DISPARITY MAP QUALITY FOR IMAGE-BASED RENDERING BASED ON MULTIPLE METRICS

Patrick Vandewalle and Chris Varekamp

Philips Group Innovation - Research
High Tech Campus 36, 5656AE Eindhoven, The Netherlands

ABSTRACT

While a wide range of disparity estimation algorithms exists, only few methods are available to analyze the quality of a disparity map. We present a new two-dimensional approach to disparity map evaluation using the stereo matching error and the temporal error. These measures give a good indication of stereo or multi-view performance on a 3D display. Results are illustrated on a ground truth disparity map with multiple distortions as well as on a commercially available disparity estimator using varying post-processing settings on a real stereoscopic video sequence.

Index Terms— disparity map quality, depth map, image quality

1. INTRODUCTION

Disparity estimation methods have received much attention in the past fifteen years. The Middlebury evaluation website lists more than 150 algorithms [10]. Disparity maps can be applied in a wide variety of applications, such as navigation of autonomous robots [6] or depth adjustment for stereoscopic displays [5, 7]. Depending on the application, the criteria for a good disparity map can be very different. For example, a robot arm needs accurate depth measurements with a minimum sample density for proper operation. This sample density is not necessarily per-pixel. However, for depth adjustment on stereoscopic displays or rendering for autostereoscopic displays, we need per-pixel disparity values, where depth accuracy is less important than temporal stability. In this paper, we will consider disparity maps for use in rendering for (auto-)stereoscopic displays.

In earlier work, Scharstein and Szeliski provided a taxonomy of stereo correspondence algorithms [10]. They measure the quality of an algorithm using the percentage of bad matching pixels B , defined as

$$B = \frac{1}{N_p} \sum_{(x,y)} (|d(x,y) - d_T(x,y)| > \delta) \cdot 100 [\%], \quad (1)$$

The authors would like to thank the anonymous reviewers for their valuable feedback.

where δ is the disparity error tolerance. The total number of pixels included in the sum is labeled N_p . Next to the disparity map d being analyzed, this measure requires a ground truth disparity map d_T for comparison. Moreover, pixels in occluded areas for the ground truth measurement can not be taken into account. At the same time, for autostereoscopic displays, a disparity value at every pixel location is needed in order to render visually pleasing 3D data.

Bosc et al. have proposed a quality metric on synthesized views using wavelet analysis [1]. They compare the original view with a rendered view after applying a block matching algorithm to compensate for the displacements between the two images. They perform the comparison using the histogram of the horizontal details subband of the wavelet domain image. Because we are most sensitive to distortions in the shape of faces and persons, a higher weight is given to those areas that contain skintones. Temporal instabilities are not explicitly taken into account.

Dumic et al. performed a comparison of multiple objective image and video quality metrics to the perceived quality on an autostereoscopic display [2]. They used 2D and depth video sequences and simulated packet losses on the depth maps as degradations. A better correlation is observed with the video quality measures on synthesized views, rather than calculating these measures directly on depth maps. The best correlation with the observed results was obtained using PSNR on the synthesized views.

We present a two-dimensional analysis of disparity map quality, using both the matching error with original stereoscopic content and a temporal instability measure. These will be described in the next two sections. Section 4 presents a method to evaluate the quality independently of the (linear) mapping of disparity values that was used. In Section 5, we evaluate our measure on a number of test scenarios, and finally Section 6 concludes this paper.

2. MATCH ERROR

In our approach, we assume that a ground truth depth map is not available. Instead, we use the disparity map in combination with the original stereo pair to evaluate its quality. From the left image and its corresponding disparity map, we

can generate a new right image using depth map based rendering [4].

The match error $E_m(n)$ for frame n is then defined as the mean absolute difference (MAD) between the original right image $R(x, y, n)$ and its warped equivalent $R_w(x, y, n)$ from the left image:

$$E_m(n) = \frac{1}{3N_P} \sum_{(x,y)} \|R(x, y, n) - R_w(x, y, n)\|_1, \quad (2)$$

where the sum is over all N_P pixels in the right image. Note that occluded pixels are implicitly predicted by extrapolating the local background.

When estimating the disparity map quality for video sequences, we also average the match error over all available frames:

$$\begin{aligned} E_m &= \frac{1}{N_T} \sum_n E_m(n) \\ &= \frac{1}{3N_P N_T} \sum_n \sum_{(x,y)} \|R(x, y, n) - R_w(x, y, n)\|_1, \end{aligned} \quad (3)$$

with n indicating the frame number, and N_T the total number of frames.

