A DISPARITY RANGE ESTIMATION TECHNIQUE FOR STEREO-VIDEO STREAMING APPLICATIONS

Sergey Smirnov, Atanas Gotchev

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
Email: firstname.secondname@tut.fi

Miska Hannuksela

Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland
Email: miska.hannuksela@nokia.com

ABSTRACT

In this paper, we propose a robust and efficient technique for frame-by-frame estimation of disparity ranges in stereo video. The proposed technique utilizes a single-layer image size reduction for faster processing and more effective noise handling. Furthermore, it applies spatial-domain non-linear filtering of both disparity and confidence maps for additional noise suppression and improved range estimation. A mechanism for supporting temporal consistency is proposed as well. Performance comparisons with recent approaches demonstrate the advantages of the proposed approach.

1. INTRODUCTION

With the success of 3D cinema, applications utilizing stereo video have raised increased interest recently. Stereo video can be used for 3D scene reconstruction where geometry information about the scene is retrieved by approaches such as structure-from-stereo and structure-from-motion. Subsequently, the found geometry information in the form of depth map sequences can be used for manipulation of the video, that is to synthesize new desired views (free-viewpoint video) or retargeting the video for different stereoscopic displays ranging from high-definition imagery for home entertainment to mobile resolution for personal use on mobile devices. In some applications, quality is of primary concern (e.g. depth estimation for effective compression); in some other applications, the requirement for real-time performance is of primary importance. Examples include streaming of stereo video and its retargeting for display on different 3D displays. For all abovementioned applications, knowledge of the disparity range of a given stereo frame is of great use. For disparity estimation, knowledge of the disparity range helps avoiding too narrow or too wide searches. In the former case, estimation errors would be imposed, while the latter case would impose longer computational time. Memory consumption is also directly related with the disparity range. For the case of retargeting applications, knowledge of the disparity range helps is preserving the geometry of the scene and its adjustment to the visual comfort zone of the targeted display.

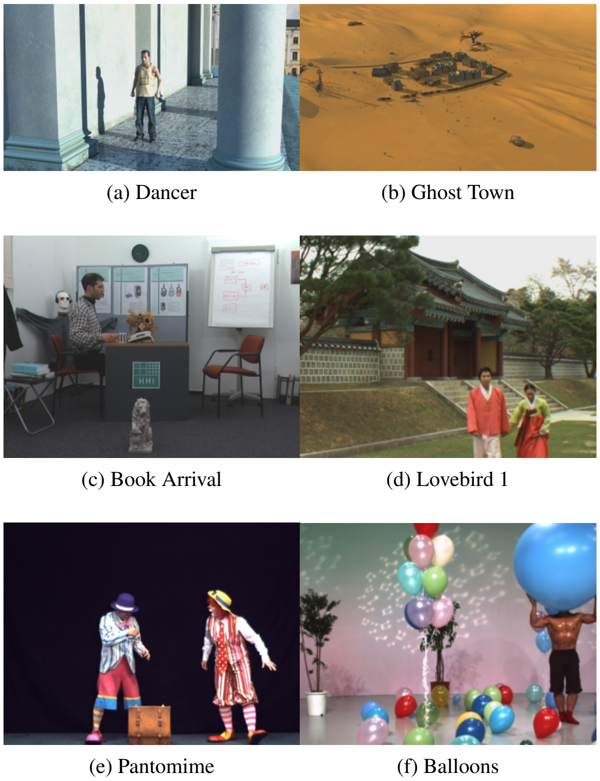

(a) Dancer
(b) Ghost Town
(c) Book Arrival
(d) Lovebird 1
(e) Pantomime
(f) Balloons

Fig. 1: Stereoscopic datasets, used in experiments

1.1. Prior art

The problem of automatic disparity range estimation has been recently addressed in a number of publications [1, 2, 3]. The approaches can be broadly grouped by the way they utilize image features for finding trustful correspondences for estimating the range of disparities. Some methods rely on multi-resolution approaches for finding dense correspondences, while other methods rely on finding sparse correspondences between scale-invariant feature points.

In [1], a coarse-to-fine range estimation approach has been proposed. The unknown disparity range is found by means of simplified stereo-matching applied on a Gaussian pyramid constructed from the input stereo pair of images. The main component of the approach is so-called Confidently Stable Matching (CSM). It aims at discarding the disparity estimate values which are either corresponding to occluded pixels or are with low estimation confidence. The disparity map at the coarsest level is used to calculate disparity histogram of the current layer, which is then adaptively thresholded in order to obtain some disparity limits. The disparity limits found at the coarser levels are used to constrain a CSM at the next (finer) pyramid layer and the process is iterated until finest layer is reached. Constraining the stereo-matching at pyramid layers allows achieving good performance with little redundancy.

