

NEW CONCEPT FOR JOINT DISPARITY ESTIMATION AND SEGMENTATION FOR REAL-TIME VIDEO PROCESSING

Nicole Atzpadin¹, Serap Askar, Peter Kauff, Oliver Schreer

Fraunhofer Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, Einsteinufer 37, D-10587 Berlin
Phone: ++49 30 31002 {-660, -610, -615, -620}, fax: ++49 30 3927200
Email: {nicole.atzpadin, serap.askar, peter.kauff, oliver.schreer}@hhi.fhg.de

ABSTRACT

We present a new concept of joint disparity estimation and segmentation developed for a real-time immersive video conferencing system, which uses segmentation and disparity estimation results to calculate a new virtual view of a conferee. For the segmentation background subtraction is used which works well under limited conditions. The main goal of the new concept is to allow more general scenarios and to improve the disparities at depth discontinuities. The goal is reached with a two-stage disparity estimation which is nearly a hierarchical approach. The main advantage of the two-stage calculation is that segmentation can be performed in between these two steps and can use depth information to decide on foreground or background and the second step of disparity estimation can use segmentation information to speed up the algorithm.

1. INTRODUCTION

The concept for joint disparity estimation and segmentation for real-time video processing presented in this paper was developed in the context of an immersive teleconferencing system. An immersive teleconferencing system enables conferees located in different geographical places to meet around a virtual table, appearing at each station in such a way to create a convincing impression of presence [1]. The purpose is to enable the participants to make use of rich communication modalities as similar as possible to those used in a face-to-face meeting (e.g., gestures, eye contact, etc) and eliminate the limits of non-immersive teleconferencing, (e.g., face-only images in separate windows, unrealistic avatars).

To realize an immersive teleconferencing system a multi-view camera set-up capturing the conferees is mounted around a large 2D display. A full 3D analysis delivers the information needed to render the remote conferees on the display of the local conferee adapted to the position of his head. To achieve realistic viewing conditions the conferees are shown in a shared virtual environment which means that they have to be segmented from the real background scene with a fast and robust foreground-background segmentation tool which is based on a background subtraction. The disparities, which represent the depth of the whole scene, are calculated between two stereo camera pairs with an efficient algorithm which is based on hybrid recursive matching (HRM) - an universal approach on real-time analysis of displacement vector fields. Figure 1 gives an overview of the 3D analysis of one stereo pair.

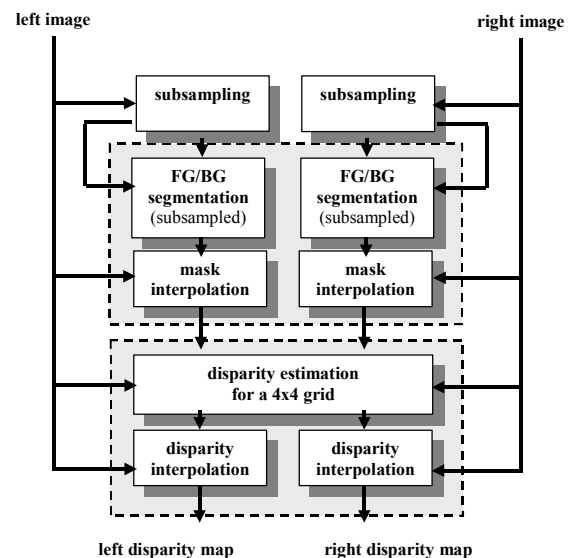


Fig. 1 : overview of the 3D analysis for one stereo pair

¹ nee Nicole Brandenburg

One of the most important modules is the foreground-background segmentation, which separates the foreground object from the background. The process is essential for the final view rendering, which integrates conferees together in a shared virtual environment. A further advantage is that the foreground object only covers half of the pixel of the whole frame which leads to less computational complexity in the subsequent processing steps. To achieve real-time the pixel segmentation itself is performed on sub-sampled images. The resulting binary mask marking each pixel as foreground or background is interpolated to full image size by a sophisticated refinement scheme[2].

The overall goal of the whole 3D analysis is to estimate correct dense disparity maps from the left to right and right to left camera input image. To reach this goal in real-time disparities are estimated for a sparse grid of 4x4, post-processed and then interpolated to dense disparity maps. The disparity algorithm is a hybrid recursive matching (HRM) approach which unites the advantages of block-recursive matching and pixel-recursive optical flow estimation in one common scheme. Its computational effort is minimised by the efficient selection of a small number of candidate vectors, guaranteeing both spatial and temporal consistency of disparities.