Instead of Scharstein and Szeliski's method of counting bad matching pixels in the disparity map [10], we measure pixel differences in the warped right image. This provides the added value of counting a bad match in a highly textured scene more than a bad matching pixel in a homogeneous region such as a blue sky. Moreover, it does not require a ground truth disparity map, which is lacking for most stereo data.

3. TEMPORAL ERROR

For the temporal error E_t , we use a very simple approach. We first compensate the previous frame for the estimated motion, and then compute the pixel-wise difference in disparity values for the two consecutive frames:

$$E_t(n) = \frac{1}{N_P} \sum_{(x,y)} |D(x, y, n) - D(x - m_x, y - m_y, n - 1)|, \quad (4)$$

where (m_x, m_y) is the estimated motion from frame n to frame $n - 1$. We used the approach by Farneback [3] to estimate the motion field. For the entire sequence, this leads to

$$E_t = \frac{\sum_{n=2}^{N_T} \sum_{(x,y)} |D(x, y, n) - D(x - m_x, y - m_y, n - 1)|}{(N_T - 1)N_P}. \quad (5)$$

Disparity maps with high temporal instability will lead to a higher temporal error E_t . Typically, such disparity maps will also show large areas of annoying flickering behavior on an autostereoscopic display.

Fig. 1. Example error surfaces for (7) with values for scaling a vertically and offset b horizontally. The optimal linear mapping values can be seen as the darkest point in the error surface.

Note that we compensated for the estimated motion in the video. This does introduce sensitivity to the motion estimation algorithm used, as inaccuracies in motion compensation also negatively affect the disparity map measure E_t . However, given that the true motion field is not available, this performs the best possible compensation.

4. DISPARITY MAPPING

In many cases, the disparity values are not available as true pixel disparities, but as an unknown linear scaling and offset of original disparities. This operation is typically done to match the available 8-bit value range, or to better fit the display hardware requirements.

We therefore introduce an estimation and compensation of such a linear mapping prior to the error analysis. We calculate the values of a and b in

$$D(x, y) = aD_0(x, y) + b, \quad (6)$$

with $D_0(x, y)$ the input disparities, and $D(x, y)$ the mapped disparities, such that $D(x, y)$ best matches the mapping from left to right image. This can be formulated as follows:

$$(a_{\text{fit}}, b_{\text{fit}}) = \arg \min_{(a,b)} \frac{1}{3N_P} \sum_{(x,y)} \|R(x, y) - R_w(x, y; a, b)\|_1. \quad (7)$$

As these calculations can be performed offline, we determine a_{fit} and b_{fit} using a full search over the parameter space. Some error surfaces are shown in Figure 1. The error surfaces tend to be very smooth, indicating that a more efficient optimization procedure could also be performed.

5. EVALUATION

For our first experiment, we use the Tank sequence made available by Richardt et al. [9]. This is a 100-frame computer-generated stereo sequence of 400x300 pixels with ground truth disparity maps. Next to the ground truth disparity map, we also used 6 sets of distorted disparity maps:

Fig. 2. Error analysis for ground truth disparity map vs blurred and noisy disparity maps.

- adding noise (standard deviation 1, 2 and 3 on the 8-bit disparity values, clipped to the 8-bit range $[0,255]$),
- blurring the disparity maps with multiple gaussian kernels (7×7 , 15×15 , 25×25).

The error measures for each of these sequences are shown in Figure 2, with the match error on the horizontal axis and the temporal error on the vertical axis. The ground truth disparity maps (marked with ‘o’) have the lowest match error. Even for the ground truth, it is non-zero, due to differences in the disocclusion areas, where previously invisible areas become visible in the synthesized view. We do not discard these areas, because the way in which disocclusion areas are filled also influences the overall perceived 3D performance. Match errors for the blurred versions of the disparity map (marked with diamonds) become larger as the kernel size increases, and the same happens for the noisy disparity maps as the noise level increases (marked with ‘+’). As expected, the noisy or blurred disparity maps also have a larger temporal error. Examples of the distorted disparity maps and their rendered right views are shown in Figure 3. They correlate well with the analysis from the error plots. The rendered view for the noisy disparity map has more irregular edges, while the rendered view for the blurred disparity map is very smooth, but is deformed around depth transitions such as the tank barrel and the tank antenna.

