In contrast, global methods incorporate assumptions about depth discontinuities and estimate disparity values by minimizing an energy function over all pixels using techniques like Graph Cuts [3][4] or Hierarchical Belief Propagation (HBP) [5]. Generally local methods tend to be fast but lack the accuracy of global methods. However, straightforward implementations of most global methods tend to be extremely slow. A thorough review of stereo matching techniques can be found in [6].

For *video sequences*, on the other hand, solutions to the stereo matching problem are few and far between. Largely due to the computational bottleneck of dealing with multi-dimensional data, lack of any real datasets with ground-truth, and the unclear relationship between optimal spatial and temporal processing for correspondence matching, few have ventured to present viable solutions to the video disparity estimation problem. The ones that have tried typically do so by extending existing methods for images to videos, with the debilitating drawback of computational times that make the methods impractical for most applications.

In an attempt to build off the successful HBP approach, [7] extends the matching cost representation to video by forming a 3-dimensional Markov Random Field (MRF). The ap-

email: ramsin@ucsd.edu
website: http://videoprocessing.ucsd.edu

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

proach intuitively makes sense but, as we mentioned earlier, computational times make it unusable in most cases. They report algorithmic run times as high as 947.5 seconds for a single 320×240 frame on a powerful computer.

Scene Flow [8] and related work [9] try to utilize motion flow fields to enforce temporal coherence. [8] defines a 3D motion vector field whereas [9] uses median filtering along vectors from traditional Optical Flow. However, the process of flow field estimation itself introduces unnecessary errors into the framework and, for high accuracy, requires significant computational time as well. The median filtering technique requires the flow fields to be computed as a pre-processing step, and so it suffers from the same drawbacks.

Perhaps the most promising technique that we have come across is that of [10], which shows practical, real-time usability via a GPU implementation of HBP using an approximation to locally adaptive support weights [11]. Their video method, TDCB, integrates temporal coherence in a similar way to [7], reporting computational times of 90 ms per frame (11.1 fps) on a 400×300 image with 64 levels of disparity. They also provide a synthetic dataset with ground-truth disparity maps. We believe such datasets are a crucial fundamental step in advancing the maturation of video disparity estimation. Even with their promising findings, we show that our algorithm is capable of further refining their results in Section 4.

Additional methods have also been proposed, but they normally require specific hardware, such as time-of-flight sensors [12], or constraints on the data, such as static scenes [13], that go beyond the scope of this paper.

3. PROPOSED METHOD

We avoid formulating the video disparity problem in space-time, with the realization that such attempts become computationally impractical. Furthermore, image-based techniques have been thoroughly studied and are much more advanced in their implementations. We wish to leverage the breakthroughs made in that research area. Yet the difficulty with operating on each frame of a video sequence independently is that we lose the consistency between consecutive frames. The noisy estimates for each frame create a flickering effect over time that is highly bothersome to the human visual system.

To improve the temporal consistency, we present a novel, fast, and efficient method based on an augmented Lagrangian method for total variation (TV) image restoration presented in [14]. Chan et al. show that the method is faster than state-of-the-art methods and yields superb results.

The proposed method consists of two steps. First, we compute disparity estimates for each frame individually using our image-based method, which is discussed in Section 3.1. Then between neighboring pixels in space-time, we enforce the disparity smoothness assumption that values should vary smoothly except at object boundaries. With this assumption, we can treat the temporal (and spatial) inconsistencies as a signal corrupted with noise. By solving a new three-dimensional TV minimization problem, the spatial and temporal consistency is improved. The TV minimization is detailed in Section 3.2.

3.1. Image-based Disparity Map Estimation

Our approach for static images is a global method using Hierarchical Belief Propagation (HBP) for inferencing. HBP maintains the accuracy of global methods, such as traditional Belief Propagation and Graph Cuts, but rivals local methods in computational time.