2. NEW CONCEPT

The whole 3D analysis system introduced in the last chapter works well under ideal conditions. Ideal conditions exist if the background is static without any moving objects and foreground object and background are not of similar colour. But a videoconferencing system cannot always provide ideal conditions so that other information has to be included into the decision process of the segmentation to allow a more general scenario. Another problem with the existing system is the non-consideration of the window size problem in the disparity estimation as mentioned in [3]. The intensity window must be large enough to include enough intensity variation for matching but small enough to avoid the effects of projective distortion. A lot of solutions were proposed in the last years, one of them is a hierarchical approach [4], where disparities are estimated from coarse to fine, starting with large windows, which become smaller from one step to the next.

To overcome the problems of the existing approach we recommend a system with joint disparity estimation and segmentation as it is shown in Fig. 2. The disparity

estimation is split into two parts. In the first part disparities are estimated for the foreground and background in the sub-sampled images. In the second part the disparities are improved only for the foreground object. The splitting of disparity estimation has the advantage that disparities and therefore depth information is available for the foreground-background segmentation. As a by-product the disparity estimation can solve the window size problem as it can use different window sizes in the two steps.

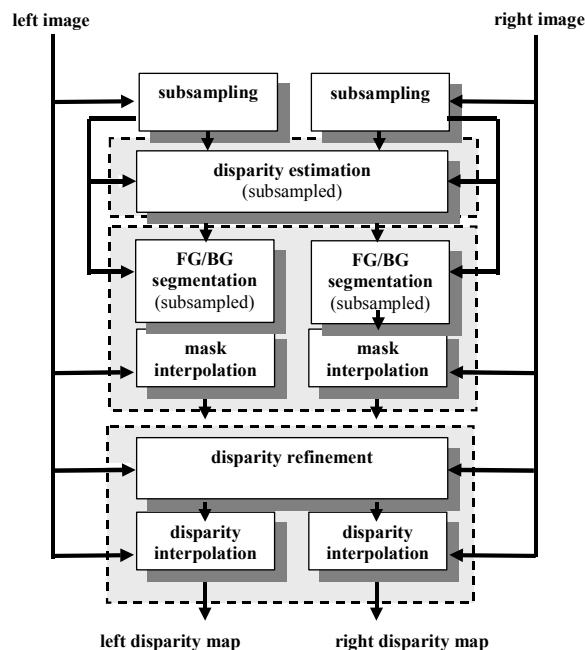


Fig. 2 : joint disparity estimation and segmentation

The different processing stages will now be described in more detail.

2. HYBRID RECURSIVE MATCHING

The hybrid block- and pixel-recursive disparity estimator unites the advantages of block-recursive disparity matching and pixel-recursive optical flow estimation in one common scheme[5].

The main idea of hybrid block- and pixel-recursive matching is to use neighbouring spatio-temporal candidates as input for the block-recursive disparity estimation. The rationale is that such candidate vectors are the most likely to provide a good estimate of the disparity for the current pixel. In addition, a further update vector is tested against the best candidate. This

update vector is computed by applying a local, pixel-recursive process to the current block, which uses the best candidate of block-recursion as a start vector.

The whole algorithm can be divided into three stages:

1. three candidate vectors (two spatial and one temporal) are evaluated for the current block position by recursive block matching;
2. the candidate vector with the best result is chosen as the start vector for the pixel-recursive algorithm, which yields an update vector;
3. the final vector is obtained by comparing the update vector from the pixel recursive stage with the start vector from the block-recursive one.

Only three candidates are tested in the recursive block matching to find the best match between the current pixel position in the left image and the related pixel position in the right image. A shape-driven displaced block difference (DBD) is taken as criterion. Large window sizes are required in order to estimate high reliable disparities which can be inaccurate due to projective distortions but not completely wrong as it could happen with smaller window sizes. Unfortunately large window sizes are connected to high computational complexity. A sub-sampling of the original images avoids this extra costs, because here also small window sizes lead to reliable results.

