MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Disparity Estimation of Misaligned Images in a Scanline Optimization Framework

Rzeszutek, R.; Tian, D.; Vetro, A.

TR2013-025 May 2013

Abstract

Modern, state-of-the-art disparity estimation techniques are able to very accurately estimate the disparity for a wide variety of scene types. However all of these methods assume that the input images are epipolar rectified. When an image pair is not rectified, it must be pre-processed before any estimation can be done. In this paper we propose a disparity estimation scheme that is able to handle non-rectified images without requiring a rectification step. We show how a minor modification to an existing estimation framework can allow for any disparity estimation framework to produce disparity maps for non-rectified images.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DISPARITY ESTIMATION OF MISALIGNED IMAGES IN A SCANLINE OPTIMIZATION FRAMEWORK

Richard Rzeszutek[1†] *Dong Tian*[2*] *Anthony Vetro*[2]

[1] Department of Electrical and Computer Engineering, Ryerson University, Toronto, Ontario, Canada

[2] Mitsubishi Electric Research Laboratory, Cambridge, Massachusetts, USA

ABSTRACT

Modern, state-of-the-art disparity estimation techniques are able to very accurately estimate the disparity for a wide variety of scene types. However all of these methods assume that the input images are epipolar rectified. When an image pair is not rectified, it must be pre-processed before any estimation can be done. In this paper we propose a disparity estimation scheme that is able to handle non-rectified images without requiring a rectification step. We show how a minor modification to an existing estimation framework can allow for *any* disparity estimation framework to produce disparity maps for non-rectified images.

Index Terms— Stereo Image Processing, Stereo Vision, Spatial Filters, Optimization

1. INTRODUCTION

While there are a large number of different methods available for estimating the inter-pixel disparity between a stereo image pair, the fundamental assumption of all of these methods is that the input pair is epipolar rectified [1]. This assumption simply states that when given rectified images, the relationship between the same feature in both images is a two-dimensional translation. This relationship is then further restricted so that the translation is only along the horizontal or 'x' axis of the image coordinate system. The purpose of this requirement is so that searching for equivalent features only has to be done along horizontal rows, or "scanlines". Physically this corresponds to the two cameras being positioned at the same height with parallel lines-of-sight.

In practice this requirement cannot usually be met for natural images as there will always been some degree of misalignment with physical camera setups [2]. However, modern disparity estimation methods are robust enough such that small misalignments have little effect on the final result. This is a natural consequence of being robust against errors such as mismatched colours, camera sensor noise, reflections, etc. But, as the misalignment grows larger so does the likelihood of error in the disparity estimate.

One approach to deal with this is to pre-process the input images such that it appears as if they were taken by aligned cameras. This process, known as rectification, is straightforward if the cameras are "calibrated", i.e. their intrinsic parameters are known. However, if these parameters are unknown then an *uncalibrated* rectification must be used. This is a very difficult problem to solve and can require the solution to a non-linear system [3]. Furthermore, depending on the cameras' physical configuration and the nature of the scene, it is quite possible for the rectification process to produce useless transforms. E.g. if the estimate of the fundamental matrix results in the epipole being inside of the image then a rectification is not possible.

Another approach is to perform a two-dimensional search. This removes the need for an initial rectification step and the ambiguities that may results from it. In this case the disparity estimation procedure begins to resemble optical flow [4]. However, while they are related problems, disparity and optical flow are not the same. Optical flow attempts to characterize the apparent two dimensional motion of pixels between images while disparity is the apparent offset along epipolar lines. This small distinction means that disparity estimation methods can be much more efficient then optical flow methods as they only need to perform a one dimensional search and do not require solving large linear systems.

There are a number of instances where misaligned image handling is very useful. For instance, in depth-based image rendering [5] the disparity map is necessary in order to render novel viewpoints. Obtaining rectified stereo pairs is possible under controlled conditions but in cases such as automated processing of user-generated content this is not the case. Optical flow can be used to handle misaligned images but its computational burden makes it difficult to use. However, much work has been done on real time disparity estimation and it would be useful to extend existing approaches to handle misaligned images to minimize the computational cost.

Very little prior work has been done in estimating the disparity of misaligned images. Nalpantidis et al [6] used the hierarchical block matching scheme from MPEG video coding to replace the normal one-dimensional search with a two-dimensional search. The authors dropped this into a simple Winner-Take-All (WTA) framework to obtain the final

*Please direct all correspondences to tian@merl.com. †This work was done while R. Rzeszutek was on an internship at MERL.

disparity values. Thévenon et al [7] performed a full two-dimensional search and utilized dynamic programming (DP) to improve the quality of the final disparity map. This helps to resolve some of the issues with WTA-based methods, namely that they suffer in regions with little to no texture.