The approach in [3] also employs some disparity histogram thresholding while the histogram values are calculated from disparities found between matched feature points. Speeded Up Robust Features (SURF) [4] are used and they are calculated at the full-resolution stereo image. To mitigate possible outliers in the estimated histogram, hard-thresholding is applied prior to the range estimation. The relatively low number of feature points used ensures low complexity of the approach. The approach is then extended from still images to image sequences to provide also some temporal consistency.

Preliminary disparity range estimation can help the improvement of general disparity estimation methods. In [2], an improvement of the graph cuts global optimization approach has been proposed utilizing a reduction of the search space. An extension to Markov Random Fields (MRF) stereo has been proposed in [5]. A similar approach for Belief Propagation has been proposed in [6]. For all these approaches, improvements are due to having precise disparity range estimate before the actual execution.

2. PROPOSED TECHNIQUE

Given a stereo video sequence, the problem is to find per-frame lower and upper limit of the disparities between the left and right view images. The procedure should be fast, automatic and content-independent. For streaming applications it should be also working in real time with possible initial buffering to use previous frames in the estimation procedure, if needed. The disparity range estimate relies on collecting a number of correspondences between points in the left and right image frames and calculating an empirical disparity histogram which needs to be properly thresholded to eliminate erroneous disparity estimates at the both ends of the range.

Our technique aims at combining the approaches proposed in [3] and [1] and modify them so to achieve faster and better performance. While comparing the two state-of-the-art approaches, one can conclude that the approach suggested in [3] is faster due to the fact that it uses a sparse set of stereo correspondences between feature points in the two views. In contrast, the approach in [1] relies on dense correspondence matching though achieved in a coarse-to-fine manner. However, it is precisely the dense set of correspondences found which makes the method more robust against outliers and provides more stable disparity histogram. Therefore, we adopt the later approach as a starting point of our modified technique.

2.1. Downsampling

While the general coarse-to-fine approach is reducing the computational cost, it is also a source of inter-layer error propagation. Finer layers use a disparity search range which depends on the disparity range estimate at the coarser layer. There are situations where the disparity range, inevitably shrunk at the coarse layer, cannot be fully expanded at the finer layers. Even two Gaussian pyramidal layers can cause range shrinkage propagation errors. In the original approach, the authors have suggested to use a course pyramidal layer with a sufficient size (in the spatial range of 100 pixels) in order to avoid over-smoothing of details and hence missing some valuable details for finding stereo correspondences [1]. In our proposal, we suggest performing only a single layer size reduction with a controllable downsampling parameter. This provides a flexible mechanism to trade-off computational complexity for precision of details and avoiding error propagation. The downsampling factor is controllable by some preliminary knowledge of the disparity range. In the case of video, this can be the disparity range estimated from the previous frame which determines the search range and correspondingly the best scale where this can be done without causing over-smoothing. Furthermore, at the coarse layer, we reduce the disparity histogram threshold (γ_{hist}) with approximately 0.1 - 0.2% in order to avoid range shrinkage.

2.2. Spatial filtering

Disparity estimates made with a single-layer coarse-to-fine approach are prone to matching errors, the so-called outliers. The method in [1] handles outliers by a CSM mechanism. Still, the resulting disparity histogram might contain spurious peaks, which lead to erroneous histogram thresholding. An empirical analysis of the spatial distribution of correctly and wrongly matched pixels and their confidence values demonstrated a different statistical behavior, which suggested an optimal spatial filtering approach for the removal of erroneous matches. In our approach, we use small-kernel non-linear spatial filtering applied to both disparity and confidence maps, prior to histogram estimation. More specifically, our empirical study has shown that 2D medial filtering is optimal for the type of noise presented in the disparity and confidence estimates. Contrary, mean filters (e.g. Gaussian) lead to smoothing edges in disparity maps. More comprehensive non-linear filters (e.g. bilateral) bring marginal improvement for the price of much higher computational cost.

