SCENE FLOW CONSTRAINED MULTI-PRIOR PATCH-SWEEPING FOR REAL-TIME UPPER BODY 3D RECONSTRUCTION

W. Waizenegger[1,2]*, I. Feldmann*[1]*, O. Schreer*[1] *and P. Eisert*[1,2]

[1]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany
[2]Humboldt University, Berlin, Germany

ABSTRACT

We present a high performance real-time approach for robust 3D structure estimation of human bodies. Our idea extends state of the art high precision patch-sweep techniques by exploiting temporal information. By parameterizing 3D patches in the spatio-temporal domain we gain increased quality and robustness, while the computational complexity decreases drastically. The key to these improvements is the utilization of a global scene flow like prior called *spatio-temporal object* (*STO*). The target application of our method reaches from video communication and virtual eye contact scenarios till future digital cinema 3D post-production and real-time person modelling and modification. The result of this work extends a novel 3D scene representation developed in the EC research project SCENE.

Index Terms— scene flow, real-time, 3D reconstruction, upper body

1. INTRODUCTION

Realistic rendering of persons in other perspectives than captured by available cameras is still a challenging research topic. A quite prominent application is video communication, where the eye contact problem is tackled by 3D scene reconstruction and virtual placement of a camera in the display. In current video communication systems, either high-end commercial tele-presence systems by Cisco, HP or web application like Skype video, the cameras are mounted on top of the display, while the user is looking onto the display at the remote participant. This difference in viewing angle leads to the fundamental problem of missing eye contact, which is considered as a major drawback of current video communication. Several attempts have been made to face this problem e.g. [1][2], but the rendering quality did not gain user acceptance. An improved solution has been proposed in [3], but the focus was on realistic rendering of the face region. Rendering of upper body representations of human beings is not only relevant for video communication, but also in other fields of application. New advances in computing technology as well as algorithm development offer the opportunity to achieve real-time capability and the necessary rendering quality. Therefore, chat applications or social media services, which currently represent users by artificial avatars, request for more natural appearance and realistic human body motion [4]. In the context of news broadcast, concepts like virtual studios are investigated, where the journalist is not anymore overlaid in a 2D scene, but fully integrated in the scene based on 3D scene analysis and realistic 3D rendering and compositing of the journalist itself [5]. New concepts of media walls either in 2D or 3D are discussed offering a completely new interface at home or in the office supporting natural interaction and communication with remote people, friends, family members or even business colleagues. Due to recent advances in depth estimation and 3D reconstruction, real-time view rendering methods have been developed [6][7]. However, there are still restrictions on the scenario such as single person and still background and challenges such as high-fidelity reconstruction, wide field of view analysis and complex 3D scene structures. In this paper, we propose a combined image and model-based approach, whereby the focus is on spatio-temporal consistency of reconstructed objects. Flickering and inconsistency of object shapes over time are the most recognizable and disturbing artifacts. The model definition used herein is a flexible and generalized approach consisting on texture, depth and confidence as major properties, the so-called *spatio-temporal object* (*STO*) as illustrated in figure 1.

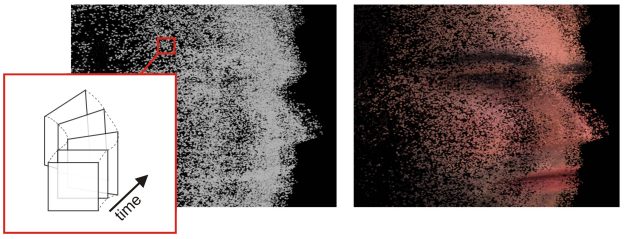


Fig. 1. Joint parameterization of 3D patches in space and time leads to spatio-temporal objects.

The paper is organized as follows. Section 2 outlines an algorithmic overview for the proposed approach. In section 3 we review the basic idea of constrained patch sweeping and refer to state of the art. In section 4 *spatio-temporal objects* are conceptually introduced and a concrete numerical procedure for their approximation is discussed. Section 5 describes

our novel multi-prior extension for Patch-Sweep and the integration of *spatio-temporal objects* as additional priors. The quantitative and qualitative evaluation of our approach is discussed in section 6.

