Content Generation for 3D Video/TV

Nicole Brosch, Asmaa Hosni, Asmaa Hosni, Geetha Ramachandran, Liu He, Margrit Gelautz

A lack of suitable 3D content currently constitutes a major bottleneck for transferring the recent success of 3D cinema to our home TV. In this paper, we take a look at state-of-the-art techniques to generate 3D content from existing 2D or newly captured 3D content. In particular, we present a method to convert original 2D image sequences to 3D content by adding depth information with only little user support. Furthermore, we show results of a stereo algorithm which provides the basis for automatic conversion of stereoscopic film material for viewing on different types of displays. In this context, we also discuss the potential of inpainting techniques for filling in image regions that were originally occluded.

Keywords**:** 3D video, stereo vision, 2D to 3D conversion, inpainting

1. Introduction and Motivation

The new generation of media displays focuses on moving from 2D to 3D. In this context, the entire processing chain of 3D content generation - from the acquisition of suitable film/video material to conversion routines for transmission and display on different types of 3D devices - needs to be revisited (Mendiburu, 2009). The key idea behind the generation of 3D content is to be able to provide the viewer with an illusion of depth as seen in the real world. In this paper, we propose a comprehensive 3D TV content generation approach, covering its key components, namely creation of depth maps, depth-based rendering and interpolation for novel view generation. We present state-of-the-art techniques for the different processing steps and demonstrate high-quality results obtained in experiments with our test data.

In general, the stereoscopic depth experience emerges when watching two slightly shifted views of the same scene, each with one eye. In 3D cinemas, the separation of the two views is accomplished using special glasses that direct each presented view to the corresponding eye. The human brain processes these images yielding a depth perception by exploitation of the geometric differences (denoted as *stereo disparities*) between the two images.

There are several ways for generating 3D content that can be viewed on stereoscopic devices. An obvious way would be to use a stereo camera during the original video shooting. In principle, the acquired stereo video (consisting of two synchronized video streams) could be displayed directly on a suitable stereo monitor. In many cases, however, further processing steps are required to adjust the stereo content to different types of displays and viewing distances. For example, a stereo video that had been shot for display on a 3D cinema screen would – in its original form - yield an uncomfortable viewing experience on a small-size mobile 3D display. The viewing freedom, encompassing the viewing distance as well as the position of viewing, plays an important role in determining the number of views required. Hence, a key requirement for 3D content adaptation is the generation of novel views that simulate virtual cameras that were not available during the original video acquisition. A particular need for novel view generation comes up in the context of (multi-user) autostereoscopic displays, which rely on multiple views to enable glass-free 3D viewing. The related conversion procedures typically require the computation of a *depth* (or *disparity*) *map* as intermediate product.

In many applications, the depth map is computed from two views of the same scene using stereo vision methods. However, if only one view is available, such as for existing monocular videos, 2D-to-3D conversion methods can be considered as an alternative solution. To compensate for the missing stereo information, the 2D-to-3D conversion typically relies on user input to make the problem tractable. In this paper we will present state-of-the-art solutions for both scenarios.

An overview of the processing steps for 3D content generation is given in Figure 1. In Section 2, we present the principles of a stereo vision algorithm that automatically computes high-quality depth maps from

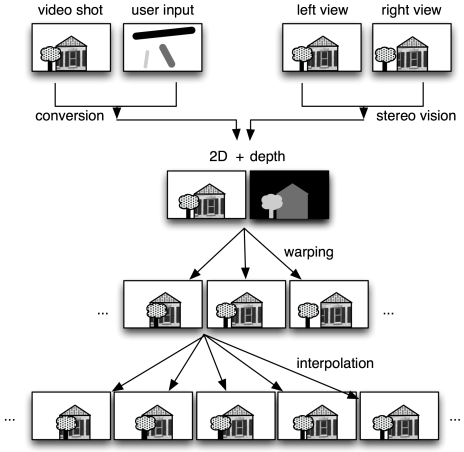

Fig. 1: 3D content generation process.

