Unsupervised object segmentation for 2D to 3D conversion

Matthias Kunter, Sebastian Knorr, Andreas Krutz*, and Thomas Sikora*

imcube Media,
Technische Universität Berlin,
Einsteinufer 17, 10587 Berlin, Germany
{kunter, knorr}@imcube.de

* Communication Systems Lab,
Technische Universität Berlin,
Einsteinufer 17, 10587 Berlin, Germany
{krutz, sikora}@nue.tu-berlin.de

ABSTRACT

In this paper, we address the handling of *independently moving objects* (IMOs) in automatic 2D to stereoscopic 3D conversion systems based on *structure-from-motion* (SfM) techniques. Exploiting the different viewing positions of a moving camera, these techniques yield excellent 3D results for static scene objects. However, the independent motion of any foreground object requires a separate conversion process. We propose a novel segmentation approach that estimates the occluded static background and segments the IMOs based on advanced change detection. The background estimation is achieved applying 2D registration and blending techniques, representing an approximation of the underlying scene geometry. The segmentation process itself uses anisotropic filtering applied on the difference image between original frame and the estimated background frame. In order to render the segmented objects into the automatically generated 3D scene properly, a small amount of user interaction will be necessary, e.g. an assignment of intra-object depth or the object's absolute z-position. Experiments show that the segmentation method achieves accurate mask results for a variety of scenes, similar to the masks obtained manually using state-of-the-art rotoscoping tools. Though, this work contributes to the extension of SfM-based automatic 3D conversion methods for the application on dynamic scenes.

Keywords: 2D to 3D conversion, object segmentation, image registration, structure-from-motion

1 INTRODUCTION

With the ongoing development of stereoscopic displays, 3D projection techniques, and 3D cinema, the stereoscopic content generation process attracts a great deal of attention. Rendering a second stereoscopic view from a monocular image sequence, which is also known as 2D/3D conversion, is a promising way to achieve high quality stereo motion pictures. Over the past years it has been shown that for certain recording conditions the conversion can be achieved fully automatically [1]-[8]. These approaches can be classified in those generating dense depth maps [1]-[5] for *depth image based rendering*, and approaches that generate a stereo view using other techniques [6]-[8], e.g. planar image transformation. The former class of conversion methods is relatively error-prone due the high probability of misestimated regions in the depth map and the problem of occlusion handling. The latter class of approaches yields visually better results, but only in the absence of *independently moving objects* (IMOs), i.e. a recording of fixed scene objects. In [8] we presented the *realistic stereo view synthesis* (RSVS) approach which is based on *structure-from-motion* (SfM) [9] and *image-based rendering* (IBR) techniques [10] to create virtual stereoscopic views.

In this work, we present a mainly automatic 2D/3D conversion of scenes containing IMOs captured with a moving camera. For the conversion of the static background RSVS is used, thus, a camera movement with a minimal portion of horizontal translation is necessary [8]. In order to render a second view for the IMOs, a motion-based object segmentation technique is applied. We achieve the separation by locally estimating the background image for every frame using a robust blending criterion for registered neighboring images [11], [12]. The image registration process is approximated by a 2D image transformation, which does not exactly represent the underlying geometry but yields sufficient accurate background models. The final segmentation is processed on the difference image between the actual frame and its background model using anisotropic filtering techniques for homogenization and morphological operators.

The overview of the main system is shown in Figure 1, which is an extension of our 2D/3D conversion system presented in [13]-[15].

The organization oft this paper is as follows. The next section briefly describes the fundamental steps of a robust *SfM* analysis in order to estimate the camera path and a sparse reconstruction of the static background objects. The data obtained by this procedure represents the basis of any following step. In Section 2, the image blending and the moving objects segmentation process are outlined. The robust conversion process of the background using RSVS is explained in Section 4. Finally, experimental segmentation results are presented and a conclusion is drawn.