2. METHODOLOGY

Our estimation system is based on the semi-global matching (SGM) method developed by Hirschmuller [8]. We use a different cost-aggregation strategy to minimize the number of paths that need to be integrated. However, the fundamental improvement that allows us to handle misaligned images can be applied to *any* disparity estimation system, not just an SGM-based one. A complete flowchart of our method is presented in Fig. 1.

2.1. Disparity Estimation

We first describe the operation of our system for the case of aligned cameras. Given a reference image I_{ref} and target image I_{tgt}, we first compute a cost volume $C(x, y|d)$ such that

$$C(x, y|d) = \text{MC}\left(I_{ref}(x, y), I_{tgt}(x - d, y)\right), \qquad (1)$$

where $\text{MC}(\cdot)$ is the matching cost function. We then filter the cost volume using guided image filters [9] in the exact same manner as was developed by Hosni et al [10]. We also utilize the same cost function and ask the reader to refer to that paper for a full derivation. However, unlike Hosni et al we do not obtain the disparity map $d(x, y)$ by solving

$$d(x, y) = \underset{d}{\text{argmin}}\{C'(x, y|d)\}, \qquad (2)$$

where $C'(x, y|d)$ is the filtered cost volume. Rather we find the optimal disparities by applying Hirschmuller's SGM method to $C'(x, y|d)$.

The advantage to this approach is that it allows us to simplify the SGM stage by decreasing the number of paths through the cost volume. In the original method matching was done on a per-pixel basis and the integration was done in eight directions. With pre-filtering the number of paths can be reduced to the four cardinal directions. This also allows us to enforce inter-scanline consistency, something that is difficult to do with DP-based methods.

Another advantage is that it makes the estimation method itself more robust to noise. While guided image filters are very powerful, in the end the method presented in [10] is a simple WTA approach. It is still susceptible to errors in occluded and untextured regions.

2.2. 2D Search

The crux of our method is the two-dimensional search that we use for misaligned stereo pairs. For a one-dimensional search, the cost volume is three-dimensional. However this volume becomes four-dimensional for a two-dimensional search because the matching cost now depends on both a horizontal *and* vertical offset such that

$$C(x, y|d_x, d_y) = \text{MC}\left(I_{ref}(x, y), I_{tgt}(x - d_x, y - d_y)\right). \qquad (3)$$

The new terms, d_x and d_y, now refer to the horizontal and vertical displacements, respectively.

It is still desirable to process this higher dimensional cost volume using an SGM scheme. An underlying assumption of SGM is that the minimum cost path is traversed on a 2D slice of the cost volume. By moving to the higher dimensional cost volume, a "slice" is now a 3D volume, breaking this assumption.

To allow us to use an SGM scheme we construct a compressed 3D cost volume $\tilde{C}(x, y|d_x)$ from the original 4D cost volume. This is done by

$$\tilde{C}(x, y|d_x) = \underset{d_y}{\text{argmin}}\{C(x, y|d_x, d_y)\}. \qquad (4)$$

We also record the associated vertical disparity and construct a compressed cost table that maps a horizontal offset to its lowest cost vertical offset. The compressed cost volume is then processed with the SGM optimization to produce the final disparity map. Because each horizontal offset had an associated vertical offset, this information is also part of the disparity map. This allows us to keep the vertical offsets coupled to the horizontal offsets.

This process is visually shown in Fig. 2. For any given pixel located at (x, y), there is an associated 2D cost volume (Fig. 2a). After compression, the cost volume is now 1D

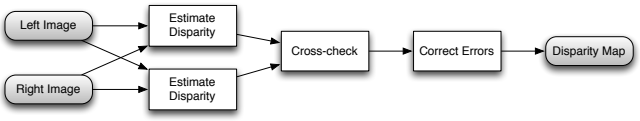
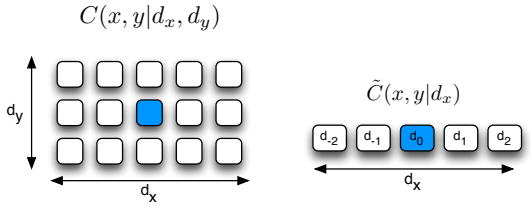


Fig. 1: Flowchart outlining the proposed disparity estimation method.



(a) Original (b) Compressed

Fig. 2: A graphical example of the cost compression for a particular pixel before compression (Fig. 2a) and after compression (Fig. 2b).