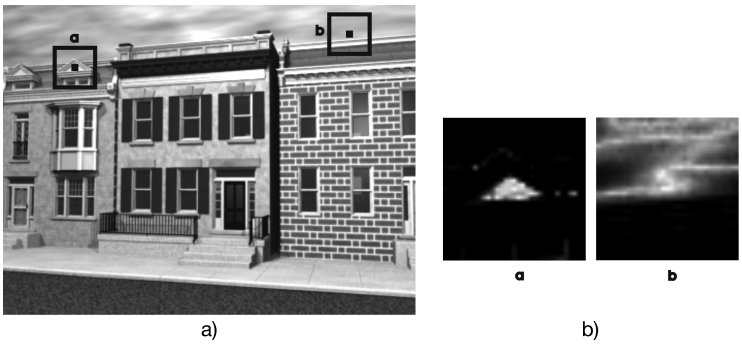

Fig. 2: Support regions for selected windows. Bright pixels represent high support weights and dark pixels otherwise. Our segmentation method gives relatively low support weights to pixels whose disparity is different from that of the center pixel. Original "Street" image from (Richardt et al., 2010).

stereo images as input. Section 3 focuses on the alternative approach of a 2D-to-3D conversion technique that propagates depth information - given in the form of a few user-scribbles on a monoscopic input sequence - to the entire video shot. The depth maps computed by these techniques are then utilized for novel view generation using suitable interpolation methods in Section 4. In this context, we address the need for image inpainting techniques in order to fill in regions that were occluded in the original video material. The results obtained in the different processing steps are discussed and illustrated using a variety of test images including a "Girls" sequence recorded in our lab.

2. Stereo Vision

The concept of stereo vision is to reconstruct 3D information, using two images that capture the same scene, but are recorded from slightly different viewpoints. The key challenge in stereo vision is to compute the *disparity map* - that is, to find a corresponding pixel in the right image for each pixel of the left image. This problem is known as the *stereo matching* problem and represents a crucial step in 3D video/TV content generation. The simplest strategy for solving the stereo matching problem is a so-called local approach. Here, a support (square) window is centered on a pixel of the reference (left) image. This support window is then shifted in the matching (right) image to find a point of maximum correspondence. As opposed to local approaches, global stereo techniques seek to apply an optimization scheme to the whole scene. Global approaches usually require much more computation time and are not further considered in the context of this paper.

It should be noted that the computed disparity map, which encodes the pixel shift between left and right image, is closely related to the depth map. An image pixel with a high disparity value has experienced a large geometric shift between the two images, because it is located close to the (stereo) camera. Contrarily, scene points that are far away from the camera are characterized by low disparity values. For the sake of simplicity, we use the terms disparity map and depth map interchangeably in this paper, although they are actually inversely proportional.

Traditional local stereo approaches apply an implicit assumption that all pixels within the support window are assumed to have the same depth (or, equivalently, disparity) values. This assumption is systematically violated in areas that are close to disparity borders (which often coincide with object boundaries.) As a consequence, the well-known *foreground fattening* effect rises up. This foreground fattening effect represents the inherent problem of standard local methods. On the other hand, a precise reconstruction of the object outlines is