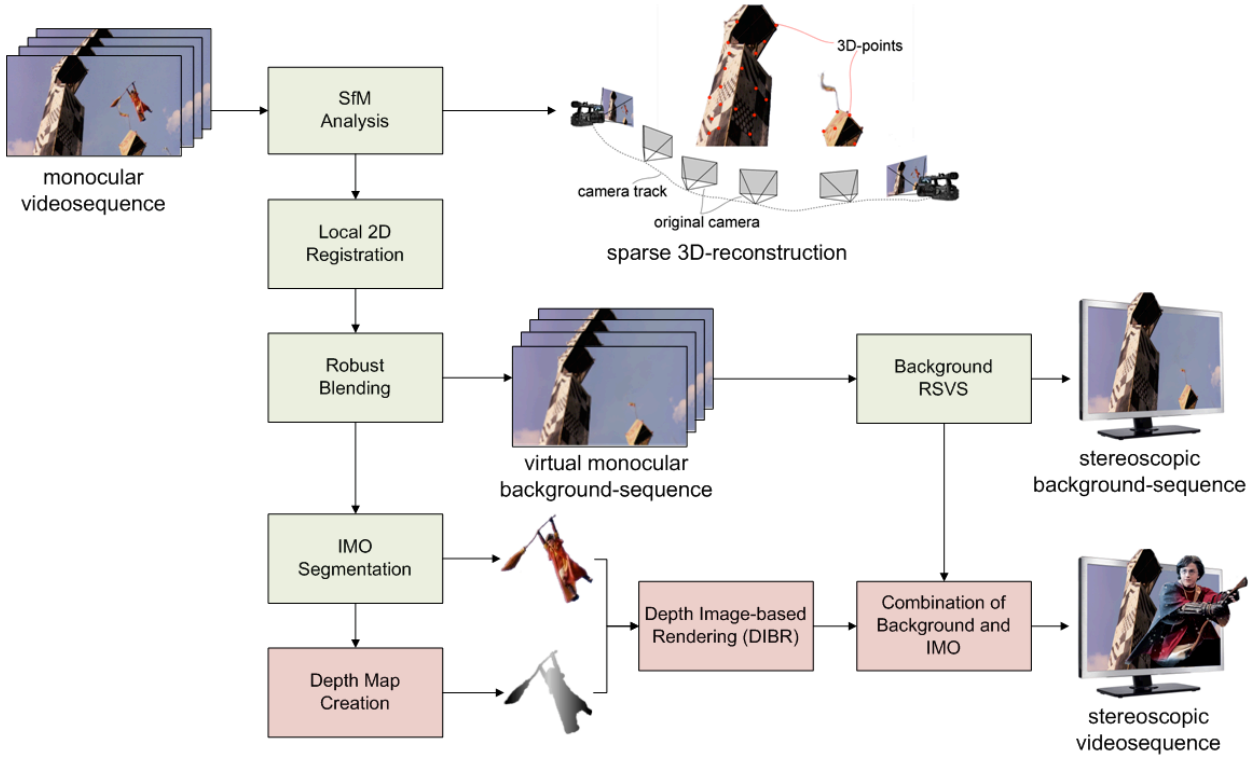

Figure 1: Flowchart of the overall system for 2D to stereoscopic 3D video conversion of scenes with independently moving objects (IMOs) captured by a moving camera. The paper mainly addresses the techniques described in the green blocks.

2 STRUCTURE-FROM-MOTION ANALYSIS

The structure-from-motion analysis consists of two stages. First, a robust feature tracking and key frame selection process is performed. Thus, camera path and a sparse 3D point reconstruction can be achieved even with an amount of non-supporting regions, e.g., IMOs [16]. The key frames are used for sparse reconstruction of the scene background. In order to obtain a satisfying reconstruction result, key frames are selected to guarantee an optimal trade-off between a sufficient distance of their camera centers and a sufficient amount of joint feature points. Here, we use the Geometric Robust Information Criterion (GRIC) as a robust model selection criterion [17].

Second, a joint self calibration and reconstruction process is executed. The reconstruction follows the classical approach of a robust camera matrix estimation and 3D point estimation for two key frames and refinement with further key frames [18]. A more detailed description of the whole SfM approach can be found in [15].

3 IMAGE BLENDING AND INDEPENDENTLY MOVING OBJECT SEGMENTATION

3.1 Background image estimation

In this approach, we rely on the fact that independently moving foreground objects move significantly different compared to the camera motion. Thus, occluded areas or parts of occluded areas will be discovered somewhere in the sequence. In order to estimate the background of any frame F_n of a shot, the 2D perspective transformation (homography $\mathbf{H}_{n,k}$) between F_n and its adjacent frames F_k with $n\text{-}\Delta \leq k \leq n\text{+}\Delta$ is estimated. This is achieved by linearly solving the mapping

$$m_n^i = \mathbf{H}_{\mathbf{n,k}} \cdot m_k^i , \tag{1}$$

where m_n^i denotes the projection of the i-th 3D feature point M_n^i into frame F_n and m_k^i denotes the projection of the same feature point into frame F_k. Note, that all points are written in homogenous coordinates. The number Δ of adjacent frames to be mapped is limited by an average Euclidian distance error for the mapped 2D feature points. It is quite obvious that the applied 2D mapping actually only complies with camera motions that have no translational component. However, for common recording settings the distance of neighboring views of a camera path can be neglected compared to the object's distance. Thus, the mapping of the background content is still satisfying for a limited number of neighboring frames.

Using a robust blending approach, we estimate the background image for each frame F_n. This is achieved by warping the image content of all adjacent frames into the reference coordinate system of F_n according to the determined transformation parameters. Robustness during the pixel blending can be achieved by modeling the color distribution for all candidate pixels [19]. In many cases, especially with fast foreground objects, it is sufficient to select the median pixel value. The principle of the whole blending process is depicted in Figure 2. The resulting background image is the source for the object segmentation process and the realistic stereo view synthesis (RSVS) of the image background.