(Fig. 2b). Each horizontal disparity has a best matching vertical disparity d_n, where n is the index of the horizontal offset. In effect, this creates a look-up-table for the horizontal disparities.

2.3. Cross-checking and Error Correction

After the disparity estimation is complete there are still some errors in the disparity map. These errors are often a result of two things: mismatch errors and occlusion errors. While the optimization process can reduce mismatch errors, it cannot eliminate them and cannot remove occlusion errors at all.

To handle this we use the same strategy as in a number of different estimation methods: obtain left-to-right and right-to-left disparity estimates and compare the results. This is known as "cross-checking" and we do this using the same method outlined in [8] (it is in fact a very common method). Once the errors have been identified, we use a simple two-stage approach.

First we median filter the two disparity maps to eliminate any mismatch errors. These tend to be very small and only a few pixels or so in area. After this is done we replace any pixel in an occluded area with the *minimum* disparity of the two maps. The rational behind this is that this region often contains background which will be of a lower disparity value. The result is our final disparity map.

2.4. Relation to Prior Work

As we discussed in Section 1, very little work has been done regarding disparity estimation of misaligned images. Typically most estimation has focused on aligned image pairs as this is a simpler problem to handle. Of the work that has been done, our work most closely resembles that of Thévenon et al [7]. The major difference is that their method did not take inter-scanline consistency into account and required a relatively complicated DP optimization scheme.

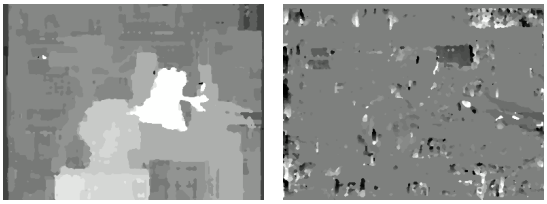
Our extension to an existing method is much simpler and can be used with *any* estimation method, not just one based on SGM. In fact, in a WTA framework our method very closely resembles that of Nalpantidis et al [6]. But unlike Nalpantidis et al we formalize our method in terms of the matching cost volume. This gives our approach a degree of flexibility not found in prior work.

3. RESULTS

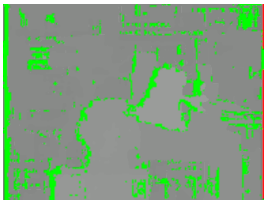
We demonstrate the efficacy of our method in this section. This is accomplished by comparing the disparity maps that are produced when the vertical disparity estimation is enabled versus when it is disabled. The only parameter that we change is the size of the matching window. For smaller images, too large of a window can cause small details to disappear. Conversely, for larger images, too small of a window can result in more noisy disparity maps. For the purposes of this paper we do not employ sub-pixel disparity refinement. All of the results presented contain integer-valued disparities.

3.1. Controlled Misalignment

Our first example is the well-known Tsukuba test image [11]. The disparity maps produced by our method on the unmodified images are shown in Fig. 3. Because of the relatively small size of the image (384x288) we used a matching window size of 9x9.



(a) Horizontal Disparity (b) Vertical Disparity



(c) Occlusions

Fig. 3: The output of the proposed estimation method on the "Tsukuba" test image.

Note that for even a perfectly rectified image pair, the vertical disparity map (Fig. 3b) has non-zero values. These occur along object edges where depth ambiguities naturally occur. As can be seen in the occlusion map (Fig. 3c), the errors in the vertical disparities occurs in roughly the same location as occluded and disoccluded pixels. This is expected as these are regions where reliable disparity estimates are not possible.

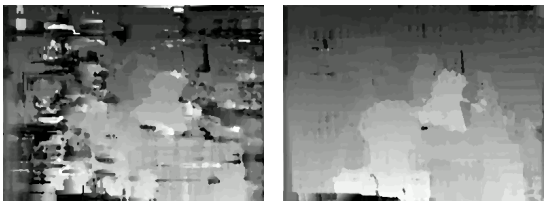
To show the effect of distortion, we rotate the right stereo image by five degrees. The rotated image has also been cropped so that it is the same dimensions as the left image. Fig. 4 shows the new stereo pair.

(a) Left Image (b) Right Image

Fig. 4: Distorted "Tsukuba" stereo image pair.

Fig. 5 shows a comparison between a disparity estimation without the vertical processing and one with the vertical

processing. Without any sort of processing the resulting map (Fig. 5a) is of limited use and very noisy. When the processing is enabled (Fig. 5b) the structures in the image pair are clearly visible. There is some distortion, namely the apparent gradient running from bottom to top. However this is much more useful than the result in Fig. 5a. Fig. 6 shows the vertical disparity map. Note the relatively smooth gradient from left to right, as is expected by a simple rotation.