2.3. Temporal consistency

In the case of 3D video, temporal stability and smoothness in the disparity range estimate is an important issue [3]. Visualization of 3D scenes usually requires a rather smooth disparity transition between scene cuts. This gives one more opportunity to speed up the processing by imposing some initial search range based on previous frames. In our approach, we define a range extension parameter (γ_{ext}) which defines the initial search range as follows:

$$mindisp_{i+1}^{initial} = mindisp_i - \gamma_{ext} \tag{1}$$

$$maxdisp_{i+1}^{initial} = maxdisp_i + \gamma_{ext} \tag{2}$$

This provides some flexibility in finding the range limits, starting at a reasonably-wide initial range. The extension parameter also allows an adaptation to ranges in scenes with more abrupt cuts. For such scenes, the convergence is satisfactory fast.

3. EXPERIMENTAL RESULTS

3.1. Dataset

The stereoscopic datasets used in the experiments are illustrated by thumbnails in Figure 1. The following sequences were used: ”Dancer” and ”Ghost Town” by Nokia Research Center; ”Book Arrival” by Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI); ”Pantomime”, and ”Balloons” by Nagoya University; ”Lovebird 1” by Electronics and Telecommunications Research Institute (ETRI). For the synthetic scenes ”Dancer” and ”Ghost Town”, grand true depth data is available, therefore true disparity limits can be found as well. For the indoor scenes ”Book Arrival” and ”Balloons” we assumed fixed disparity ranges. The ground true was estimated by a very precise dense stereo matching. For the scenes ”Pantomine” and ”Lovebird 1”, the disparity limits were first found automatically and then inspected, to manually remove unnatural peaks.

3.2. Experimental settings

Along with our technique, we have implemented also the coarse-to-fine and the feature-based approaches, as described in the related papers [2], [4]. The approach in [4] was applied in two forms with and without temporal filtering. The histogram thresholding approach described in [2] was found more reliable and better performing than the one in [4] therefore we applied it to both histogram estimates. Our approach is denoted in the figures as ’proposed’.

All experiments were performed with the following computer configuration: DELL OPTIFLEX 960 with 4Gb of RAM, 3Ghz Intel Core 2 Duo CPU, Microsoft Windows XP OS. The control (framework) scripts were written in MATLAB, while the main algorithms were written in C/C++ and compiled to MEX-function in order to be run from the MATLAB environment. Our feature-point based algorithm implementation uses OpenSURF C++ code from http://code.google.com/p/opensurf1. As an underlying matching method, the coarse-to-fine and the proposed technique implementation use a constant-complexity square window SAD stereo matching written in C++. The programs were compiled by Microsoft Visual Studio 2008, with highest performance settings. OpenMP optimizations in all codes were used where possible.

Dataset	Coars.-fine	SURF	Proposed	Proposed (1st frame)
Dancer	77.7	11.1	6.4	117.8
Ghost Town	81.1	15.5	8.8	120.6
Book Arrival	17.3	4.7	3.9	43.5
Lovebird 1	17.1	3.3	3.5	41.8
Pantomime	33	6.4	7.8	67.33
Balloons	17.2	2.7	3.6	41.8

Table 1: Average computational times in seconds per frame for considered approaches

3.3. Performance evaluation

We evaluate the performance of the three methods by average absolute error, calculated as follows:

$$AAE = \frac{1}{N}\sum_{i=1}^{N}\{|maxdisp_i - ma\hat{x}disp_i| + \\ + |mindisp_i - min\hat{d}isp_i|\},$$

where $mindisp_i$ and $maxdisp_i$ are true limits of i-th frame for current dataset, and $min\hat{d}isp_i$ and $ma\hat{x}disp_i$ their estimates. N is the number of frames in the dataset.

A technical tweak can additionally improve the estimates. We have used what we call a guard interval parameter to tackle problems with ’too aggressive’ histogram thresholding. Namely, we reduce the $mindisp$ by 1 and increase $maxdisp$ by 1. This lead to improvement in all compared approached for all datasets. Higher guard intervals were rather content-dependent and we did not take them into accounts in comparisons.