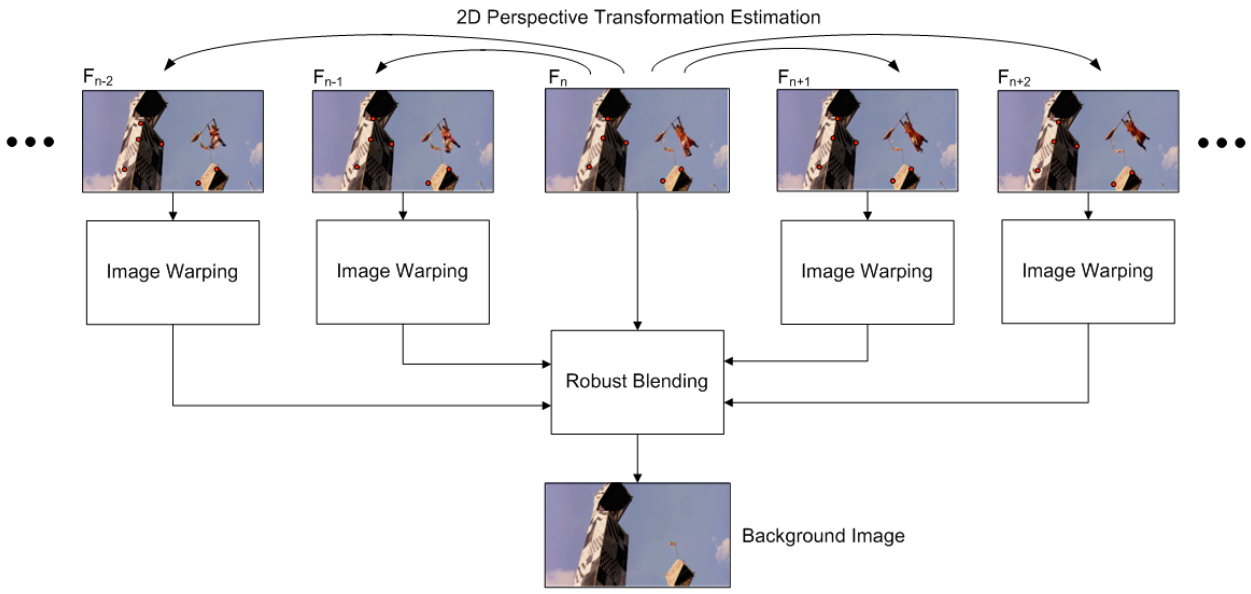

Figure 2: Principle of the background image estimation for object segmentation. Since the independently moving objects do not comply with the camera motion, occluded regions will be visible in other images. This model holds especially for small changes between neighboring camera positions.

3.2 Moving object segmentation

For the segmentation process we apply a change detection algorithm on the luminance difference image. It adopts the techniques described in [12] and identifies significant regions of change between the original frame F_n and its estimated background image. The change detection process follows the typical scheme depicted in Figure 3. It consists of a

threshold operation, which coarsely classifies the pixels into foreground or background classes, and additional morphological operations in order to homogenize the classification result and eliminate small classification errors.

The main part of interest of the change detection is the threshold calculation, which is conducted as follows. First, the difference image D_n is filtered using an anisotropic diffusion filter solving iteratively

$$D_{nt} = dD_n / dt = \mathrm{div}\left(g\left(|\nabla I_n|\right)\nabla D_n\right), \tag{2}$$

where I_n represents the intensity image of frame F_n. The term g(s) is also called edge stopping function and is designed to suppress the homogenization at strong edges of the original image. Second the threshold is calculated adaptively using

$$\tau_c = \mathrm{mean}\left(D_s(x,y)\right) + 0.1 \cdot \left(\max\left(D_s(x,y)\right) - \mathrm{mean}\left(D_s(x,y)\right)\right). \tag{3}$$

Here, D_s represents the smoothed difference image as output of the iterative anisotropic diffusion process. The output of the change detection process is a binary foreground mask. In this approach it is important to extract the foreground objects completely. Therefore, we concentrate on low false negative rate rather than on a low false positive rate. See Section 5 for segmentation results.

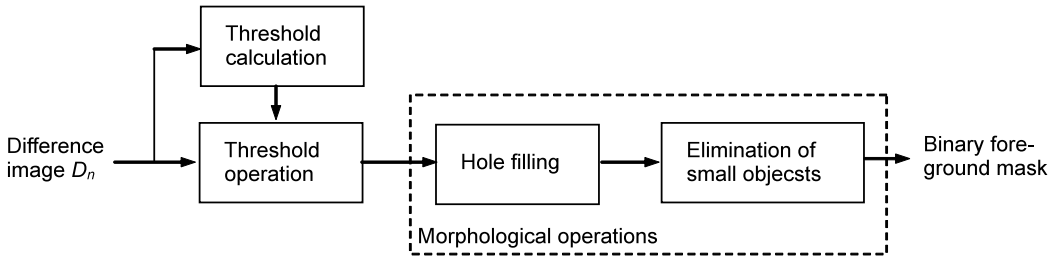

Figure 3: Flow chart of a typical change detection process for the segmentation of foreground objects based on luminance frame difference

