MULTI-VIEW DEPTH ESTIMATION BASED ON VISUAL-HULL ENHANCED HYBRID RECURSIVE MATCHING FOR 3D VIDEO CONFERENCE SYSTEMS

I. Feldmann[†], *N. Atzpadin*[†], *O. Schreer*[†], *J.-C. Pujol-Acolado*[*], *J.-L. Landabaso*[*], *O. Divorra Escoda*[*]

[†] Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut, Berlin, Germany, http://ip.hhi.de
[*]Telefonica Research, Barcelona, Spain, http://www.tid.es

ABSTRACT

This paper discusses the problem of high quality depth map estimation for real-time systems. Our work is based on the European FP7 project 3DPresence which aims to build a multi-view and multi-user 3D videoconferencing system. Based on new multi-view autostereoscopic display technology the remote conferees will be rendered as an integral part of a three dimensional virtual shared environment. In order to create the related views for the 3D displays as well as to virtually correct the eye contact problem robust depth maps are required. For this purpose, in this paper we will discuss the fusion of two competing approaches which have, from a camera configuration point of view, contrary to each other properties. Namely, we will combine the volumetric Visual Hull (VH) approach with the stereo matching based Hybrid Recursive Matching (HRM) to a new method which benefits from the advantages of both techniques and discards their weak points.

Index Terms— Depth estimation, Visual-Hull, HRM, stereo-matching, real-time

1. INTRODUCTION

The target application of this paper are immersive 3D videoconferencing systems. In difference to traditional high-end commercial 2D solutions, such as Ciscos TelePresence, Polycoms TPX, and HPs HALO, in a 3D tele-conferencing system the conferees are cut out of the real scene and virtually placed into a 3D shared table environment together with the remote conferees (see fig. 1). The advantage of these solutions is that eye contact and gesture awareness can be created by adapting virtually the 3D perspective and 3D position of all remote conferees on each of the terminal displays.

From a research point of view the major challenge of such systems is the generation of real-time high quality and high resolution depth maps or 3D models of the captured scene and the conferees [1]. Recent approaches use usually highly optimized disparity estimation techniques based on stereo block matching [2]. A more detailed review on state of the art stereo correspondence algorithms is provided in [3]. Nevertheless, in the past years an increasing number of approaches can be found which solve the problem of real-time 3D scene reconstruction by employing volumetric reconstruction techniques such as [4, 5, 6, 7].

However, both general approaches have major drawbacks. On one hand, for stereo correspondence search the depth reconstruction quality is strongly limited by the camera base-line of the system. Large camera distances will introduce artifacts, caused by occlusions and disocclusions [1]. On the other hand, volumetric reconstruction systems offer an increasing scene reconstruction quality with increasing camera distances. Vice versa, small camera baselines lead in this case to a low reconstruction quality. Further on, a major drawback of volumetric silhouette based solutions is their dependency on the object shape. In general, concave 3D object regions cannot be detected.

Fig. 1. Example for a shared virtual 3D video conference environment as proposed by the European research project 3DPresence.

The main idea of this paper is to combine the advantages of both general approaches into one common system. In order to extract the depth and 3D structure of the scene a first set of images from different positions is processed to obtain a 3D volume estimate known as Visual Hull (VH). This volumetric information is converted to a depth map as if it had been seen from a particular point in space. A second set of camera images is used to obtain the depth of the scene based on stereo-matching techniques. The novelty of our method is, that the depth maps of the first VH based step will be implicitly integrated into the correspondence search of the Hybrid Recursive Matching (HRM) module. In this way, a robust reconstruction can be achieved which combines the advantages of both methods.

The paper is organized as follows. The next two sections are devoted to the algorithmic details of the volumetric-based approach and the plain hybrid recursive matching. Section 4 describes the proposed combination of both techniques. Section 5 presents experimental results. The paper concludes in section 6 with a discussion of future research directions.

2. VISUAL HULL BASED DEPTH ESTIMATION

A fast approximation of the real depth for several views is needed. Our idea is to obtain an approximation of the real volume and then

This work has been performed within the framework of EU FP7 Project 3DPresence.

