Dense wide-baseline disparities from conventional stereo for immersive videoconferencing

Spela Ivekovic, Emanuele Trucco
Heriot-Watt University
EECE, EPS
Edinburgh EH14 4AS
{si1, e.trucco}@hw.ac.uk

Abstract

We propose an algorithm creating consistent, dense disparity maps from incomplete disparity data generated by a conventional stereo system used in a wide-baseline configuration. The reference application is IBR-oriented immersive videoconferencing, in which disparities are used by a view synthesis module to create instantaneous views of remote speakers consistent with the local speaker's viewpoint. We perform spline-based disparity interpolation within non-overlapping regions. Regions are defined by discontinuity boundaries identified in the incomplete disparity map. We demonstrate very good results on significantly incomplete disparity data computed by a conventional correlation-based stereo algorithm on a real wide-baseline stereo pair acquired by an immersive videoconferencing system.

1 Introduction and related work

Videoconferencing systems have recently been reported relying on multi-view computer vision (e.g., VIRTUE, the HP Coliseum, NTII [2, 8, 4, 3, 1]), some of which immersive, i.e., aiming to achieve telepresence (or tele-immersion) [8, 1]. The number of views, and consequently of the input colour video channels, is currently limited in immersive systems by the need of frame-rate processing of high-quality images on mostly conventional platforms [4, 3] and small maximum latency (approximately 200 ms, after which conversation-style communication is disrupted and structured turn-taking becomes necessary). As a consequence, no more than four or six cameras are normally used.

Figure 1. The VIRTUE immersive videoconferencing station. Four cameras observing the local participant are mounted around the screen.

The cameras observing the local participant must be placed around the screen (Figure 1). The local participant is therefore observed from a different viewpoint than the remote participant's one, creating the need for view reprojection, in turn needing disparity data [8]. If the screen is large for the purpose of immersion (say 60 inches or more) and at a short distance from the speaker (typically around 0.5 m), the cameras generate a wide-baseline multi-view configuration. Figure 2 shows, for example, a stereo pair acquired by the left stereo pair of cameras of the VIRTUE prototype (Figure 1), for which the baseline was approximately 80cm, and the subject at approximately 1 metre from the screen. In such a configuration, disparities may cover large intervals and local appearance be severely altered by scale, perspective and occlusions. This is exacerbated by gestures bringing arms and hands much closer to the cameras than the body.

As a consequence, fast, conventional stereo algorithms result in very incomplete disparity maps, especially after enforcing left-right consistency. Simple, fast interpolation schemes can be used to recover dense disparities, but at the cost of introducing artifacts in the synthesised images. Figure 3 shows the left-right disparity map computed for the pair in Figure 2 and the result of straightforward linear interpolation to complete the missing regions. The result ignores boundaries between different depth regions (e.g., arms and body), creates outliers which do not belong to any of the real depth regions, and the hand is not recovered properly. The consequences for view synthesis are illustrated eloquently by Figure 7. Clearly, algorithms estimating dense disparity

Figure 2. VIRTUE stereo pair (cameras on left and top of the screen, see Figure 1).

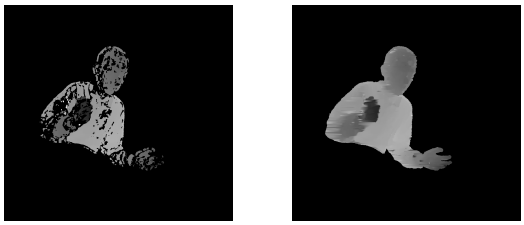

Figure 3. Left, disparity map after the left-right consistency check; Right, the result of the postprocessing using straightforward interpolation.

maps must preserve region discontinuities and limit depth variations within regions to avoid outliers.

Notice that recent, robust algorithms for dense, wide-baseline stereo [6, 11] are inapplicable here as either not designed for real-time applications (e.g., [11] reports execution times of 15 minutes with 1500×1000 images, arguably equivalent to 2 mins with 300×250 images), or designed for sparse correspondences only [6].