2. ALGORITHMIC OVERVIEW

The proposed algorithmic scheme consists of three major real-time components. A *Hybrid Recursive Matching (HRM)* module, which is a CPU based stereo block matcher with a sophisticated local temporal component [8]. A module for estimating the upper body globally consistent in spatio-temporal space denoted as *STO* estimator. And finally, a GPU based Patch-Sweeping module for high precision depth estimation [3]. This novel approach extends the work described in [9] in order to decrease the computational complexity to allow the algorithm to run on a broader range of hardware and to increase quality and temporal consistency at the same time.

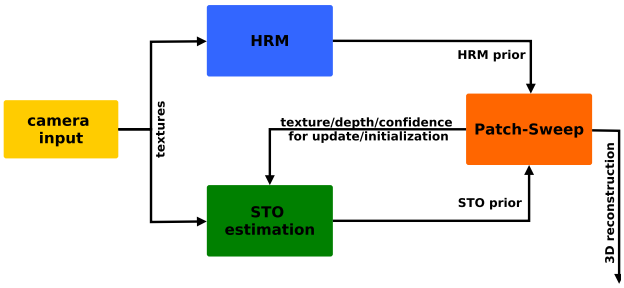


Fig. 2. Module interaction and data flow of the proposed algorithm.

The input for both *HRM* and *STO* are three synchronized and calibrated video streams. They compute complementary priors for the Patch-Sweep module, which outputs high precision 3D reconstruction and serves the feedback loop for *STO* initialization and update. A detailed discussion of each component will be presented in the following sections.

3. PRIOR CONSTRAINED PATCH-SWEEPING

Patch-Sweeping is a brute force high precision depth estimation algorithm [3]. It is designed with focus on parallelization and therefore perfectly suited for GPU implementation. Utilizing the massive computational power of modern GPUs, Patch-Sweeping can be executed at video frame rate even for high resolutions. However, reducing the sweeping range by applying a reasonable prior decrease the hardware demands significantly, while increasing the reconstruction quality at the same time. For prior selection, it was pointed out that priors, which are computed in an algorithmically different way turned out being very beneficial [9]. In the very same work, the authors propose *HRM* for prior computation. The resulting approach offers a great improvement, but still misses a global constraint that enforces temporal consistency on object level. For this task an additional novel prior is introduced that fits well into the concept presented in [9] and can be considered as a consequent extension to the idea of different prior consolidation.

4. SPATIO-TEMPORAL OBJECT

Conceptually, a given 3D reconstruction of an upper body sequence can be considered as *STO* with a fixed topology. Therefore, its temporal evolvement can be described with an appropriate set of deformation and spatial transformation parameters. A precise estimation of these parameters based on an initial 3D representation for a certain time instance will enable the computation of a valuable additional Patch-Sweep prior. Please note, that the computation of the temporal transition parameters can be considered as consistent scene flow estimation for a single 3D object. While the *HRM* prior is based on a spatially local estimation with a temporal component, the *STO* prior emerges as a complementary, globally oriented scene flow constraint.

Clearly, since we are targeting at real-time processing, our goal is to minimize the computational load. Therefore, we avoid a full estimation of deformation and spatial transformation parameters. Instead, the *rigid object* assumption is introduced, which means that for small time instants the evolvement of the *spatio-temporal object* is approximated by a rigid spatial transformation. Especially, considering upper body video sequences, significant deformations are very unlikely between two time instances. However, the handling of exceptions is quite important. Please refer to section 5 for a more detailed discussion.

In the following, we focus on the iterative numerical estimation of the transition parameters for a *spatio-temporal object* $\boldsymbol{O}$ from one time instance t to the next $t+1$. For this task, color constancy between two time instants is assumed and a minimization target based on the squared differences of color intensities is formulated:

$$J(\boldsymbol{R}_t, \boldsymbol{t}_t) := \sum_i \int\limits_{\boldsymbol{X} \in \boldsymbol{O}_t} \left(T(\boldsymbol{X}) - I_i(P_i(H(\boldsymbol{X}, \boldsymbol{R}_t, \boldsymbol{t}_t)))\right)^2 d\boldsymbol{X},$$

where H is the function, that applies the rotation $\boldsymbol{R}_t$ and the translation $\boldsymbol{t}_t$ to $\boldsymbol{O}_t$ such that $\boldsymbol{O}_{t+1} = H(\boldsymbol{O}_t, \boldsymbol{R}_t, \boldsymbol{t}_t)$, T assigns a texture value to each $\boldsymbol{X} \in \boldsymbol{O}_t$, $P_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection function that maps spatial points onto image i and I_i assigns a texture value to each coordinate on image plane i. Please note, that in this manner the minimization is directly carried out with respect to the squared texture differences in 3D. The optimization parameters are consequently the translation vector $\boldsymbol{t}_t$ and the 3D rotation $\boldsymbol{R}_t$. Numerically, this problem can be solved via simple gradient decent. However, with a focus on efficiency it is desirable to apply a numerical scheme with a higher order of convergence. For this reason it is assumed that the 3D rotation is small. Therefore, its first

order approximation, namely an element of the group of 3×3 skew-symmetric matrices (the Lie group of infinitesimal 3D rotations), is considered sufficient for optimization. This approximation enables a more powerful Newton-Raphson numerical scheme proposed by Zimmermann et al. [10].