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

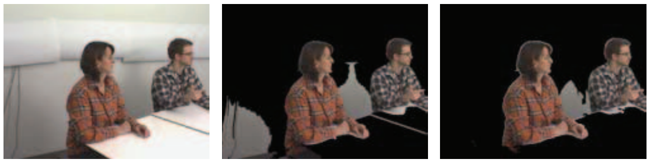

Fig. 2. Segmentation results for a specific view, **left)** original image, **middle)** single view segmentation, **right)** segmentation after the VH projection. The background pixels are plotted in black whereas the foreground pixels are displayed in their original color.

use it to calculate the depth from a desired view position. In order to do so, a voxel-based Shape from Silhouette technique (SfS) is chosen due to simplicity [8, 5]. The obtained volume by this algorithm is the VH, which is the maximum volume consistent with a set of calibrated input silhouettes. The basic SfS technique establishes a bounding voxel-based volume around the space to be reconstructed. Then, every voxel is labeled as "occupied" if the pixel of all images where it projects to is part of the foreground. Otherwise, it is labeled as "empty". Since our main objective is not to obtain a perfect 3D reconstruction but a fast approximation, SfS fits our needs.

The silhouettes are obtained using the technique described in [9]. It uses a probabilistic Gaussian model per pixel to describe the background, learned from a training sequence. For new incoming images, every single pixel is first compared to the Gaussian model and classified as "foreground" or "background". If it is labeled as "foreground", it will be applied further on to a shadow module detector which is based on color vector alignment to the background model. Afterwards, a global operator is used to remove *salt and pepper* noise and to fill holes. We tune the thresholds for this algorithm so that false negatives are minimized, The SfS technique will be set up accordingly. The results after the segmentation are shown in fig. 2 (middle). Note that the "shadow" pixels have been included into the "background" pixels. In this particular example, shadows are not removed satisfactory since the *no false negatives* condition leads to many false positives.

When all the images have been segmented, the SfS reconstructs the volume according to them. Since we made sure that no false negatives will exist in the segmentation, the "survival voxel test" is based the voxel projection to image "foreground" in all of the input images. This simple test condition leads to a fast implementation. The resulting volume can handle properly most of the false positives, such as strong shadows, as they are not labeled as "foreground" in all of the silhouettes. In this way, the volume projection to the image plain can be used as a segmentation refinement which especially improves the robustness of shadowed regions significantly as illustrated in fig. 2 right.

An example of reconstructed volume is plotted in fig. 3 (left) and the corresponding depth map can be found on the right-hand side of the figure. The figure shows that a high quality 3D reconstruction of the conferees could be obtained. Compared to stereo matching approaches the depth resolution is very high. Further on, the scene is reconstructed in high detail. For example, the fingers of the conferees are estimated correctly and in good quality.

Nevertheless, unwanted "ghost" artifacts may appear in the result as illustrated in fig. 3 in front of the left-hand person. These artifacts are based on one hand on a wrong number or wrong geometrical configuration of the cameras. On the other hand, segmentation failures, as illustrated in fig. 2 right (remaining shadow artifacts in the background), may cause those artifacts. Further on, as mentioned beforehand, concave regions cannot be detected by the current approach. Due to these limitations, a depth refinement step will be introduced which is based on a stereo matching method, namely the Hybrid Recursive Matching (HRM). It will be introduced in the next sections.

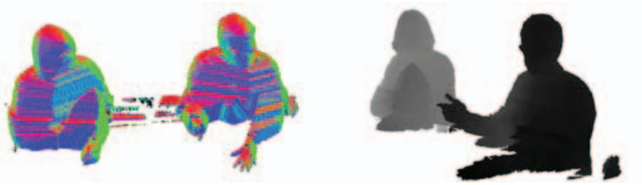

Fig. 3. **left)** Example for a reconstructed volume for a voxel size of 5^3mm. Nine cameras were used, **right)** depth map generated from the reconstructed volume using the SfS algorithm