We present an algorithm generating dense disparity maps from incomplete ones consistently with the inter-region and intra-region constraints pointed out above. The input is formed by disparity maps computed by a conventional, fast correlation-based stereo algorithm [9] enforcing left-right/right-left consistency. The resulting disparity maps contain a large number of missing regions (Figure 3 left). Smaller inconsistencies are removed by a simple median filtering. Larger holes undergo a more thorough approach. To guarantee that regions with no disparities are filled on the basis of consistent data (i.e., no interpolation occurs between regions with very different depth values), we segment the disparity map into depth layers. We also use hand segmentation. Hands are important as they suggest the location of the arms, and are usually the closest parts to the screen, generating the highest disparities and the sharpest discontinuities. The individual segmented depth regions are processed with cubic splines. No-data regions where the scene exercises sharp depth discontinuities, are filled using a direct interpolation scheme constrained by depth layers suggested by [7]. Any remaining no-data region is filled by extrapolation.

Disparity post-processing has been implemented in various IBR-oriented applications. Notably, Chang and Zakhor [5] mention briefly cubic spline interpolation as one of the steps in post-processing a disparity map, but give no details explaining the benefits of spline-based processing. The constrained interpolation scheme proposed by Schreer et al. [7] is primarily concerned with hand regions creating hand-body occlusions and sharp depth discontinuities. They use binary hand masks obtained from a hand-tracking algorithm locating the hand regions in the video images, then process any other regions with a simple texture-based inter- and extrapolation using texture similarity, but not the surrounding disparities.

The remainder of the paper is structured as follows. Section II step by step describes our algorithm for disparity map post-processing. Section III elaborates on the experimental results. Section IV describes our conclusions and ideas for future work.

2 The algorithm

2.1 Starting point

The starting point for the postprocessing is the result of a pyramidal correlation [9] including left-right consistency check (Figure 3 left) run on a wide-baseline stereo pair which has been previously foreground-segmented. The resulting disparity map contains a number of areas where the consistency check failed (or "holes", shown in black in Figure 3 left). A 7×7 median filtering is first run to close minor holes and remove the obvious, isolated outliers.

2.2 Depth segmentation

The resulting disparity map is segmented into different depth regions using multi-modal histogram segmentation. Depth segmentation does not produce homogeneous regions; due to the nature of the correlation algorithm, the segmented regions are interspersed with sparse values belonging to other regions. In the region homogenising step such values are removed using a 19x19 homogenising window and overwriting the regional outliers with the majority region index. At this point, the disparity map has been segmented into disjoint homogeneous depth regions.

2.3 Spline fitting within the depth regions

Next, cubic-spline approximation is used to recover the missing areas within individual regions. Notice that spline fitting is only performed *within* the segmented regions and not across region borders, in order to preserve region discontinuities. Before the actual spline fitting, disparity values within each region are processed with a RANSAC-like algorithm removing the ambiguous outlying local values -

disparities *within* the individual depth regions are expected to change smoothly.

2.4 Completing the hands

In videoconferencing, frequent gestures move the hands in front of the body and therefore close to the cameras. This generates much higher disparities than the body, as well as significant occlusions. The correlation algorithm falls prey of such problems, yielding wrong disparities which fail the consistency check. We use a binary hand mask obtained by the hand-tracking algorithm in [7], which recovers the hand area using a constrained interpolation approach suggested by [7].

2.5 Spline fitting across the region boundaries

The hands can generate the highest disparities, as they are the part of the body which usually comes closer to the camera. In the remainder of the disparity map, neighbouring regions are separated by smaller depth discontinuities. It may therefore make sense to use cubic-spline approximation across the borders of such limited discontinuities.

2.6 Finalising the disparity map

The previous step does not close all remaining holes in the disparity map. If the boundary of a region includes the figure's boundary (i.e., the figure-background boundary), data supporting spline approximation is available on one side only. Such regions are therefore filled by expanding the internal region disparities towards the boundary.