4 AUTOMATIC BACKGROUND CONVERSION (RSVS)

In order to generate a second – left or right – background view for each original view, we define a virtual camera in the scene. Here, we apply a parallel camera setup. Thus, the obtained stereo image pairs have no vertical parallax. Since the SfM results are only defined up to a scale, the user has to give a normalization constraint, i.e., the distance of the first original camera in the camera path to a specific 3D point in the scene. Using this metric normalization, the virtual camera position $C_{n,v}^m$ can be defined by

$$C_{n,v}^m = C_n^m + \mathbf{R}_n^{-1} \begin{bmatrix} \pm t_x \\ 0 \\ 0 \end{bmatrix}. \tag{4}$$

Here, C_n^m represents the metric position of the n-th camera (frame) and $\mathbf{R}_n$ the respective rotation matrix. t_x is the average human eye distance. Calculating the projection matrix $P_{n,v}^m$ with

$$\mathbf{P}_{n,v}^m = \mathbf{K}\mathbf{R} \cdot [\mathbf{I} \,|\, -\widetilde{C}_{n,v}^m], \tag{5}$$

where $\widetilde{C}_{n,v}^m$ represents the virtual camera center in inhomogeneous coordinates, we project the determined 3D points onto the image plain of the virtual camera. Using these projected points, the virtual image can be rendered applying planar image transformations and warpings. Therefore, our *realistic stereo view synthesis* (RSVS) belongs to the class of *image based rendering* (IBR) methods [10].

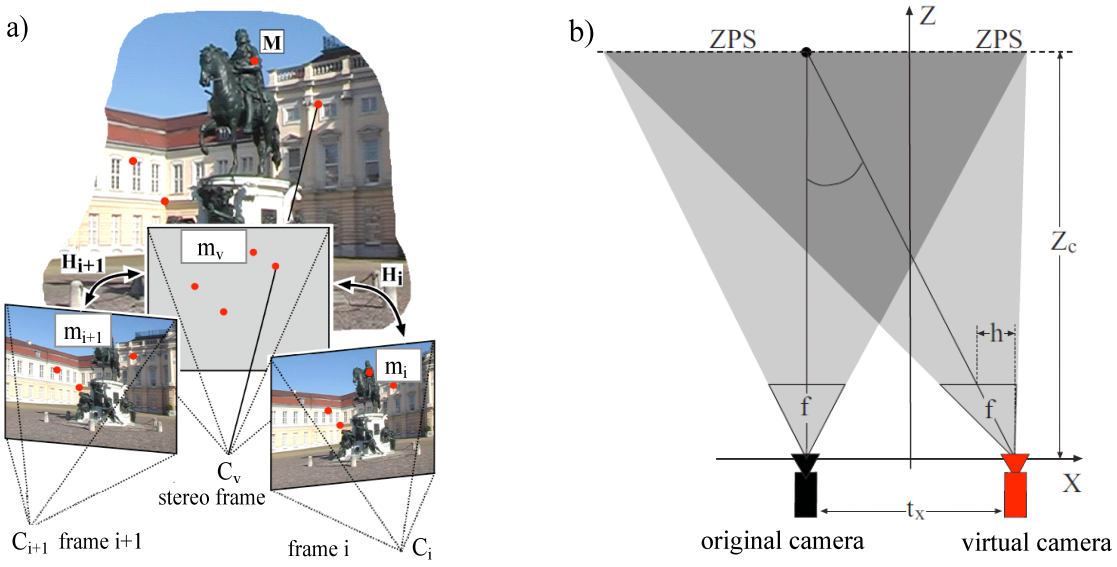

Figure 4: **a)** Schematic illustration of the image based stereo view synthesis using planar transformation, **b)** Parallel camera setup with shift sensor approach in order to select the *zero parallax setting* (ZPS)

4.1 Virtual camera image generation

The two advantages of IBR methods for stereoscopic view synthesis are the relatively small computational load and the property of photo-consistency. Thus, for recording setups that are compliant with the conditions for RSVS no occlusion handling is necessary. In order to render the image of the virtual camera, we look for the closest camera view in the previously estimated original camera path. Between these views a 2D perspective transformation based on the projected point correspondences using nonlinear optimization is determined. Figure 4a illustrates the transformation estimation. The resulting 2D homography $\mathbf{H}_{v,i}$ is then used to warp the image of Frame F_i into the coordinate system of the virtual image. Since original view and virtual view usually have different viewing angles, the generated image will not be fully covered by one warping process. Hence, undiscovered regions will be filled with content from next closest views.

Generally, a parallel camera setup yields only image pairs with negative parallax. However, a free choice for the *zero parallax setting* (ZPS), i.e., the location of the screen in the perceived 3D scene, is desirable. We apply the shift sensor approach to tackle this issue [20], [21]. The principle of this approach is depicted in Figure 4b. In the setup for the virtual view the position of the sensor is shifted perpendicular to the optical axis of the camera, which results in a change of the perceived depth for the whole scene.

5 EXPERIMENTAL RESULTS

5.1 Segmentation evaluation

For the evaluation of the segmentation results we apply classical objective pixel based measures. In order to do so, ground truth data has to be generated manually. As commonly known, the manually created ground truth masks highly depend on the editing individual and therefore, even those objective measures may vary within a certain error tolerance.