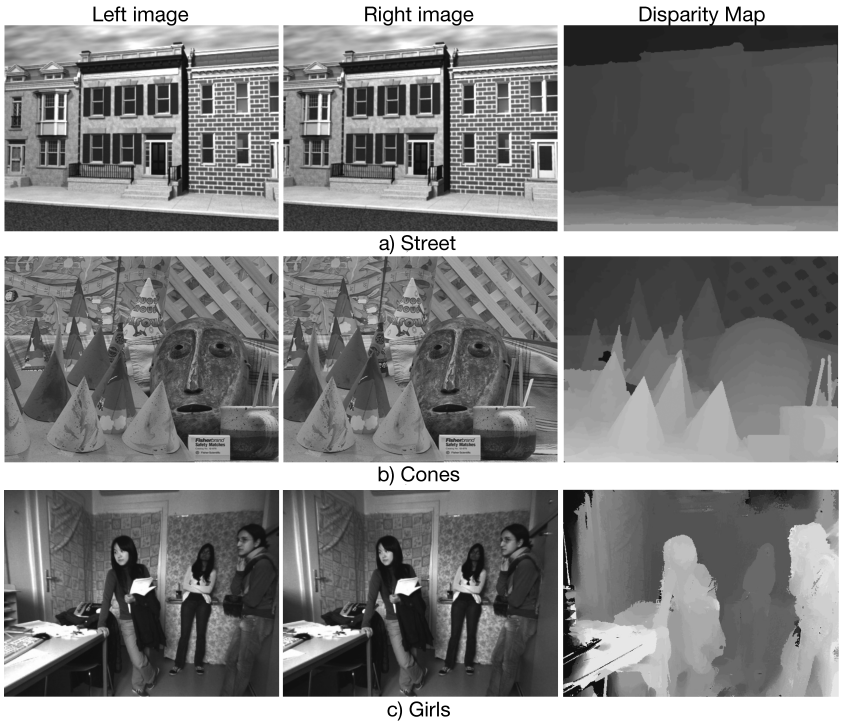

Fig. 3: Sample images and their corresponding disparity maps. a) "Street" image taken from (Richardt et al., 2010). b) "Cones" image taken from (Scharstein, Szeliski, 2002). c) Frame taken from our live system ("Girls").

particularly important for 3D content generation, since the human eye is very sensitive to artifacts along object boundaries which prevent a clear 3D impression.

In our algorithm, this problem is solved by explicitly identifying those pixels of the support window that are most likely to share the same depth value. For example, consider Figure 2 a) where we want to compute an "optimal" support window for the central pixel surrounded by a rectangle. In Figure 2 b), bright pixels represent pixels that are likely to have the same disparity values as the center pixel. Our method only uses those bright pixels in the correspondence search (matching process) and hence avoids the foreground fattening problem described above.

The key question that we have to answer is: "How can we extract those pixels that lie on the same disparity with the center pixel?" We solve this problem by using the concept of self-similarity, i.e. pixels that are close in color and spatial distance to the center pixel of the support window are most likely similar in disparity (because they are likely to lie on the same object). This concept has originally been used in (Yoon et al., 2005), where the likelihood that two pixels have the same disparity value is computed by comparing their color values and spatial positions. These two cues (color and spatial distance) are motivated by the *Gestalt theory*. In our work, we introduce a third Gestalt cue, namely *connectivity* (Hosni et al., 2009). We state that two pixels should be connected in the image by a path along which the color does not change sharply. This connectivity cue leads to improved adaptive support weight windows and hence to improved matching results (with improved disparity maps).

Examples of disparity maps that are generated by our stereo matching method are shown in Figure 3. As can be seen from this figure, our algorithm performs well in the reconstruction of disparity borders, while it also finds correct disparities for regions of poor texture, which are a challenge for local stereo methods. Our algorithm is evaluated by the Middlebury Stereo Vision Benchmark (http://www.middlebury.edu/stereo/), a well-known trusted evaluation system (Scharstein and Szeliski, 2002). According to this evaluation, our algorithm currently outperforms all other competing local approaches.

3. 2D-to-3D Conversion

Conversion of existing monocular video material (conventional videos) to 3D content is a promising alternative to the production of 3D media using stereo views (see Figure 1). Such 2D-to-3D conversion methods either