5. MULTI-PRIOR PATCH-SWEEP

Our multi-prior extension accepts an arbitrary number of prior inputs for Patch-Sweep. All priors are handled equally. First, for every prior a sweeping range is computed according to a predefined sweeping range parameter, e.g. a few millimeter. The prior is assumed to be in the center of the sweeping range. Second, different patch configurations are evaluated for each of them and the result is chosen with respect to the best matching score, cf. [3]. Third, the patch with the best matching score among all the prior induced sweeping ranges is selected as final result.

According to figure 2, two priors for upper body 3D reconstruction are used. The first one is *HRM* and the second is *STO* estimation. The interplay of *HRM* and Patch-Sweep is well documented in [9].

The novel *STO* prior needs to be initialized with a valid 3D reconstruction. This is done as soon as the first Patch-Sweep result is available, cf. figure 2. However, textures and depth and confidence maps from the Patch-Sweep module are of high resolution and might contain too many values for efficient *STO* estimation. Therefore, since the focus is on real-time processing, we reduce the amount of input values for initialization by rescaling the textures and the depth and confidence maps by a user defined scaling factor. Among the resulting reduced set of initialization values we select the most promising values for initialization according to a user defined confidence threshold. In this way, we ensure to pick up values for estimation that are already considered as reliable. Please note that a copy of the original depth input is kept for prior computation.

For *STO* estimation, the input textures are rescaled according to the scaling factor used for initialization in order to get a suitable input. Afterwards, the spatial transition parameters are computed according to section 4. Certainly, for the purpose of limiting the computational load, the maximum amount of iteration steps for estimation is fixed to a predefined value. Once the transition parameters are available, namely an infinitesimal 3D rotation and a spatial translation, the prior is computed by applying these parameters on the last *STO* update and finally transmitted to the Patch-Sweep module.

Clearly, *STO* estimation may fail in case of fast movements, perspective changes or may get biased by object deformation over time. For this reason, we adopt the concept of key-frames from video coding in order to reinitialize the estimation after a fixed amount of time. For this purpose the *spatio-temporal object* is reset at each key-frame and (re-) initialized as described above.

6. EVALUATION

Our approach will be evaluated on four upper body sequences with a length of 100 frames containing three different persons performing *very fast, fast* and *slow* upper body movements. They are recorded with a synchronized single baseline trifocal camera system. The recorded image size was full HD, where the recorded upper bodies cover approximately 70% of the image. As a preprocessing step, all images where segmented by a simple foreground background subtraction algorithm. The evaluation of the 3D analysis was performed with respect to resulting segmentation masks. Please refer to figure 3 for brief overview of the datasets.