We evaluated three shots from well- known entertainment material. For a visual evaluation we depict the segmentation masks in a color coded manner following the scheme of Figure 5. Here, the true positives (*TP*) represent the number of correctly detected foreground pixels. The false negatives (*FN*) represent the number of pixels that are falsely marked as background and false positives (*FP*) are pixels falsely detected as foreground. The rest of the image is marked as true negatives (*TN*), i.e., pixels that are correctly detected as background. The pixel classification for some exemplary frames can be seen in Figure 6. As already pointed out we tried to minimize the number of false negatives (red) with the burden of a slight higher false positive rate (blue). This is motivated by the fact that all foreground objects should be extracted completely in order to obtain a good conversion result.

		ground truth	
		foreground	background
foreground segmentation	positive	TP	FP
	negative	FN	TN

Figure 5: Pixel classification for segmentation quality assessment

To describe the accuracy of the segmentation, we utilize three measures: precision P, recall R, and the F-measure F. The precision P is the ratio between the correctly detected foreground pixels and all pixels detected as foreground,

$$P = \frac{TP}{TP + FP}. \tag{6}$$

The recall R is the ratio between the correctly detected foreground pixels and all relevant foreground pixels,

$$R = \frac{TP}{TP + FN}, \tag{7}$$

while the F-measure represents the *harmonic mean* between R and P,

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{8}$$

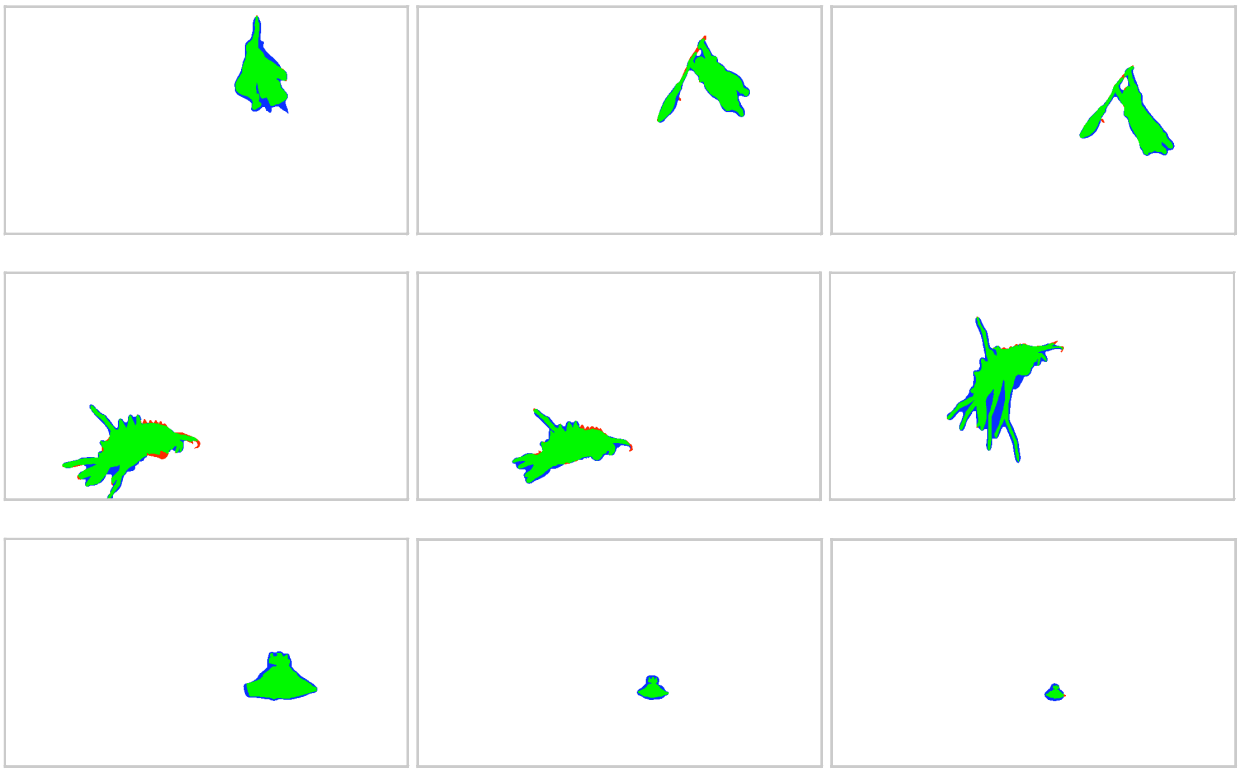

Figure 6: Segmentation results for shots of feature film “Harry Potter and the Philosopher’s Stone”, courtesy of Warner Bros. (72 frames, SDTV); nature documentary “Planet Earth”, courtesy of BBC (103 frames, SDTV); feature film “Star Wars – Episode IV”, courtesy of 20th Century Fox (109 frames; SDTV)