3. HYBRID RECURSIVE MATCHING

The estimation of suitable depth maps from stereo or multi-view camera systems is certainly one of the most challenging tasks in the given context. The disparity estimation itself is based on the Hybrid-Recursive-Matching (HRM) algorithm as described in [2]. The main idea of the hybrid recursive stereo matching algorithm is to unite the advantages of block-recursive disparity matching and pixel-recursive optical flow estimation in one common scheme. The block-recursive part assumes that depth does not change significantly from one image to the next and that depth is nearly the same in the local neighborhood. Obviously this assumption cannot be fulfilled in all image areas - especially not in areas with high motion and at depth discontinuities. To update the results of the block-recursive stage in these areas, the pixel recursive stage calculates the optical flow by analyzing gradients and gray value differences. In more detail, the structure of the whole algorithm can be outlined in three subsequent processing steps (see fig. 4):

1. Three or four candidate vectors 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 testing if the update vector from the pixel recursive stage is of higher quality than the start vector from the block-recursive one.

The idea of the block recursion is to use information of both, the previous image and the spatial neighborhood. This kind of recursion forces temporal and spatial consistency and it additionally reduces the local search range to a few pixel as known from many other matching algorithms. Usually, calculation of three matching scores per considered pixel is fully sufficient to achieve results comparable to a full search method. The pixel-recursive stage is a low-complexity method, which calculates dense displacement fields using a simplified optical flow approach. Following the principal of the optical flow, an update vector is calculated on the basis of spatial gradients and gradients between the two frames. Due to its recursive structure the HRM algorithm produces extremely smooth and temporally consistent "per-pixel" disparity maps. Hence, they contain highly redundant information and have almost no random noise

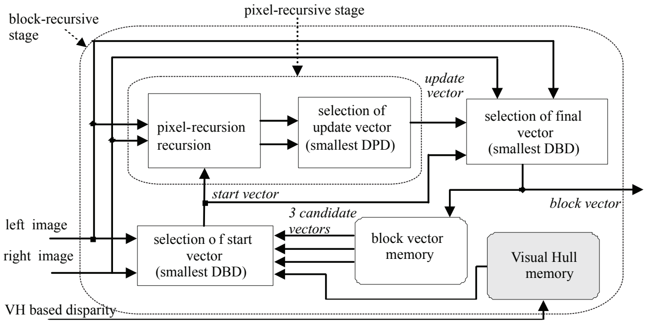

Fig. 4. Outline of the HRM algorithm

- a property that is essential for efficient coding of depth maps. As any matching algorithm, HRM usually generates failures and mismatches in critical image areas. These mismatches are detected and corrected by sophisticated post-processing.

4. AN IMPROVED DEPTH ESTIMATION ALGORITHM: VISUAL-HULL ENHANCED HRM

In our set-up, the depth maps obtained with the HRM technique are usually more robust than those ones which can be extracted from the VH. However, there are some particular situations where HRM cannot provide the best possible quality. Since the system has been designed to operate in real-time, our particular implementation of HRM makes an intensive use of the depth information obtained from previous frames. Therefore, there are some inaccuracies in those regions corresponding to fast moving parts. In contrast, the quality that the VH provides is invariant to fast moving regions. This fact is used to fuse the depth maps derived from the VH under fast motion situations. As mentioned before, one problem of the VH approach is that it suffers from the ”ghost” artifact problem. Further, SfS based VH cannot resolve concave scene object shapes. On the other hand, the HRM is limited to relatively small baselines. Further, homogeneous or periodic textures in the image may cause artifacts in the reconstruction result.

Due to the completely different approaches to calculate depth information, HRM and VH produce depth failures in different image regions. To exploit the advantages of both approaches, the HRM uses the results of VH to enhance the quality mainly in regions of fast motion and in occluded regions. In the original HRM algorithm three candidates are evaluated, which are defined by disparities from the previous and the current image. The HRM with VH also tests a disparity value defined by VH. The following spatial, temporal and external candidates are tested for this purpose:

- A horizontal predecessor, taken from the left or right position in the actual frame.
- A vertical predecessor, taken from the bottom or top position in the actual frame.
- A temporal predecessor, taken from the same position in the previous frame.
- An external disparity, taken from the same position in the disparity map derived from VH

Although the quality of disparities is enhanced by the combination of the two approaches mismatches occur. Usually, there are two reasons for the detected mismatches: ambiguities during matching (homogeneities, similarities, periodicities, etc.) or occluded areas. These two failure categories have completely different origins. Ambiguities are caused by an ill-posed matching problem; i.e., point-correspondences exist but could not be found correctly by the matcher. In contrast, point correspondences do not exist at all in occluded areas and cannot be matched on principle therefore. Due to the different perspectives used by VH, regions which are detected as occluded in HRM with VH are directly substituted by the VH results.