3 Experimental results

We have evaluated the proposed algorithm on stereo pairs from the VIRTUE videoconferencing station [2, 8]. The input stereo pair is shown in Figure 2.

Figure 4 shows the corresponding left and right disparity maps after the consistency check. The missing disparity data regions are marked in black. Figure 5 top left shows the left disparity map split into disjoint depth regions and 5 top right the result of spline fitting within individual depth regions. The result of the next step (hand completion) is shown in Figure 5 bottom left, and that of inter-region (across-boundary) spline fitting in Figure 5 bottom right. The final extrapolation step generates dense disparity maps shown in Figure 6 left and 6 right (left-right and right-left disparity maps, respectively).

To illustrate comparatively the benefits brought about by our algorithm to view synthesis, we show in Figure 7 top left a novel view synthesised using the original disparity map (no post-processing), the dense disparity maps achieved by

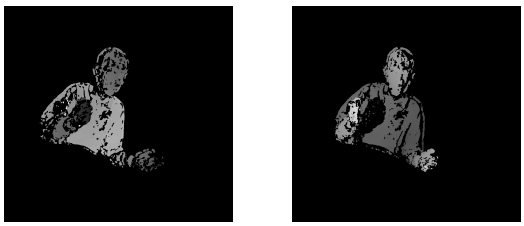

Figure 4. Left and right disparity maps after the consistency check. There are a number of regions with no data which need completing.

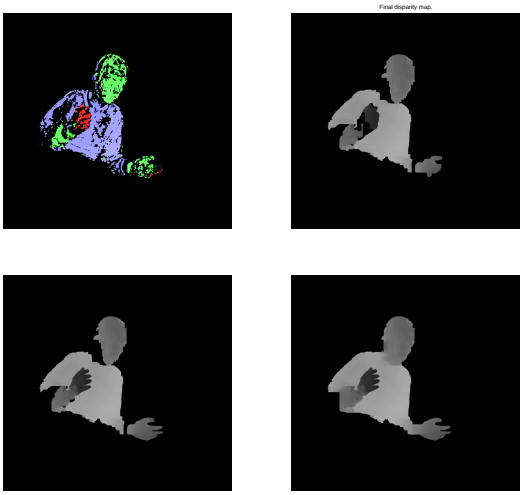

Figure 5. Top left: disjoint depth regions of the disparity map. Top right: result of spline fitting within the depth regions. Bottom left: completing the hands. Bottom right: fitting the splines across the depth regions.

straightforward linear interpolation (Figure 7 top right), and the one generated by our algorithm (Figure 7 bottom). For view synthesis, we used our own version of the trilinear reprojection algorithm [**?**]. The unprocessed disparities are obviously unusable. The linear interpolation yields a plausible result, but containing a serious amount of artifacts concentrated around depth region boundaries. Using disparities generated by our algorithm, artifacts have been very much reduced, as boundaries between depth regions are clearer and depth variations within regions are limited by spline fitting, bringing an apparent, significant improvement. The quality of our result must be judged considering the severely incomplete disparity maps which are input to the algorithm, and that only limited assumptions on the scene are used.

4 Conclusions

We have presented an algorithm generating dense disparity maps from severely sparse disparity data. We recover homogeneous depth regions from the video frames and the

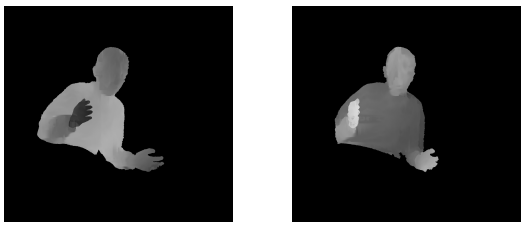

Figure 6. Final post-processed left and right disparity maps.