Next, we applied our error measures to the disparity estimator built into Matlab (v2013a, with computer vision toolbox [11]). We ran the estimator on a 1920×1080 stereo video sequence ‘Ticketline’, using the default algorithm settings and a disparity range of $[-32,32]$ pixels. The resulting quality measures are shown in Figure 4 (marked with ‘o’). An example depth map and rendered frame are shown in Figure 5(a) and (b), respectively. As these results are quite noisy, we applied several post-processing filters on these depth maps. We used a standard Gaussian blur filter (‘+’, 51×51 Gaussian

Fig. 3. Distorted disparity maps and corresponding rendered views. (a), (d) and (g), and (b), (e) and (h) show ground truth, noisy version (std 3.0) and blurred version (25×25) for a depth map and a rendered view, respectively. (c), (f) and (i) show details of (b), (e) and (h).

filter with $\sigma = 11$), as well as a median filter (‘x’, 51×51 window) and a cross-bilateral grid filter (‘*’, as described by Paris and Durand [8] and implemented in Matlab by Jiawen Chen). As parameters in the bilateral grid, we used $\sigma_{\text{spatial}} = 1080/32 = 33.75$ and $\sigma_{\text{range}} = 0.05$, where the pixel values were scaled to $[0,1]$. From Figure 4 we can see that all post-processing methods improve both the temporal and match error compared to the original estimator. The median filter results in the lowest match error. This can be explained by the fact that the median filter smoothens the disparity maps without inserting new disparity values, and thus it minimally distorts the stereo correspondences. The temporal error is lowest for the median filter and the bilateral filter, while it is somewhat higher for the blur filter. The images shown in Figure 5 and 6 are consistent with the results in Figure 4. The median filtered image looks most natural, with the least distortions.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a new method to analyze the quality of disparity maps. We compare the disparity maps on two aspects: match quality and temporal stability. This results in a visually very insightful plot showing how these two quantities are differently balanced in different algorithms. It can help in analyzing the differences between multiple disparity estimation algorithms, or in verifying how big the improvement of an updated algorithm is.

Fig. 4. Error analysis for disparity map estimated using Matlab's disparity estimator vs post-processed versions using a Gaussian blur, median filter and bilateral grid.

Fig. 5. Distorted disparity maps and corresponding rendered views for ticketline sequence. (a)-(b) Matlab disparity estimator. (c)-(d) Blur post-processing. (e)-(f) Median filter post-processing. (g)-(h) Bilateral grid post-processing.

Fig. 6. Details of Figure 5 (b), (d), (f) and (h), respectively.

In future work, it would be interesting to evaluate more algorithms using the proposed analysis. Also, the current analysis cannot handle disparity maps that are nonlinearly scaled, or in which a single object is changed in depth with respect to the rest of the scene. It would be interesting to study invariance to such changes, as they tend to generate visually pleasing images on autostereoscopic displays.

7. REFERENCES

- [1] E. Bosc, F. Battisti, M. Carli, and P. Le Callet. A wavelet-based image quality metric for the assessment of 3d synthesized views. In *Proc. SPIE/IS&T Electronic Imaging: Stereoscopic Displays and Applications*, volume 8648, 2013.
- [2] Emil Dumic, Sonja Grgic, Luis Alberto da Silva Cruz, and Pedro Assuncao. Objective quality measures comparison of impaired 3d video sequences from the uc3d database. In *Proc. IEEE 3DTV-Conference*, 2014.
- [3] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. *Lecture Notes in Computer Science*, 2749:363–370, 2003.
- [4] C. Fehn. Depth-image-based rendering (dibr), compression and transmission for a new approach on 3d-tv. In *Proc. SPIE Stereoscopic Image Processing and Rendering*, volume 5291, 2004.
- [5] H. J. Kim, J. W. Choi, A.-J. Chang, and K. Y. Yu. Reconstruction of stereoscopic imagery for visual comfort. In *Proc. SPIE/IS&T Electronic Imaging: Stereoscopic Displays and Applications*, volume 6803, 2008.

- [6] D. Murray and J. J. Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8(2):161–171, 2000.
- [7] H. Pan, C. Yuan, and S. Daly. 3d video disparity scaling for preference and prevention of discomfort. In *Proc. SPIE/IS&T Electronic Imaging: Stereoscopic Displays and Applications*, volume 7863, 2011.
- [8] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. In *Proc. European Conference on Computer Vision*, 2006.
- [9] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *Proc. European Conference on Computer Vision (ECCV)*, pages 510–523, 2010.
- [10] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7–42, 2002.
- [11] The Mathworks. Matlab, 2013. v2013a, including Computer Vision System Toolbox.