Let $\mathcal{P}$ be the set of pixels in an image and $\mathcal{L}$ be a finite set of labels. The labels correspond to quantities that we want to estimate at each pixel (i.e., the disparity). A labeling f assigns a label $f_p \epsilon \mathcal{L}$ to each pixel $p \epsilon \mathcal{P}$. As with traditional global methods, for each pixel we designate an energy function that indicates how well that label fits:

$$E(f) = \sum_{p \epsilon \mathcal{P}} D_p(f_p) + \sum_{(p,q) \epsilon \mathcal{N}} V(f_p - f_q) \tag{1}$$

$D_p(f_p)$ is referred to as the data cost and $V(f_p - f_q)$ is commonly called the smoothness cost in some literature, but it is more accurate to refer to it as a discontinuity cost. Intuitively, the data cost captures how well the labeling fits the node (how well the disparity estimate matches the stereo information). The discontinuity cost enforces the assumption that labels should vary slowly almost everywhere except for significant changes along object boundaries. Neighboring pixels in neighborhood $\mathcal{N}$ are penalized according to how large the difference is between their labels.

In our implementation, the data cost is computed over a large window for each pixel using Yoon and Kweon's locally adaptive support weights [11], so that only points with a high probability of belonging to the same object contribute significantly to the cost calculation. For the discontinuity cost, we use the commonly employed truncated weighted linear model, $V(f_p - f_q) = min(\alpha\,|f_p - f_q|\,, \beta)$, where f_p and f_q are the labels we wish to assign to pixels p and q.

We use the method of Felzenszwalb et al. [5] to minimize the energy over the entire image in a coarse-to-fine manner. It iteratively passes messages from all pixels to their neighbors in parallel. The message vector represents the minimal-energy labeling of each node (pixel) and all the information coming into it through the connected nodes. This current labeling, or belief, of each pixel is passed to its neighbors. The idea is that after T iterations, information will have propagated across the image and the minimization will lead to a globally optimal disparity labeling across the image.

3.2. Video Disparity Estimation - Temporal Consistency

In the previous step, disparity maps are computed for each frame individually. To enhance the temporal consistency of the disparity maps, we make the observation that disparity should generally be a piecewise smooth function in time, except for discontinuities at object borders (in which case the value may drastically change). This is because objects do not simply disappear from one frame to the next. However, this smoothness assumption is normally violated in most initial disparity maps, as there are inevitable estimation errors.

Realizing the fact that disparity maps have to be piecewise smooth, we propose to consider the sequence of disparity maps as a space-time volume, i.e., a three-dimensional function $f(x, y, t)$ with (x, y) being the spatial coordinates and t being the temporal coordinate. By applying a denoising algorithm to this space-time volume, we seek a piecewise smooth solution which, on one hand, has less temporal noise, and, on the other hand, preserves the disparity information as much as possible. To this end, we consider the following l_1-minimization problem:

$$\underset{\mathbf{f}}{\text{minimize}} \quad \mu \|\mathbf{f} - \mathbf{g}\|_1 + \|\mathbf{D}\mathbf{f}\|_2 \tag{2}$$

where $\mathbf{f}$ denotes the unknown disparity map (the vectorized version of $f(x, y, t)$), $\mathbf{g}$ is the initial disparity map from the previous step, and $\mathbf{D} = [\beta_x \mathbf{D}_x^T,\ \beta_y \mathbf{D}_y^T,\ \beta_t \mathbf{D}_t]^T$ represents the forward difference operators along the horizontal, vertical and temporal directions. The parameters $(\beta_x, \beta_y, \beta_t)$ control the relative emphasis being put on the spatial and temporal terms. An l_1-norm is chosen for the objective function, $\mathbf{f} - \mathbf{g}$, because the target solution $\mathbf{f}$ is ideally piecewise smooth. The regularization term $\|\mathbf{D}\mathbf{f}\|_2$ is the TV-norm on $\mathbf{f}$.

Problem (2) is solved by first introducing two intermediate variables $\mathbf{r} = \mathbf{f} - \mathbf{g}$ and $\mathbf{u} = \mathbf{D}\mathbf{f}$, and transforming the unconstrained problem into an equivalent constrained minimization problem. Then an augmented Lagrangian method is used to handle the constraints, and an alternating direction method (ADM) is used to solve the sub-problems iteratively. Due to limited space, we refer the reader to [14] for the details of this method.

There are three significant features of the proposed algorithm. First, disparity maps are now considered as a space-time volume, instead of individual frames. This allows us to handle *both* spatial and temporal consistency simultaneously, by tuning the parameters $(\beta_x, \beta_y, \beta_t)$. Second, as opposed to most of the existing methods that try to enhance temporal consistency by heuristic means, the proposed algorithm is guaranteed to find the global minimum of problem (2), because (2) is convex. Third, the proposed algorithm is fast. Typical run time for a 300×400 resolution sequence is approximately 2 seconds per frame on MATLAB, which implies the possibility for real-time processing with a compiled language.