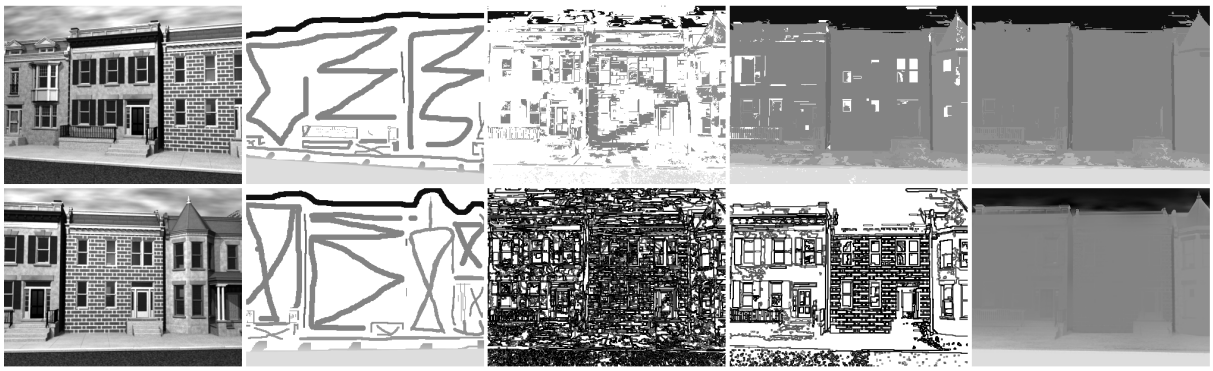

Figure 4: Propagation and segmentation process. a) Keyframes of input video. b) User scribbles (black: back, light gray: front, white: unknown). c) Pixel-wise over-segmentation of a middle frame (bottom) and corresponding disparity map (top). White pixels have not been assigned a disparity yet. d) Graph-based segmentation of regions to super-regions (bottom) and corresponding disparity map (top). e) Assignment of disparities to the missing regions (top). Final disparity map after refinement step (bottom). Original video from (Richardt et al., 2010).

automatically estimate the required disparity information by analyzing a video's content or propagate disparities given by a user. While the first option limits the choice of video material (e.g. static scenes), the incorporation of user input provides more flexibility. In semi-automatic conversion techniques, the general approach is to define disparity values in key frames (e.g. with scribbles, see Figure 4 a)-b)), which are then propagated to the entire video sequence. Here, the main challenge is to minimize the time a user needs to annotate key frames and to obtain high-quality disparity maps. More precisely, the result should be temporally coherent (no flickering) and contain smooth temporal disparity changes of moving objects. Spatial edges in the disparity map should be consistent with the input video and maintain the disparity discontinuities at object outlines.

The strategy of propagating sparse user input assumes that neighboring pixels that are similar in color, share the same or at least similar disparities. This concept is related to spatio-temporal video segmentation, where the goal is to group pixels that are similar in a certain feature space (e.g. color) into regions. Furthermore, segmentation provides additional information about object borders, which is often difficult to preserve (see e.g. (Guttmann et al., 2009)). Based on these observations our approach propagates disparities simultaneously with segmenting the input video.

Our approach (Brosch et al., 2011) builds upon a graph-based video segmentation algorithm suggested by (Grundmann et al., 2010), which we extend to incorporate depth information. The algorithm comprises two steps, a pixel-based over-segmentation (Figure 4 c)), and the subsequent merging of adjacent regions into super-regions (Figure 4 d)). In the first step, the segmentation algorithm compares spatially and temporally neighboring pixels in a fixed order and merges them into regions, if they are similar in color. During this process, we propagate the disparity information derived from the user scribbles that were drawn on the first and last frame of the video sequence. In case a pixel without disparity information merges with a pixel of known disparity, the known disparity value is propagated. To enable slanted surfaces and disparity changes in time, it is also possible to merge pixels with conflicting disparities. In this case, the original disparities of the individual pixels are kept within the merged region.

In the second step, neighboring regions, which were derived in the previous step, are compared. Again, similar regions are merged. Here, instead of expressing similarity by the pixels' color difference, color histograms and per-frame-motion histograms are used (Grundmann et al., 2010). During this process, the known disparity values are further spread among neighboring regions, while suitable strategies are applied to resolve possible ambiguities in depth assignment.

When applying this merging and assigning process iteratively, more and more pixels are assigned to disparity values (see Figure 4 a)-d)). To speed this process up, we abort it after several iterations (e.g. 25) and assign the remaining regions by using the disparity of the most similar neighboring region (see Figure 4 e), top). As a result we obtain a full disparity video, which contains abrupt