The hybrid analysis scheme has two main advantages in comparison to common approaches. The recursive structure speeds up the analysis dramatically. The combined choice of spatial and temporal candidates yields reliable disparity vectors and spatially and temporally consistent disparity fields due to an efficient strategy of testing particular vector candidates. The latter aspect is important to avoid temporal inconsistencies in disparity sequences, which may cause strongly visible and very annoying artefacts in virtual views synthesised on the basis of these disparities.

As a post-processing step an efficient consistency check comparing the disparities from the left to right and right to left image discards all inconsistent disparities.

3. FG-BG SEGMENTATION

The segmentation algorithm is based on a background subtraction, comparing the current image of a particular conference with a pre-known background reference image (see Fig. 3).

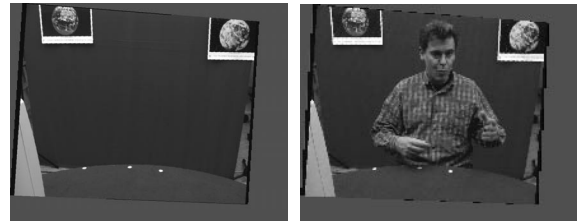


Fig. 3 : pre-known background reference image and current image

The structure of the segmentation module is shown in Fig. 4. The background reference buffer is captured during an initialisation phase at the beginning of the conference session where a short background sequence is analysed. This provides further statistical information about noise. From this a pixel wise threshold buffer is derived, which is used for segmentation. During the conference session global illumination changes are compensated with an adaptive offset buffer.

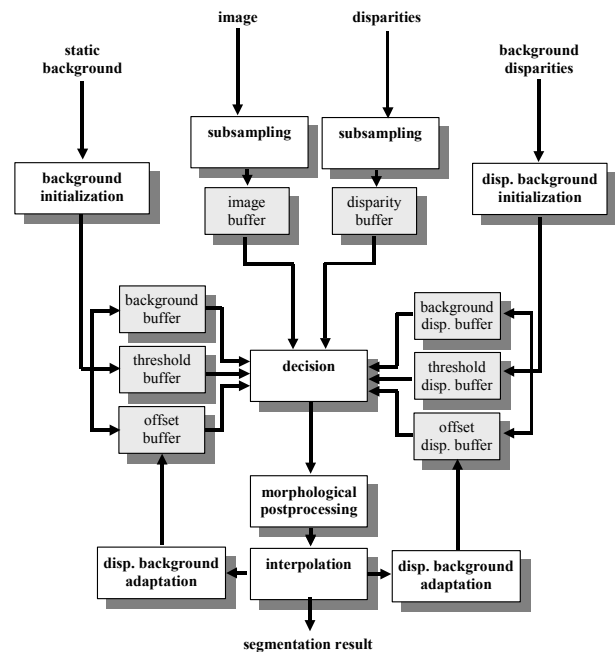


Fig. 4 : outline of the foreground-background segmentation

The background subtraction for colour images is enlarged to a background subtraction for colour and depth images. Disparities are estimated for the pre-known reference images in the initialisation phase and compared with the current disparities. Fig. 5 shows examples of consistent disparities. A background adaptation is also reasonable for disparities because it allows little movements in the background.

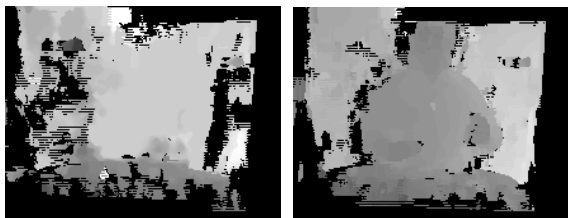


Fig. 5 : pre-calculated background disparities and current disparities (inconsistent disparities are black)

The differences of the three colour channels y , u , and v and the disparity difference are evaluated in the decision process to find a final result. The disparity difference is stronger weighted than one of the colour channels. If either the disparity of the background or the foreground are inconsistent only colour information is used for the decision process.

The result of segmentation is a binary mask marking each pixel as foreground (i.e. object) or background. Because this mask is usually corrupted by noise different kinds of well-known non-linear and morphological post-processing algorithms are applied to smooth the binary mask and get one closed object region, which represents the conferee. Fig. 6 shows results for the segmentation using only colour and for the segmentation using colour and depth without any post-processing.