4. ANALYSIS AND DISCUSSION

In Fig.1, we show six frames of a zoomed-in region of the left view from the ”Old Timers” sequence [15], along with per-frame disparity map estimates and refinements using our TV method. Notice that the image-based results contain both spatial noise and temporal inconsistencies, particularly in the background area on the right side. After refinement, these errors are removed while object edges are still preserved.

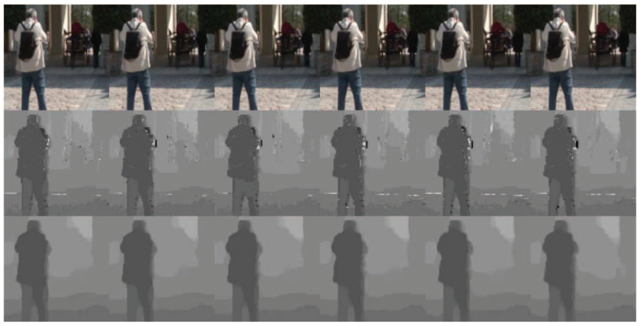

Fig. 1. Disparity Refinement for Old Timers. Top: Stereo Left View. Middle: Initial Disparity Estimate. Bottom: Processed Spatio-temporally Consistent Estimates.

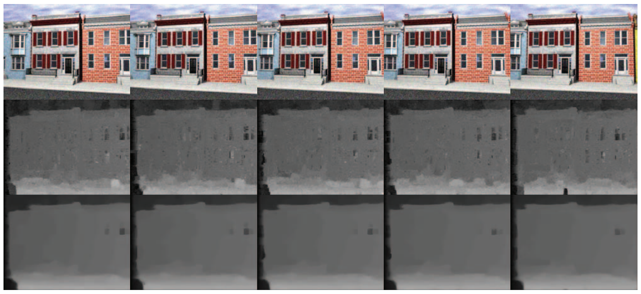

Fig. 2. Disparity Refinement for Synthetic Street Sequence. Top: Original. Middle: Disparity. Bottom: Processed.

While visually the results are clearly better after refinement, it is difficult to quantitatively assess the performance. Because of the lack of stereo sequences with ground truth disparity maps, effective evaluation of video disparity estimation techniques has been limited. We believe this may be a pivotal reason why so few methods currently exist to handle videos. Fortunately, Richardt et al. [10] share our sentiments and have created a set of five synthetic stereo sequences with associated disparity maps, which we use for evaluation. Fig. 2 contains five frames of ”Street”, one of these sequences, processed with our method. The results again are a set of consistent estimates.

To simulate real sequences, we add Gaussian noise, distributed as $\mathcal{N}$(0, 20), to the synthetic videos. We add this noise because stereo images from real cameras will not completely match each other due to a variety of reasons, such as luminance or sensor response differences. We run our method, HBP-TV, on all 5 sequences to compare with the two top image methods presented by Richardt et al., DCB and DCB2, and their spatio-temporal method, TDCB. Table 1 illustrates that our method achieves superior results on nearly every dataset by a significantly large margin using the percentage of bad pixels metric. A bad pixel is defined as any pixel that has an estimated disparity that deviates from the true value by an amount larger than a certain threshold (set at 1 in our case).

Table 1. Comparison of methods with noise $\sim\mathcal{N}(0, 20)$. Average percent of bad pixels (threshold of 1) for all frames.

Technique	Book	Street	Tanks	Temple	Tunnel
HBP-TV	**26.97**	**17.69**	**26.50**	**18.01**	29.50
TDCB	38.95	24.17	29.34	29.89	33.01
DCB	47.24	30.91	33.56	37.59	**24.04**
DCB2	53.92	38.02	45.67	40.97	31.19

Table 2. Versatility of TV for the various disparity methods. Average percent of bad pixels (threshold of 1) for all frames.

Technique	Book	Street	Tanks	Temple	Tunnel
TDCB-TV	**27.10**	**17.45**	**23.25**	**21.94**	**32.21**
TDCB	38.95	24.17	29.34	29.89	33.01
DCB-TV	**35.31**	**22.45**	**23.00**	**27.38**	**22.41**
DCB	47.24	30.91	33.56	37.59	24.04
DCB2-TV	**48.66**	**31.91**	**41.28**	**32.14**	**30.43**
DCB2	53.92	38.02	45.67	40.97	31.19

We next evaluate the efficacy of our TV method in improving arbitrary disparity estimates. We take the output of each of the three competing methods and apply TV enhancement. Table 2 tabulates these results. In every instance, we enhance the disparity estimation technique, even with TDCB, the method that already incorporates temporal information.

