AUTOMATIC 3DS CONVERSION OF HISTORICAL AERIAL PHOTOGRAPHS

F. Malapelle [1], *A. Hast*[2], *A. Fusiello*[1], *B. Rossi*[3], *P. Fragneto*[3] *and A. Marchetti*[4]

[1]Università degli Studi di Udine, DIEGM – Udine, Italy
[2]Uppsala University, IT Dept. – Uppsala, Sweden
[3]STMicroelectronics, AST Lab. – Agrate Brianza, Italy
[4]Consiglio Nazionale delle Ricerche, IIT – Pisa, Italy

ABSTRACT

In this paper we present a method for the generation of 3D stereo (3DS) pairs from sequences of historical aerial photographs. The goal of our work is to provide a stereoscopic display when the existing exposures are in a monocular sequence. Each input image is processed using its neighbours and a synthetic image is rendered, which, together with the original one, form a stereo pair. Promising results on real images taken from a historical photo archive are shown, that corroborate the viability of generating 3DS data from monocular footage.

***Index Terms*—** 3DS Conversion, View-synthesis, Image-based rendering, Historical aerial photography.

1. INTRODUCTION

Since the birth of modern aviation, aerial photos have been a rich source for understanding the historical development and geospatial changes. They offer a unique way to go back in time, exploring things as they were and therefore they have been used for aerial archaeology [2, 3, 4, 5]. An archive with several millions of such historical aerial photos is maintained by the Aerofototeca Nazionale (AFN) of the Italian Ministry of Cultural Heritage, in Rome. This archive portrays the Italian territory since the end of the nineteenth century, before its transformation due to the post-war reconstruction, the economic boom and changes by natural disasters such as earthquakes and floods.

During World War II stereoscopic images played an important role in the success of missions. Pilots from the photographic reconnaissance units took several consecutive photos over Europe that covered a long line of each flight route. A sample sequence of exposures is shown in Fig. 1. By meticulously photographing the ground it was possible for the photographic interpreters to visualize the area covered in 3D stereo (3DS), which can only be obtained if there is a substantial overlap between pairs of images.

Viewing aerial photos in 3DS is important, for the depth cue gives a much better understanding of the scene since the perceived depth information adds clues that are not available in a single exposure. This was important for the photographic interpreters, as it made possible to distinguish among objects such as houses, trees and the ground and especially estimating their height. Today, such stereo images can be a valuable tool for digital heritage research (e.g. extensions to projects such as [6] to also include stereo images).

In [7, 8] 3DS visualization is applied to historical aerial photographs with compelling results. These works concentrated on aspects such as pairs selection, geometric and illumination corrections in order to produce a 3DS display *with the available exposures*. This implies that the images must be viewed in such a way that one eye sees one exposure and the other eye sees the next one in line. In other words, the stereoscopic baseline (i.e., the line joining the two eyes) is parallel to the line of flight. This solution provides valid results for a static view, but falls short when trying to visualize the whole flight as a 3DS video. As a matter of fact, when moving along the line of flight it will be necessary, at a certain point, to switch from one stereo pair to the next in the sequence. However, there exists no such natural continuation between the stereo pairs, leading to a sudden switch between pairs that is perceptually disrupting.

The solution proposed in this paper is to generate a 3DS video from the monocular sequence using *view-synthesis* techniques. This entails that the baseline is *orthogonal* to the line of flight: one eye sees the existing stream of exposures, and the other one sees a synthetic stream of images, corresponding to a virtual eye displaced from the other one orthogonally w.r.t. the flight trajectory.

View-synthesis or *Image Based Rendering* is the generation of novel images as if they were captured from virtual viewpoints, starting from a set of actual images. Applications include the generation of a 3DS video from a monocular one [9, 10, 11, 12] and the upsampling of video sequences in order to achieve slow-motion effects (e.g., [13, 14, 15]). The rendering of virtual images requires some geometry information, either explicit (depth) or implicit (depth-proxies), and suitable warping functions. When cameras are *calibrated*, i.e. when both internal and external parameters are available, given the depth of an image point, it is straightforward