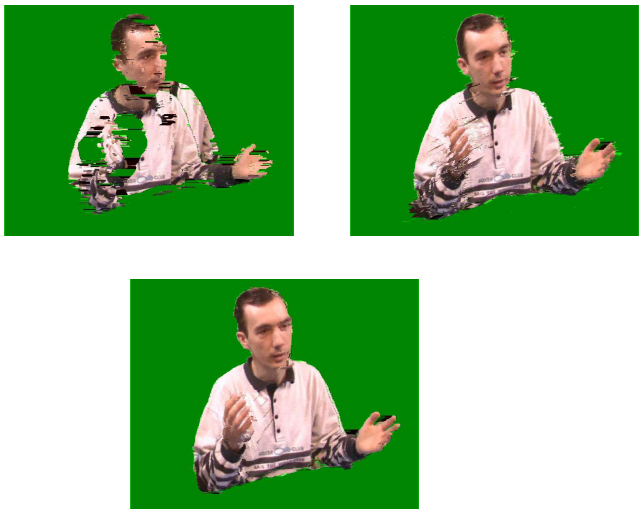

Figure 7. Top left: synthetic view using the original, non-processed disparity map. Top right: synthetic view using the linearly interpolated disparity map. Bottom: synthetic view using the algorithm post-processed disparity map.

sparse depth data, and use them to force depth interpolation only with consistent data. We use spline fitting which guarantees that interpolated depths vary gently within regions.

Several research avenues are open for future work, including achieving real-time, good quality disparities with wide-baseline stereo, and using full body models to guide discontinuity detection and replace disjoint, local continuity assumptions.

Acknowledgments

Thanks to Francesco Isgrò, Mino Bernardi, Oliver Schreer, Ralph Martin, and Roger Hubbold for various support and discussions. The VIRTUE data were acquired by Emile Hendriks's group at the Delft University of Technology.

References

[1] *www.advanced.org/teleimmersion.html*

[2] Virtue homepage: *www.virtue.eu.com*

[3] H. H. Baker, D. Tanguay, I. Sobel, D. Gelb, M. E. Goss, W. B. Culbertson, T. Malzbender. "The Coliseum Immersive Teleconferencing System", Hewlett-Packard Laboratories Technical Report HPL-2002-351, December 2002.

[4] A Criminisi, J Shotton, A Blake, P H S Torr, "Gaze manipulation for one-to-one teleconferencing", Proc IEEE Int Conf on Comp Vision, Nice, 2003.

[5] N. L. Chang and A. Zakhor. A Multivalued Representation for View Synthesis. In *Proceedings of ICIP'99*, October 1999.

[6] D. Chetverikov, J. Matas, "Periodic Textures as Distinguished Regions for Wide-Baseline Stereo Correspondence", *Texture02*, pp. 25-30.

[7] O. Schreer, N. Brandenburg, S. Askar, P. Kauff. "Hybrid Recursive Matching and Segmentation-Based Postprocessing in Real-Time Immersive Video Conferencing" In *Proceedings of VMV 2001*, 2001.

[8] E. Trucco, C. Plakas, N. Brandenburg, P. Kauff, M. Karl, O. Schreer, "Real-Time Disparity Analysis for Immersive 3-D Teleconferencing by Hybrid Recursive Matching and Census Transform", Proc. IEEE ICCV Workshop on Video Registration, Vancouver, 2001.

[9] F Isgrò, E Trucco and L-Q Xu, "Towards teleconferencing by view synthesis and large-baseline stereo", Proc IAPR/IEEE Int Conf on Image Analysis and Processing, Palermo (I), 2001, pp. 198-203.

[10] C. Strecha, T. Tuytelaars, L. van Gool, "Dense matching of multiple wide-baseline views", *Proc IEEE Int Conf on Comp Vision*, Nice, 2003.

[11] C Strecha, T Tuytelaars, L van Gool, "Dense matching of multiple wide-baseline views", Proc IEEE Int Conf on Comp Vision, Nice, 2003.