To validate the robustness of our method, we allow the additive noise to range from a σ of 0 to 100 as [10] does to evaluate performance under a large range of signal degradation. This time we leave out the left-right consistency post-processing step. Fig. 3 illustrates the results we achieve. Again, our method nearly always lower bounds the competing method. For the sake of space, we only show the results of the overall best competing method (TDCB). For more results and video sequences, please refer to our website: http://videoprocessing.ucsd.edu/~ramsin/research/disparity.

5. CONCLUSION

We have proposed a robust video disparity estimation technique that enforces spatio-temporal consistency among all frames. Our two-part method first performs image-based disparity estimates, and then, treating each frame as part of a space-time cube, we apply 3D total variation minimization to remove noise and enhance results. We have shown the method to be resilient even in large amounts of noise. Furthermore, we have illustrated that in general our method can be used to refine the results of any disparity estimation technique suffering from impulsive noise or estimation errors.

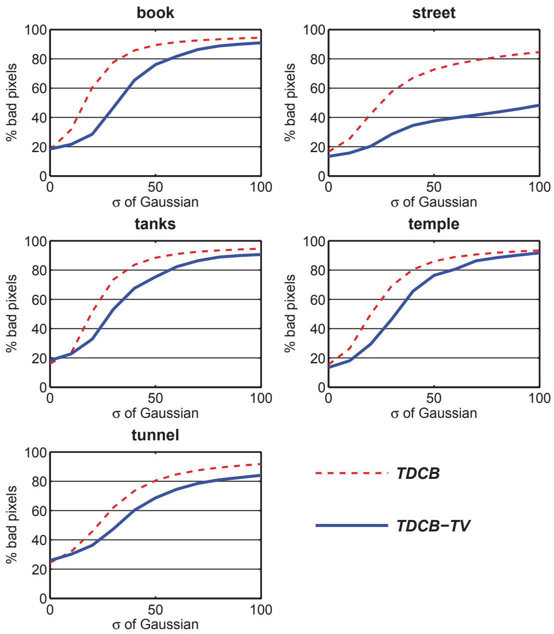

Fig. 3. Error plots for increasing Gaussian noise. Our method consistently produces less error when applied to TDCB.

6. REFERENCES

[1] P. An, Z. Zhang, and L. Shi, "Theory and Experiment Analysis of Disparity for Stereoscopic Image Pairs," in *ISIMP*, 2001, pp. 68–71.

[2] H. Hirschmuller and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching," in *CVPR*, 2007.

[3] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *PAMI*, vol. 23, no. 11, pp. 1222–1239, February 2004.

[4] V. Kolmogorov and R. Zabih, "Computing Visual Correspondence with Occlusions via Graph Cuts," in *ICCV*, 2001, pp. 508–515.

[5] P. Felzenszwalb and D. Huttenlocher, "Efficient Belief Propagation for Early Vision," in *CVPR*, 2004, pp. 261–268.

[6] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *IJCV*, vol. 47, pp. 7–42, April 2002.

[7] O. Williams, M. Isard, and J. MacCormick, "Estimating Disparity and Occlusions in Stereo Video Sequences," in *CVPR*, 2005.

[8] F. Huguet and F. Devernay, "A Variational Method for Scene Flow Estimation from Stereo Sequences," in *ICCV*, 2007.

[9] M. Bleyer and M. Gelautz, "Temporally Consistent Disparity Maps from Uncalibrated Stereo Videos," in *ISPA*, 2009.

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

[11] K. J. Yoon and I. S. Kweon, "Locally Adaptive Support-Weight Approach for Visual Correspondence Search," in *CVPR*, 2005.

[12] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps," in *CVPR*, 2008, pp. 1–8.

[13] G. Zhang, J. Jia, T. T. Wong, and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence," *PAMI*, vol. 31, no. 6, pp. 974–988, 2009.

[14] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen, "An augmented lagrangian method for total variation video restoration," in *ICASSP*, May 2011.

[15] "Mobile 3DTV," http://sp.cs.tut.fi/mobile3dtv/stereo-video/.