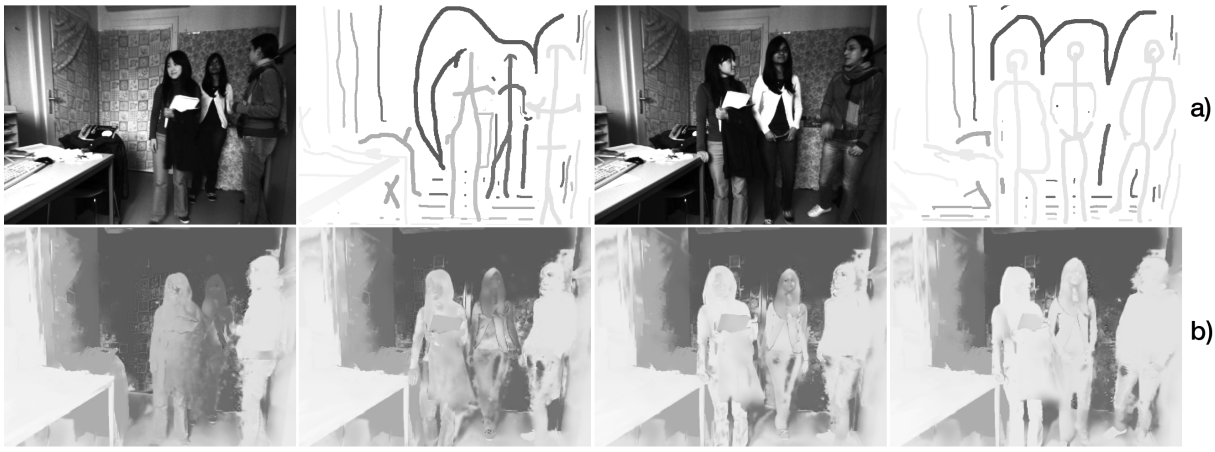

Figure 5: Depth propagation results. a) Keyframes and user scribbles (black: back, light gray: front, white: unknown). b) Example frames of obtained disparity map (black: back, white: front).

temporal changes. To interpolate disparity changes over time, we apply a generalized version of an edge-preserving smoothing filter (He et al., 2010) on each segment independently. Thereby, we use a kernel diameter as large as the temporal extent of a segment. We additionally refine disparities at object outlines. Here, we again apply the generalized guided filter, but in this case with a smaller kernel (i.e. diameter of 3 pixels) and on the entire video sequence. As this filter is sensitive to fine image structures (e.g. hairs), it captures details that have not been visible before (see Figure 4 e), bottom).

The algorithm described above delivers temporal-coherent disparity maps that contain smooth temporal disparity changes, while disparity edges at object borders are preserved. A more detailed evaluation performed in (Brosch et al., 2011) shows that our algorithm performs favorably in comparison to a state-of-the-art disparity propagation method developed by (Guttmann et al., 2009). In Figure 5, we present some results from the self-recorded "Girls" sequence, which demonstrate that our algorithm handles video shots that contain partial occlusions and motion.

4. Novel View Generation

As outlined in Figure 1, the generation of novel views that can be associated with virtual cameras not present in the original scene is a key component for generating high-quality and versatile 3D video content. A particular need for novel views comes up in the context of recent progress in the field of autostereoscopic displays. The idea behind multi-view autostereoscopic displays is to allow the consumer to have the same comforts as that provided by 2D TV presently; namely to be able to have a viewing position of choice and not to be encumbered by having to wear special glasses.

In principle, multi-view autostereoscopic displays rely on multiple images of the same scene, taken from different viewpoints, to render a 3D picture to the viewer. The different images are mapped to different pixel columns on the display matrix in an interleaved way. Optical elements embedded into the screen then properly focus the individual images and direct them towards different viewing angles. As a result, multiple viewers gathered in front of the screen receive a 3D impression corresponding to their individual viewing positions, without the need for stereo glasses.

A straightforward approach to render multiple views would be to interpolate between two existing (stereo) video streams. This approach is based on the motion compensation techniques used in (Jain and Jain, 1981). Similar methods have been used in the past, for example, by (Raya and Udupa, 1990) and (Saito et al., 1999). We illustrate the results of such an approach in Figure 6 where the left and right view of the Cones images pair (Scharstein, Szeliski, 2002) are used to generate the central view.

The algorithm starts by local block matching of the left and right frames to model the disparity in the frames. Linear interpolation is used to place the matched block from the right frame into the novel view. This process is repeated until the first version of the intermediate frame is obtained (Figure 6 a)). The novel view then undergoes an iterative refinement to fill in holes. This process involves obtaining intensity values from neighboring pixels as well as from the stereo

a) b)

Figure 6: Interpolation results to generate novel views. a) Shows the novel view after the initial search for corresponding blocks. It contains holes (black). The final results after the refinement steps are shown in b).

views. The final result of the refinement steps is shown in Figure 6 b).

An alternative approach for novel view generation relies on disparity maps as generated in Sections 2 and 3. An original view along with its associated disparity map ("2D + depth") can be used to synthesize high-quality new projections by applying suitable image warping and inpainting procedures, which we explain in the following.