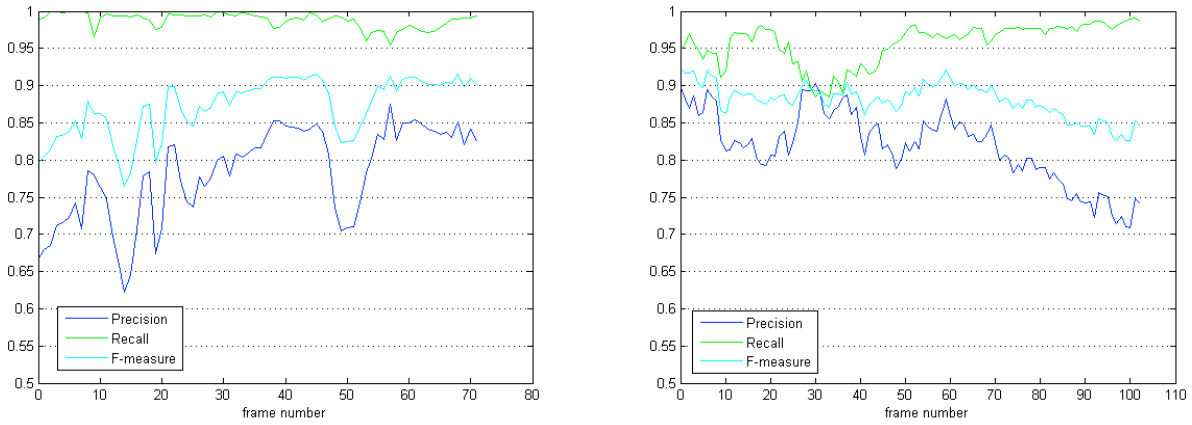

Figure 7: Segmentation quality measure for sequences "Harry Potter" (left) and "Planet Earth" (right).

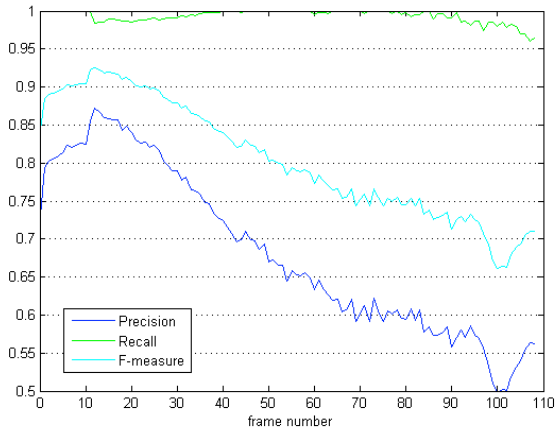

Figure 8: Segmentation quality measure for sequences "Star Wars". Note that due to a shrinking size of the foreground object the relative false positive rate increases, which has an impact on the precision value. Actually, the absolute number of false positive (FP) pixels stays constant.

The per-frame results for the three measures applied on all shots can be found in Figure 7 and Figure 8. We obtain an average recall value of 0.99 for "Harry Potter", 0.96 for "Planet Earth", and 0.99 for "Star Wars", respectively. The measured *F* value averages to 0.87 for "Harry Potter", 0.88 for "Planet Earth", and 0.81 for "Star Wars". With a recall value of almost 100% we can assure that moving foreground objects can be handled separately for 2D/3D conversion.

5.2 2D/3D conversion results

Results of all stages of the 2D/3D conversion system are shown Figure 9. The stereo view of the objects is rendered using DIBR techniques [22]. Here, a manual input of the user for one or several key frames is required in order to determine the absolute depth of the object in the scene. A gradual change in the depth over the image sequence is achieved applying linear depth tweening between the key frames. In our experiments, the intra depth of the objects is kept constant, which may result in a perception of flatness. Several approaches have been investigated that tackle this issue, e.g., in [23]. All these methods require a high degree of interactivity and go beyond the scope of this paper.

However, the generated stereoscopic sequences show that our approach meets two important demands. First, independently moving objects can be identified and extracted for more sophisticated object-based conversion techniques and second, the background handling using RSVS for 2D/3D conversion automatically takes care of the occlusion problem, i.e., newly discovered background areas due to the horizontal shift of the foreground objects. Conversion results for all three test sequences are depicted in Figure 10, Figure 11, and Figure 12.