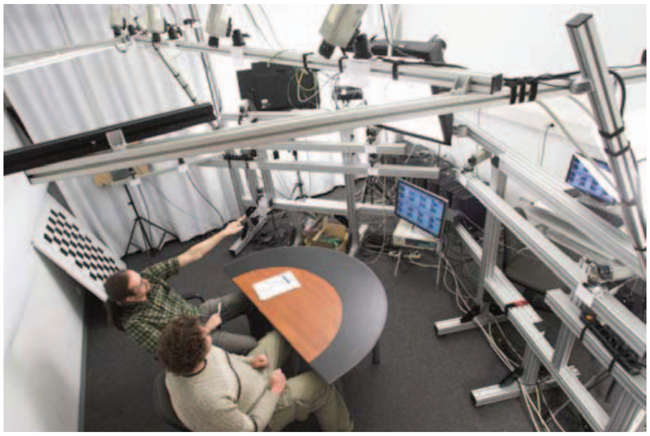

Fig. 5. Experimental setup for combined Visual Hull and HRM based analysis system in the context of 3D video conferencing

HRM uses VH depth as additional candidate for the block matching as well as for the post-processing. If additional disparity information is provided by VH the disparities have to be adapted due to the fact that VH is not able to handle concave objects. One possibility is to define a small search range around the disparity delivered by the short baseline system and to find the best match in this search area. The other possibility is to perform pixel recursion with the external disparity from VH as start vector. In practice the definition of a small search range and a search with a smaller search window showed the best compromise between quality and time for calculation. Experiments have shown that for the actual camera setup the search range is not necessary. In post-processing especially occlusions are substituted by the VH disparities.

5. EXPERIMENTAL EVALUATION

Our SfS implementation uses both CPU and GPU facilities. Look Up Tables (LUT) are used to store, for every voxel, the memory address of the pixels of the input images where it projects to. Once the data has been fetched, the survival voxel test condition is executed. Thus, our current bottleneck is set by the memory transfer speed. In order to obtain the depth map, we use OpenGL to set the specified virtual view and then render the voxels. Afterwards, the frame buffer’s depth map is read and converted to disparity. The 3D space is set to 224 × 448 × 175 voxels, that produces a 5^3mm size voxel. Although this particular configuration is not suitable for real-time, a future *all GPU* implementation will achieve a proper frame rate. The HRM is implemented on CPU only. We use an input image resolution of 1024x768 pixels. The input images are rectified in order to simplify the correspondence search.

Our experimental test setup is illustrated in fig. 5. Note, that

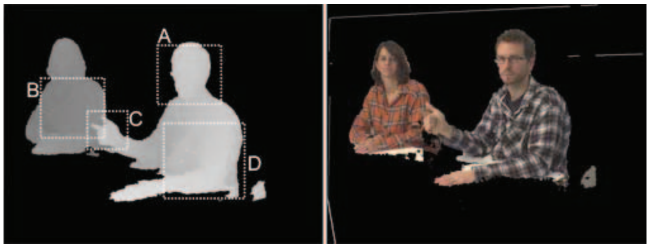

Fig. 6. **left)** Estimated depth map of the conferees based on the combined HRM and VH approach, **right**), rendered view to be integrated into the shared virtual conference environment