4.1. Image Warping

Image warping is the process in which the pixels of the reference view shift horizontally (assuming stereo pairs that have been rectified to the so-called epipolar geometry) to compose the generated view. The shifting distance of each pixel is equal to the value at this pixel position on its corresponding disparity map. The larger the disparity is, the more the pixel shifts. If several pixels shift to the same position, only the one with the largest disparity appears, because its associated 3D point is closer to the camera than the other ones and therefore occludes them. The disparity map itself can also be warped in the same way. An example of image warping is shown in Figure 7 a)-b).

In the generated view of Figure 7 a)-b), we can see gaps in black, where no pixel shifts into. This is because those parts are occluded and, hence, invisible in the reference view. To complete the generated view, we have to fill the gaps up with information from their surroundings. With the gap filling, the occluded areas appear. Therefore, this step is called disocclusion. Small holes can be filled by applying median filtering on the color image and the disparity map. Larger ones (e.g. the black gaps right to the girl's head Figure 7 a)) have to be handled with some more sophisticated methods, e.g. inpainting (see Section 4.2.).

4.2. Image Inpainting

The term inpainting is originally used to describe the artistic restoration process of a damaged painting or picture. In the digital world, this process is automatically performed by the computer. The inpainting algorithms are generally classified into Partial Differential Equation (PDE) based inpainting and texture synthesis based inpainting (Venkatesh et al., 2009). In PDE based inpainting, the colors surrounding the gaps are propagated across the boundaries and fill up the inside. The propagation process is like heat diffusion or fluid flow, so that the inpainting results are smooth. However, textures may be lost because of over-smoothing. In texture synthesis based inpainting, the target area is filled through texture replication. A texture can be copied from examples or generated procedurally from statistics over the whole image or a serial of images. The most popular texture synthesis approach is exemplar-based inpainting (Criminisi et al., 2003), in which the optimal exemplar is selected for each blank pixel by estimating the similarity between the template patch centered at the target pixel and the candidate exemplar.

While conventional inpainting methods only make use of the color information, depth-guided inpainting aims to improve the inpainting results by selecting exemplars under depth-constraints (He et al., 2011). This idea fits the application of disocclusion for 3D video well, as the disoccluded area should be the extension of its visible surroundings of some particular depth. To be more precise, the optimal exemplars for the gaps should be from the connected background, as the disoccluded regions in the generated view are the background in the