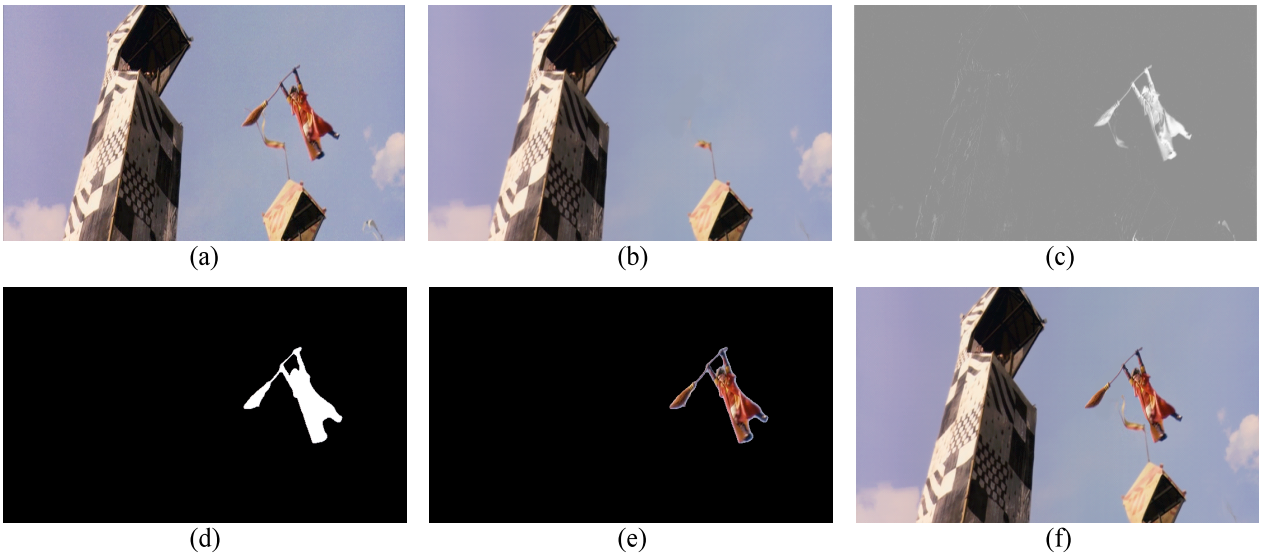

Figure 9: Segmentation and stereoscopic reconstruction results for a frame of the feature film "Harry Potter and the Philosopher's Stone". (a) original frame; (b) background image after blending; (c) luminance difference between a) and b); (d) binary object mask; (e) segmented object; (f) synthesized second view – the background image is the result of RSVS

6 SUMMARY AND CONCLUSIONS

This paper presented a new approach for the segmentation of independently moving foreground objects in sequences recorded with a moving camera, and is thus, utilizable for structure-from-motion-based 2D/3D conversion. This approach extends the applicability of our previous work, which was designed to handle static scenes only. Additionally, we drafted a system for the overall monoscopic to stereoscopic view conversion. It consists of an automatic background conversion and a manual object-based foreground conversion.

The main idea of the segmentation approach is the robust image background estimation using 2D image transformation and blending techniques. The pixel classification is performed applying change detection on the anisotropically smoothed difference image. Experimental results show that in cases where the IMO's motion is significant, high quality segmentation results can be obtained. Simultaneously, the image background estimation ensures photo consistency for uncovered regions while converting the foreground objects.

The limitations of this approach are twofold. Since SfM is used, a translation of the camera is necessary. Hence, conversion of fixed camera shots, shots from panning camera only, or shots from forward moving cameras is beyond the scope. Furthermore, the type of objects and their motion are of main importance. Multiple occluding objects or objects with only little motion are difficult to detect.

ACKNOWLEDGMENTS

The authors would like to thank Thilo Borgmann for generating the ground truth data utilized for assessing the presented segmentation technique.

Figure 10: Anaglyph image of converted frame 50 (sequence “Harry Potter”).

Figure 11: Anaglyph image of converted frame 10 (sequence “Planet Earth”).

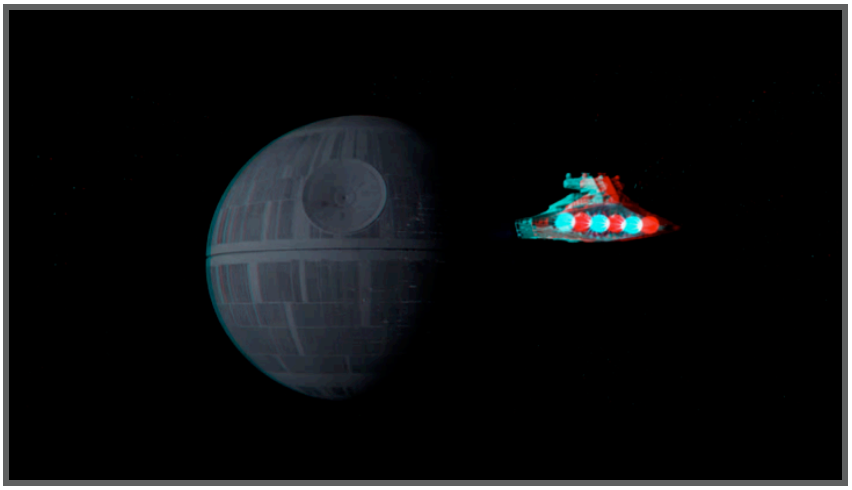

Figure 12: Anaglyph image of converted frame 5 (sequence “Star Wars”).

REFERENCES

[1] Diplaris, S., Grammalidis, N., Tzovaras, D., Strintzis, M. G., "Generation of Stereoscopic Image Sequences Using Structure and Rigid Motion Estimation by Extended Kalman Filters", Proc. of IEEE International Conference on Multimedia and Expo (ICME). Lausanne, Switzerland, 2002.