3.4. Results

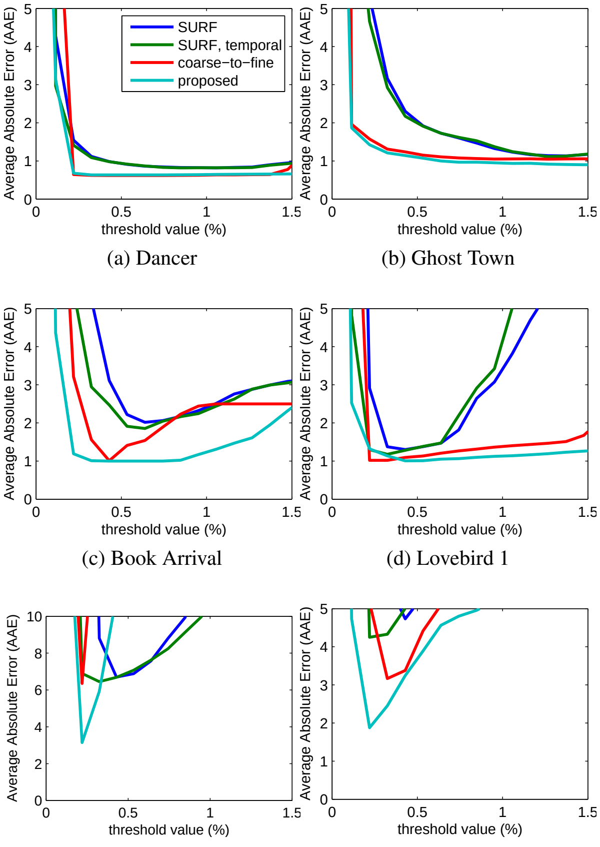
The three methods have been compared in terms of absolute error versus the histogram threshold values. Such a comparison would characterize the methods in terms of their robustness for different content. Figure 2 shows the results. For the cases of synthetic, noise-free sequences (i.e. ”Dancer” and ”Ghost Town”) the performances of the coarse-to-fine approach and the proposed approach are pretty similar. For the rest of sequences, representing real-world scenes and including noise and other imaging imperfections, the proposed approach shows a better performance compared to the coarse-to-fine one. The approach based on matching of sparse SURFs shows inferior performance for most of the sequences both in its still-image and video versions (enforcing temporal consistency).

Another noticeable advantage of the proposed approach is that the threshold used to limit the histograms is somehow content-agnostic. The best value is around 0.2% across all test sequences. This demonstrates much more consistent behavior compared with the other approaches which reach minima at different threshold points for different test sequences.

Table 1 shows the average computational times for all considered approaches. For fair comparison, the downsampling ratio was fixed to 2. Therefore, the proposed technique shows considerable computational time for the first frame after which it gets much better due to the suggested range adaptation.

4. CONCLUSIONS

In this paper, two recently-proposed approaches for automatic disparity range estimation in stereo video were analyzed and compared, namely a course-to-fine pyramidal approach and an approach utilizing feature-based sparse disparity estimation. A modification of the former one has been proposed, which increases the speed and also improves the performance. The modifications include an adaptive downsampling scale selection and proper tuning of the histogram threshold parameter. The downsampling scale is driven by information about the disparity range of the previous stereo frame in the video. This effectively tackles the problem with disparity over-smoothing and disparity range shrinkage. Furthermore using single pyramidal layer with adaptive scale selection prevents any error propagation across pyramidal layers. The approach assumes also some temporal consistency and demonstrates very consistent performance for a large set of test video sequences. Namely, it turns out that the histogram thresholding is much more content independent while compared with the other approaches. In terms of computational performance, the method is slower only for the first frame, and after that it is considerably faster which makes it a good candidate for streaming applications where fast and reliable disparity range estimation is required.



(a) Dancer (b) Ghost Town

(c) Book Arrival (d) Lovebird 1

(e) Pantomime (f) Balloons

Fig. 2: *Average absolute error* results for selected datasets

5. REFERENCES

[1] J. Kostkova and R. Sara, "Automatic disparity search range estimation for stereo pairs of unknown scenes," in *Proceedings of the Computer Vision Winter Workshop 2004 (CVWW'04)*, February 2004, pp. 1–10.

[2] O. Veksler, "Reducing search space for stereo correspondence with graph cuts," in *Proceedings of the British Machine Vision Conference (BMVC)*, September 2006, pp. 709–718.

[3] Z. Arican D. Min, S. Yea and A. Vetro, "Disparity search range estimation: Enforcing temporal consistency," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 2366–2369.

[4] T. Tuytelaars L. Van Gool H. Bay, A. Ess, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[5] H. Jin L. Wang and R. Yang, "Search space reduction for mrf stereo," in *Proceedings of the 10th European Conference on Computer Vision: Part I*, 2008, ECCV 08, pp. 576–588.

[6] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 1458–1465.