(a) Without Vertical (b) With Vertical

Fig. 5: A comparison between disparity estimation on a distorted image pair with and without vertical processing.

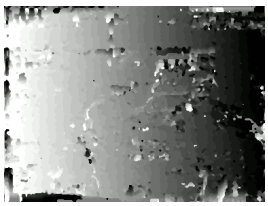


Fig. 6: Vertical disparity map associated with Fig. 5b.

This particular example is somewhat contrived in that it is highly unlikely that stereo cameras would exhibit this degree of rotational misalignment. Even a five-degree offset is quite large. However, this is an important result as it shows how effective vertical processing can be.

3.2. Uncontrolled Misalignment

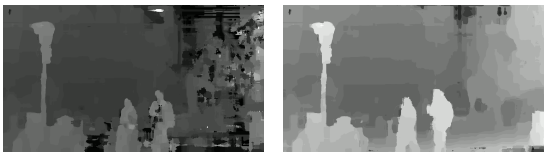
A practical application of our method is demonstrated by obtaining the disparity estimate from a frame in the "Summer in Heidelberg"[1] short film. The chosen frame is shown in Fig. 7 while the resulting disparity estimates are shown in Fig. 8. As this was from 720p footage, the matching window size was set to 25x25.

The map without vertical processing (Fig. 8a) is noticeably more noisy on the right side of the image. This is due to the "toed-in" camera configuration where the camera lines of sight are not parallel. This causes growing vertical offsets as one moves farther away from the point of focus. The map with vertical processing (Fig. 8b) is much cleaner and a better representation of scene depth. As shown, without vertical processing the trees on the right side of the frame cannot be made out but are clearly visible with vertical processing.

[1] http://3dtv.at/Movies/Heidelberg_en.aspx

(a) Left (b) Right

Fig. 7: Frame 1880 from "Summer in Heidelberg".



(a) Without Vertical (b) With Vertical

Fig. 8: A comparison between between the estimation methods for the Heidelberg frame.

The vertical disparities are shown in Fig. 9. We have also tested the proposed approach extensively on professionally-produced sports content that was captured with a toe-in camera configuration, and found the proposed disparity estimation approach to be very robust to the inherent misalignment.

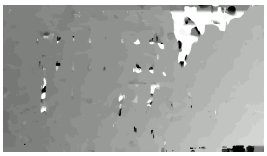


Fig. 9: Vertical disparities associated with Fig. 8b.

4. CONCLUSION

In this paper we presented a disparity estimation method that is able to obtain a disparity map for stereo images taken with misaligned or non-rectified cameras. We did this by proposing a simple modification to how the cost volume is created during the matching stage. The modification was done in such a way that makes it usable by a large class of disparity estimation methods, something that was not addressed in prior work.

There are a number of areas where our method could see some improvement. Namely, the vertical disparities are essentially found through WTA. This makes the vertical disparities noticeably noisier than the horizontal disparities. As such it would be useful to examine if a two-stage optimization would be possible: first optimize the vertical disparities and then optimize the horizontal ones.

5. REFERENCES

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*, Texts in Computer Science. Springer, 2011.

[2] S. Reeve and J. Flock, "Basic principles of stereoscopic 3d," http://www.sky.com/shop/export/sites/www.sky.com/shop/__PDF/3D/Basic_Principles_of_Stereoscopic_3D_v1.pdf.

[3] A. Fusiello and L. Irsara, "Quasi-euclidean uncalibrated epipolar rectification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, dec. 2008, pp. 1 –4.

[4] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of Mathematical Models in Computer Vision*, Nikos Paragios, Yunmei Chen, and Olivier Faugeras, Eds., pp. 237–257. Springer US, 2006.

[5] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," pp. 93–104, 2004.

[6] L. Nalpantidis, A. Amanatiadis, G. Sirakoulis, N. Kyriakoulis, and A. Gasteratos, "Dense disparity estimation using a hierarchical matching technique from uncalibrated stereo vision," in *Imaging Systems and Techniques, 2009. IST '09. IEEE International Workshop on*, may 2009, pp. 427 –431.

[7] J. Thévenon, J. M. del Rincón, R. Dieny, and J. C. Nebel, "Dense pixel matching between unrectified and distorted images using dynamic programming.," in *VISAPP (2)*, Gabriela Csurka and José Braz, Eds. 2012, pp. 216–224, SciTePress.

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

[9] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios, Eds., vol. 6311 of *Lecture Notes in Computer Science*, pp. 1–14. Springer Berlin / Heidelberg, 2010.

[10] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1, 2012.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, Apr. 2002.