[2] Moustakas, K., Tzovaras, D. and Strintzis, M.G., "Stereoscopic video generation based on efficient structure and motion estimation from a monoscopic image sequence", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 15, No. 8, pp. 1065 - 1073, August 2005.

[3] Xu, F., Er, G., Xie, X. and Dai, Q., "2D-to-3D Conversion Based on Motion and Color Mergence", 3DTV-Conference, Istanbul, Turkey, May 2008.

[4] Matsumoto, Y., Terasaki, H., Sugimoto, K.., Arakawa, T., "Conversion System of Monocular Image Sequence to Stereo Using Motion Parallax", Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems III, San Jose, CA, USA, 1996.

[5] Kim, D., Min, D., Sohn, K., "A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis", IEEE Transactions on Broadcasting, vol. 54, no. 2, pp. 188-197, 2008.

[6] Rotem, R., Wolowelsky, K. and Pelz, D., "Automatic video to stereoscopic video conversion", Proc. of the SPIE: Stereoscopic Displays and Virtual Reality Systems XII, Vol. 5664, pp. 198-206, March 2005.

[7] Garcia, B. J., "Approaches to Stereoscopic Video Based on Spatio-Temporal Interpolation", Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems IV, San Jose, CA, USA, 1997.

[8] Knorr, S. and Sikora, T., "An Image-based Rendering (IBR) Approach for Realistic Stereo View Synthesis of TV Broadcast Based on Structure From Motion", IEEE Int. Conf. on Image Processing (ICIP), San Antonio, Texas, USA, Sept. 16-19, 2007.

[9] Jebara, T., Azarbayejani, A. and Pentland, A., "3D structure from 2D motion", IEEE Signal Processing Magazine, May 1999, Vol. 16. No. 3, p. 66-84.

[10] McMillan, L., "An Image-Based Approach to Three-Dimensional Computer Graphics", University of North Carolina at Chapel Hill, Diss., April 1997.

[11] Kunter, M., Kim, J.-H. and Sikora, T., "Super-resolution Mosaicing using Embedded Hybrid Recursive Flow-based Segmentation**",** IEEE Fifth Int. Conf. on Information, Communications and Signal Processing (ICICS '05), Bangkok, Thailand, December 6-9, 2005.

[12] Krutz, A., Kunter, M., Mandal, M. , Frater, M. and Sikora, T., "Motion-based Object Segmentation using Sprites and Anisotropic Diffusion"**,** 8th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Santorini, Greece, June 6-8, 2007.

[13] Knorr, S., Kunter, M., and Sikora, T., "Super-Resolution Stereo- and Multi-View Synthesis from Monocular Video Sequences", 3-D Digital Imaging and Modeling (3DIM), Montréal, Québec, Canada, August 21-23, 2007.

[14] Knorr, S., Smolic, A. and Sikora, T., "From 2D- to Stereo- to Multi-view Video**",** 3DTV-Conference, Kos Island, Greece, May 7-9, 2007.

[15] Knorr S., Kunter, M., and Sikora T.,"Stereoscopic 3D from 2D Video with Super-Resolution Capability", Signal Processing: Image Communication, Vol. 23, No. 9, pp. 665-676, October 2008.

[16] Imre, E., Knorr, S., Alatan, A.A. and Sikora, T., "Prioritized Sequential 3D Reconstruction in Video Sequences of Dynamic Scenes"**,** IEEE Int. Conf. on Image Processing (ICIP), Atlanta, USA, October 8-11, 2006.

[17] Torr, P.H.S., Fitzgibbon, A.W., and Zisserman, A.W., "Maintaining multiple motion model hypotheses over many views to recover matching and structure", Proc. IEEE Int. Conf. on Computer Vision (ICCV), pp. 485-491, 1998.

[18] Pollefeys, M., "Tutorial on 3D modeling from images", European Conf. on Computer Vision (ECCV), 2000.

[19] Farin, D., de With, P. H. N., and Effelsberg, W., "Robust background estimation for complex video sequences", IEEE Int. Conf. on Image Processing (ICIP'03), Barcelona, Spain, September 2003.

[20] Woods, A.,Docherty, T., Koch, R., "Image Distortions in Stereoscopic Video Systems", Proc. of SPIE: Stereoscopic Displays and Applications IV, San Jose, CA, USA, 1993.

[21] Fehn, C., "Depth-Image-Based Rendering, Compression and Transmission for a New Approach on 3D-TV", Proc. of SPIE: Stereoscopic Displays and Applications, San Jose, CA, USA, January 2004.

[22] Kauff, P., Atzpadin, N., Fehn, C., Müller, M., Schreer, O., Smolic, A., and Tanger, R., "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability", Signal Processing: Image Communication, Vol. 22, No, 2, pp. 217-234, Feb. 2007.

[23] Feldman, M. H., Lipton, L., "Interactive 2D to 3D Stereoscopic Image Synthesis", Proc. of SPIE-IS&T Electronic Imaging, Stereoscopic Displays and Virtual Reality Systems XII, San Jose, CA, USA, Jan. 2005.