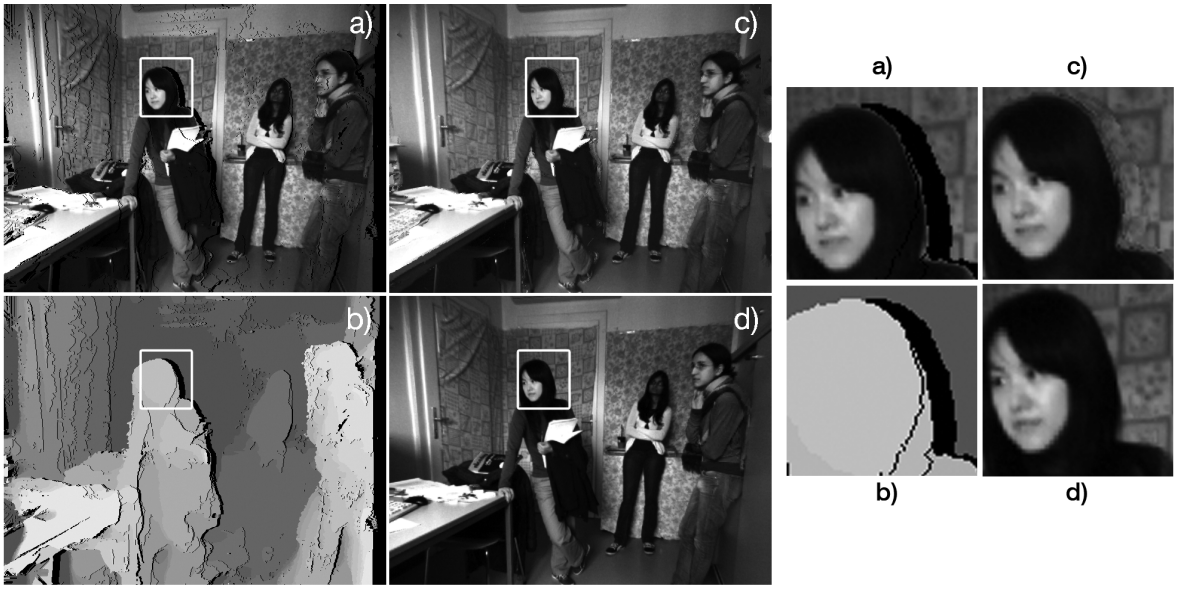

Figure 7: Image warping and inpainting for an example frame (left) and corresponding zoom-in patches (right). The generated right view a) and disparity map b) contain holes (black). In c) large holes are filled by depth-guided inpainting. d) shows the original right view for comparison.

reference view. Figure 7 c) shows the inpainted image generated by our algorithm. One can recognize that the textured wallpaper behind the girl's head is completed and the texture is preserved as well.

5. Conclusion

In this paper we have reviewed a complete 3D video/TV content generation system, which is an important requirement for the commercial success of 3D TV. We place emphasis on the key elements of a 3D content generation system - containing the creation of depth maps, depth-based rendering and novel-view generation - and how the individual parts work together. In addition to a state-of-the-art stereo technique which captures depth maps from two synchronized views, we have presented a 2D-to-3D conversion method, which is able to create high-quality depth maps from only one view using sparse and comfortable user input. Both technologies have the potential to overcome the lack of available 3D video content, which presents currently a major bottleneck in the introduction of 3D TV.

The latest generation of 3D autostereoscopic displays needs to generate novel views in order to create an enhanced depth viewing experience that can be enjoyed simultaneously by multiple viewers. In this context, we presented an interpolation technique to generate these intermediate views and demonstrated the results of a sophisticated inpainting method for filling in regions that were occluded in the original view. The suggested processing chain enables the generation of high-quality 3D content at reduced production costs, due to the comprehensive and flexible generation of depth maps and associated synthesized camera views.

6. References

Brosch N., Rhemann C. and Gelautz M. (2011): Segmentation-based depth propagation in videos. In: Proc. of ÖAGM/AAPR*, 14.*

Criminisi A., Perez P. and Toyama K. (2003): Object removal by exemplar-based inpainting. In: Proc. of CVPR 2*,* 721–728.

Grundmann M., Kwatra V., Han M. and Essa I. (2010): Efficient hierarchical graph-based video segmentation. In: Proc. of CVPR, 1–14.

Guttmann M., Wolf L. and Cohen-Or D. (2009): Semi-automatic stereo extraction from video footage. In: Proc. of ICCV, 136–142.

He K., Sun J. and Tang X. (2010): Guided image filtering. In: Proc. of ECCV, 1–14.

He L., Bleyer M. and Gelautz M. (2011): Object removal by depth-guided inpainting. In: Proc. of ÖAGM/AAPR*, 15.*

Hosni A., Bleyer M., Gelautz M. and Rhemann C. (2009): Local stereo matching using geodesic support weights. ICIP, 2093–2096.

Jain J. and Jain A. (1981): Displacement Measurement and Its Application in Interframe Image Coding, IEEE Transactions on Communications 29, nr.12: 1799- 1808.

Mendiburu B. (2009): 3D Movie Making – Stereoscopic Digital Cinema from Script to Screen, Oxford, UK: Elsevier/Focal Press Inc.: 223

Raya S.P. and Udupa J.K. (1990): Shape-based interpolation of multidimensional objects, IEEE Transactions on Medical Imaging 9, nr.1: 32-42.

Richardt C., Orr D., Davies I., Criminisi A. and Dodgson N. (2010): Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In: Proc. of ECCV 3, 510-523.

Saito H., Baba S., Kimura M., Vedula S. and Kanade T. (1999): Appearance-based virtual view generation of temporally-varying events from multi-camera images in the 3D room, In: Proc. of 3DIM, 516-525.

Scharstein D. and Szeliski R. (2002): A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 47, nr. 1/2/3: 7-42.

Venkatesh M.V., Cheung S. and Zhao J. (2009): Efficient object-based video inpainting. Pattern Recognition Letters 30, nr. 2: 168–179.

Yoon K.J. and Kweon I.S. (2005): Locally adaptive support-weight approach for visual correspondence search. In: Proc. of CVPR 2, 924-931.