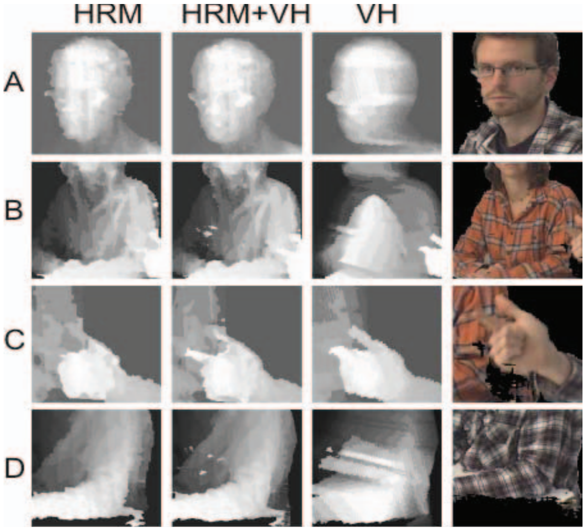

Fig. 7. Comparison of depth map estimation results for the separate and combined HRM and VH approaches. Critical regions are chosen based on the highlighted boxes in fig. 6. Note, that the depth maps were scaled in order to enhance visibility.

the final system will contain four 42 inch multi-view 3D displays. The test setup simulates the geometrical equivalents of the virtual shared table environment as illustrated in fig. 1. Four trifocal camera systems are placed on the top of each display. Four additional VH cameras are situated on the top and the sides of the scene. Further, each trifocal camera system provides one additional input for the VH module. This gives in total eight cameras available for VH.

Fig. 6 shows an example for the estimated depth map (left) of the combined HRM and VH approach. The right-hand figure illustrates the rendered view. Note, that the perspective of the view was adapted to the geometry of the shared virtual table environment of the tele-conference system. In this way, the eye contact between the conferees can be guaranteed. A comparison between the pure HRM based depth estimation and the pure VH based depth estimation on one hand and the proposed combined approach is shown in fig. 7. Several regions have been extracted and highlighted in order to illustrate the outcomes. In row A can be clearly seen, that the achieved depth resolution of the pure VH approach is much higher than for the pure HRM approach. The combined version gives here an acceptable compromise. Nevertheless, the pure VH suffers from 'ghost' artifacts as illustrated in rows B and D. Here, the HRM clearly outperforms the VH. A combined approach enhances the quality significantly. Further, an improvement in reconstruction quality can be achieved for scene details, such as the fingers in row C.

6. CONCLUSIONS

We have discussed two competing methods to estimate the 3D structure of a scene, namely the structure from silhouette based visual hull and the stereo based hybrid recursive matching. Both approaches have strengths and weaknesses in different situations. We have shown that it is possible to combine the two methods in order to improve the accuracy and robustness of the overall result. In this way it was possible to benefit from the advantages of both methods while discarding their weak points. Our idea was proved for critical situations of the proposed real-time 3D video conferencing system.

7. ACKNOWLEDGMENTS

This work is part of the FP7 project ”3DPresence”, Proposal no.: FP7-215269, which is funded by the European Commission.

8. REFERENCES

[1] O. Schreer, I. Feldmann, W. Waizenegger, N. Atzpadin, P. Eisert, and H. Belt, “3dpresence a system concept for multi-user and multi-party immersive 3d videoconferencing,” in *5th European Conf. on Visual Media Production (CVMP 2008)*, 2008, pp. 321–334.

[2] N. Atzpadin, P. Kauff, and O. Schreer, “Stereo analysis by hybrid recursive matching for real-time immersive video conferencing,” in *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications, Vol. 14, No. 3*, 2004, pp. 321–334.

[3] Daniel Scharstein and Richard Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[4] Kiriakos N. Kutulakos and Steven M. Seitz, “A theory of shape by space carving,” *International Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, 2000.

[5] A. Laurentini, “The Visual Hull: A new tool for contour-based image understanding,” in *Proceedings of Seventh Scandinavian Comperence on Image Processing*, 1991, pp. 993–1002.

[6] J. L. Landabaso and M. Pardàs, “Foreground regions extraction and characterization towards real-time object tracking,” in *Proceedings of Multimodal Interaction and Related Machine Learning Algorithms*. 2005, Lecture Notes in Computer Science, Springer.

[7] A. Hilton J. Starck, “Model-based multiple view reconstruction of people,” in *Proceedings of IEEE International Conference on Computer Vision*, 2003, pp. 915–922.

[8] A. Laurentini, “The visual hull concept for silhouette-based image understanding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 150–162, 1994.

[9] L.-Q. Xu, J. L. Landabaso, and M. Pardàs, “Shadow removal with blob-based morphological reconstruction for error correction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, March 2005, IEEE Computer Society.