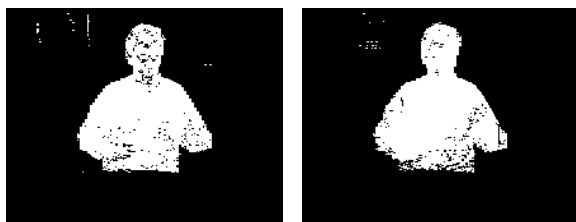


Fig. 6 : segmentation results with colour (left) and with colour and depth (right)

The results only using colour are quite good in this case due to the nearly ideal scenario. Using depth gives an improvement in parts of the image where slight

movements of the background lead to wrong segmented pixel especially around the image on the wall.

To speed up the process the segmentation is performed on sub-sampled images. Therefore a combination of interpolation and refinement is used to up-sample the final segmentation mask back to full resolution. This process is performed block wise. In homogenous areas, i.e. completely inside or outside the object region, the binary mask can be interpolated by simply copying the known pixels. At the transition regions segmentation must be particularly repeated in full resolution to preserve the contour as much as possible. This is achieved with a hierarchical filling [2]. Finally, based on the result of refinement the threshold offset buffer is updated.

4. DISPARITY REFINEMENT

A refinement of disparities resulting from the first step of disparity estimation is unavoidable due to the low resolution of disparities produced by the sub-sampling. This means it is not sufficient to multiply all disparities with the sub-sampling factor. The refined disparities have to be estimated defining a search area around the position determined by the disparity multiplied with the sub-sampling factor. Small window sizes are used for the comparison of intensities to exactly find disparity borders which coincide with the object boundaries. Fig. 7 shows results using a search window of size 8×8 pixel.

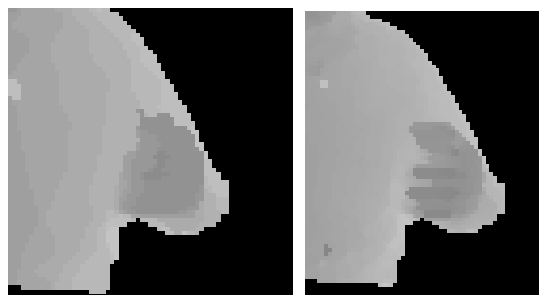


Fig. 7: sub-sampled disparities (left) and refined disparities (right)

The improvement around depth discontinuities can clearly be recognized. More details of the hands become visible. This is achieved by two search ranges in the two depth layers around the border. The disparity refinement has only to be performed inside the foreground object. To save computational costs disparities are only refined for pixel on a 4×4 grid. The remaining disparities are

interpolated with a sophisticated interpolation scheme introduced in [2].

5. CONCLUSION

In this paper we have introduced a new concept on joint disparity estimation and segmentation, which stabilizes the segmentation and extends it to a more general scenario with small movements in the background and similar colour in the foreground and background. Additionally the disparity estimation is improved especially at depth discontinuities due to a two-stage hierarchical approach refining disparities from coarse to fine. The main advantage of the new concept is that both algorithms use the results of the other algorithm. Disparity estimation needs less computational complexity in the second step because it has only to be performed inside the foreground object and the segmentation allows more generality in the scenario because of the used depth information.

6. ACKNOWLEDGEMENTS

This study is supported by the Ministry of Science and Technology of the Federal Republic of Germany, Grant-No.01 AK 022.

7. REFERENCES

- [1] P. Kauff, O.Schreer: "Virtual Team User Environments - A Step From Tele-Cubicles Towards Distributed Tele-Colaboration in Mediated Workspaces", *Int. IEEE Conf. on Multimedia and Expo (ICME 2002)*, Lausanne, Switzerland, August 2002
- [2] S. Askar, P. Kauff, N. Brandenburg, O. Schreer: "Fast Adaptive Upscaling of Low Structured Images Using a Hierarchical Filling Strategy", *Proc. of VIPromCom 2002, 4th EURASIP*
- [3] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 920-932, September 1994
- [4] J.-I. Park and S.Inoue. Hierarchical depth mapping from multiple cameras. In *Proceedings of International Conference on Image Analysis and Processing (ICIAP'97)*, pages 685-692, Florence, Italy, September 1997
- [5] P. Kauff, N. Brandenburg, M. Karl, O. Schreer: "Fast Hybrid Block- and Pixel-Recursive Disparity Analysis for Real-Time Applications in Immersive Tele-Conference Scenarios", *Proc. of WSCG 2001, 9th Int. Conf. on Computer Graphics, Visualization and Computer Vision*, Pilzen, Czech Republic, February 2001