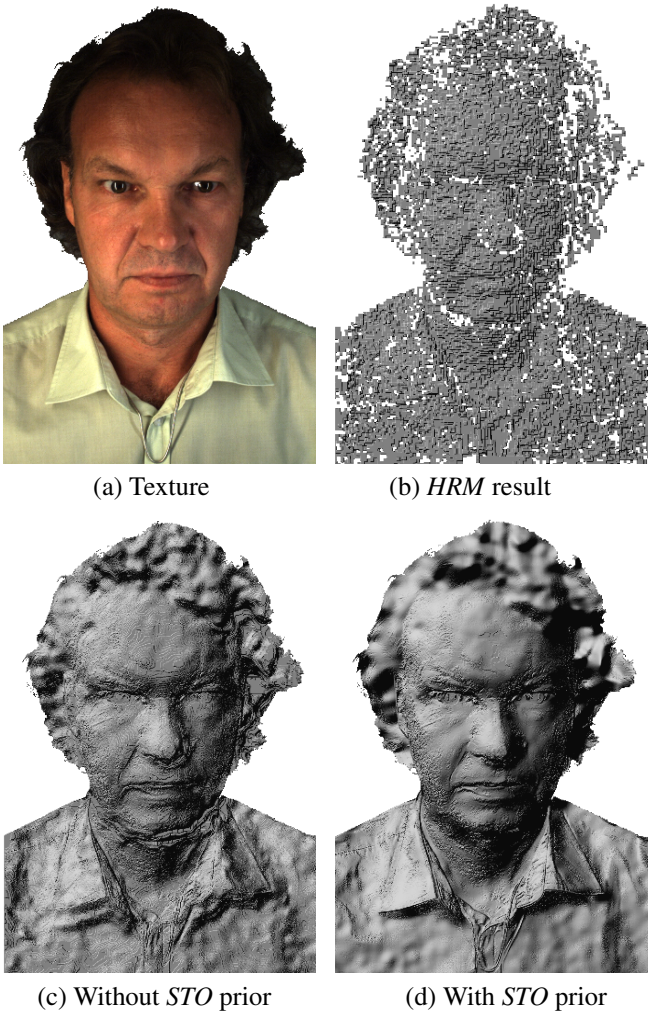
dataset	movement intensity	remarks
oliver 1	slow	calm listening person
oliver 2	very fast	heavy motion blur
florian	slow	speaking person
marian	fast	extended head movement

Fig. 3. Overview to evaluation datasets.

All our experiments where performed with a multi-threaded C/C++ implementation of the proposed approach, where GPU programming was conducted with CUDA. On a dual Hexacore Intel Xeon at 2.8GHz and two NVIDIA GTX 690 graphics cards, our current implementation runs at 10 to 15 fps in the experiments performed for this work. Please note, that the frame rate depends on the size of the segmentation mask and the chosen algorithmic parameters.

For all our experiments, we used the same algorithmic settings. Please note that some of the parameters listed below which concern *HRM* and Patch-Sweep are not explained in this work for brevity. A detailed description can be found in [8][3][9]. *HRM* was set to a 8×8 sub-grid processing with a matching block size of 8×8. Patch-Sweep was executed with three depth layers and three patch orientations with squared patches of edge length $7\mathrm{mm}$. The sweeping range was $5\mathrm{mm}$. Normalized cross correlation *NCC* was used as similarity measure for both algorithms. For evaluation and consistency purposes *NCC* was normalized to a *confidence range* of $[0 \dots 1]$ via $(NCC + 1)/2$. *Confidences* are set to zero for unavailable depth due to *HRM* mismatches or warping artifacts while generating the *STO* prior. Finally, the *STO* estimation is set to a maximal iteration count of 15, a rescale factor of 0.2, a key frame rate of 3 and a *confidence* threshold of 0.85. The evaluation is carried out by subjective quality assessment and a quantitative numeric comparison. Subjective quality improvements are discussed first.

Figure 4 shows one of the input textures and results in polygonal mesh representation for one frame of the ***oliver 1*** sequence. The top right image (b) shows the *HRM* output and the bottom left image (c) shows the output of Patch-Sweep with *HRM* prior but without a *STO* prior. It is clearly recognized, that the result is already of good quality. However,



(a) Texture (b) *HRM* result

(c) Without *STO* prior (d) With *STO* prior

Fig. 4. One of the input textures and qualitative comparison of the results.

the depth quantization is still a bit rough and different layers can be visually identified. Especially, the left part of the head, the collar and the chin suffer from those artifacts. Contrary, as depicted in the lower right image (d) most of these artifacts vanished after the application of the *STO* prior. Beside the visually significant improvement for a single frame, temporal consistency is positively affected as well. This will be demonstrated with videos at the conference presentation.

For numerical evaluation, the frame-wise mean confidence values have been calculated for all results within the foreground segmentation masks for all datasets. The outcome is visualized in figure 5. Each plot shows confidence values for one dataset with and without the *STO* prior. It can be clearly seen, that regarding confidence values the outcomes with *STO* priors are in general superior for all datasets and all frames. Moreover, for ***oliver 1*** at frame 50, ***florian*** at frame 75 and ***marian*** at frame 40 and 80 confidence drops caused by person movements occur. The duration and intensity for all of these drops is larger without the *STO* prior. Therefore, it can be concluded that the *STO* prior enables for a better recovery of 3D structure after significant scene changes. Similarly, considering ***oliver 2***, quality drops caused by heavy motion blur from very fast movement can be significantly attenuated with the help of *STO* prior.

7. CONCLUSION

In this work, we introduced a novel approach for constraining Patch-Sweep with a global scene flow like prior, via a *spatio-temporal object* estimation. We have successfully applied our algorithm to real world upper body datasets. The qualitative and quantitative improvements have been significant. Therefore, the multi-prior extension of constrained Patch-Sweep appears promising and the identification of additional appropriate priors might further enhance the 3D analysis. However, looking towards general 3D reconstruction we need to decompose the scene into several independent *spatio-temporal objects* in future work.

8. ACKNOWLEDGMENT

This research was supported by the European Commission under contract FP7-288238 SCENE.

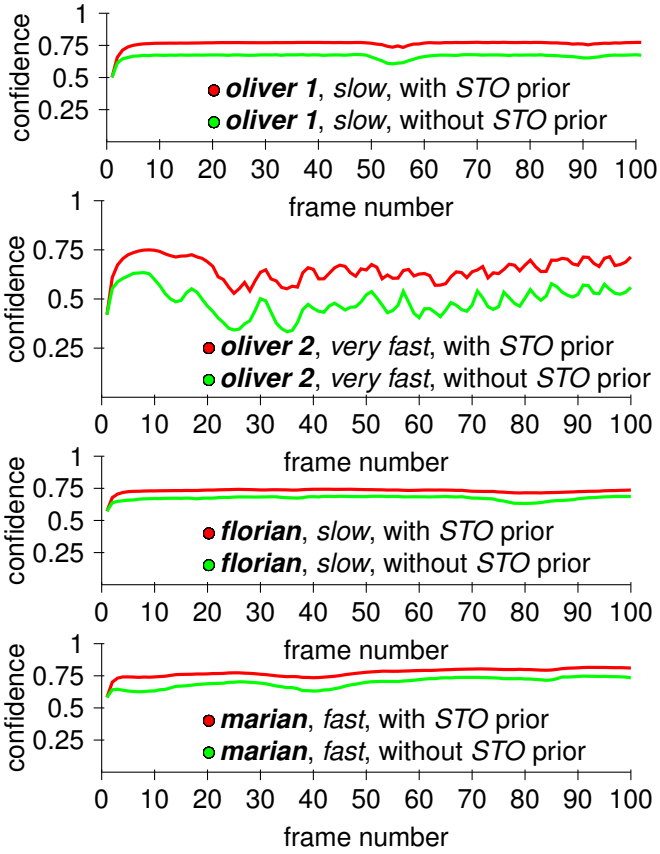


Fig. 5. Mean confidence of the results with and without *STO* prior for the four evaluation datasets.

9. REFERENCES

[1] J. Gemmell, K. Toyama, C.L. Zitnick, T. Kang, and S. Seitz, "Gaze awareness for video-conferencing: a software approach," *Multimedia, IEEE*, vol. 7, no. 4, pp. 26–35, 2000.

[2] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction," in *Proceedings of the conference on Human factors in computing systems - CHI '03*, Ft. Lauderdale, Florida, USA, 2003, p. 521.

[3] W. Waizenegger, I. Feldmann, and O. Schreer, "Real-time patch sweeping for high-quality depth estimation in 3D videoconferencing applications," in *Proc. of Real-Time Image and Video Processing*, San Francisco, California, United States, 2011.

[4] "European research project SCENE," http://3d-scene.eu/.

[5] "European research project REVERIE," http://www.reveriefp7.eu/.

[6] T. Matsuyama, Xiaojun Wu, T. Takai, and T. Wada, "Real-time dynamic 3-d object shape reconstruction and high-fidelity texture mapping for 3-d video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 357 – 369, Mar. 2004.

[7] Karsten Mueller, Aljoscha Smolic, Kristina Dix, Philipp Merkle, Peter Kauff, and Thomas Wiegand, "View synthesis for advanced 3D video systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–11, 2008.

[8] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 3, pp. 321–334, 2004.

[9] W. Waizenegger, N. Atzpadin, O. Schreer, and I. Feldmann, "Patch-sweeping with robust prior for high precision depth estimation in real-time systems," in *Proc. International Conference on Image Processing (ICIP 2011)*, Brussels, Belgium, 2011.

[10] K. Zimmermann, T. Svoboda, and J. Matas, "Multi-view 3D tracking with an incrementally constructed 3D model," in *Third International Symposium on 3D Data Processing, Visualization, and Transmission*, June 2006, pp. 488 –495.