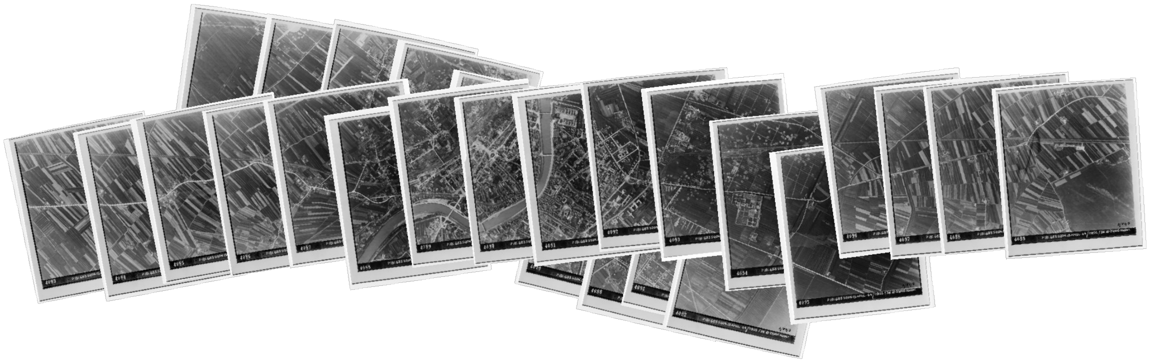

Fig. 1. An example of a sequence of exposures taken by the Royal Air Force (RAF) during WWII (from the AFN archive). The photos are placed in their relative positions by panoramic stitching, using [1].

to compute the position of the point in virtual image from any viewpoint. Techniques based on this paradigm have been extensively studied and several solutions are available in the literature ([16] and references therein).

When dealing with 3DS conversion of monocular footage, however, calibration data is hardly available, which makes the problem more challenging for several reasons. First of all, depth cannot be used, and suitable depth-proxies must be defined, together with proper warping functions based on fundamental matrices [17], trilinear tensors [18], or plane-parallax representation [19, 20]. Second, specifying the external *orientation* (position and attitude) of virtual views is unnatural, since they are embedded in a projective frame, linked to the Euclidean one by an *unknown* projective transformation. In [21, 22] an automatic method based on the planar parallax as a geometry proxy is presented: given two or more reference images, the possible uncalibrated orientations describe a 1-parametric trajectory obtained interpolating or extrapolating the relative motion among reference images.

The view synthesis approach, as opposed to using actual images for stereoscopic display as in [7], is more flexible, for the virtual viewpoint can be placed anywhere and illumination (or colour) is consistent by construction. The obvious drawback of our method is that the visual quality of the actual exposure is unmatched by any synthetic image, so, disregarding the temporal jittering, the stereoscopic display of [7] is more compelling. However our method provides a viable solution when the desired output is a *3DS video*.

The papers is structured as follows. In Sec. 2 we describe the method which consists of two main steps, the computation of depth-proxy map followed by the actual uncalibrated view-synthesis of the stereo image. Then in Sec. 3 we report and analyse some experimental results and at last in Sec. 4 we outline the conclusions and highlight possible future works.

2. PROPOSED METHOD

The aim of this procedure is the 3DS conversion of monocular aerial images, i.e. the generation of corresponding stereo images for a set of input exposures. Figure 2 shows a schematic representation of the algorithm: a triple of images is used to compute a disparity map referred to the central image which in turn allows to synthesize a stereoscopic pair. The procedure is repeated for every overlapping triplet of images to create a 3DS pair from every frame.

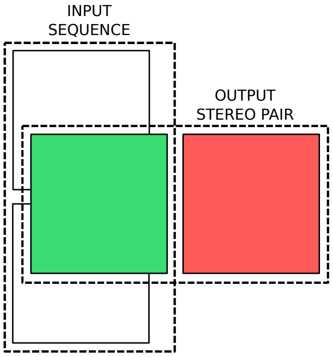

Fig. 2. Pictorial representation of the 3DS conversion for one image of the input sequence (green). The two neighbouring images are used to compute a disparity map referred to the central image which in turn allows to synthesize a stereoscopic "right" image (red).

For each input image two main steps are performed:

- computation of a parallax map using stereo matching
- rendering the novel stereo image through uncalibrated view-synthesis

2.1. Computing parallax

The key ingredient to perform the rendering of the synthetic image is planar parallax [20], a *depth-proxy* that can be profitably employed when camera parameters are not available. It can be considered as a projective surrogate of depth.

Given a pair of images (I_r, I_i) a parallax map can be computed using the following procedure:

Tie-points extraction and matching. SIFT are extracted in the two images and matched as described in [23].

Epipolar geometry. The fundamental matrix F_{ri} is estimated from the tie-points using RANSAC [24]; the epipole $\mathbf{e}_i$ of I_i is computed as the left zero-vector of F_{ri}.

Rectification. An uncalibrated rectification procedure [25] is employed which provides the rectifying homographies T_r and T_i and the infinite plane homography as $H_\Pi^{ri} = T_i^{-1} T_r$.

Stereo matching. Dense correspondences are obtained from a stereo matching algorithm. In the experiments we used the OpenCV/Matlab implementation of [26] with a left-right consistency check.

De-rectification. Dense correspondences are transferred back to the original reference images by applying the inverse of the rectifying homographies.

Parallax computation. Dense correspondences in the two images $(\mathbf{m}_r^k, \mathbf{m}_i^k)$ are related to each other by the following equation:

$$\mathbf{m}_i^k \simeq H_\Pi^{ri} \mathbf{m}_r^k + \mathbf{e}_i \gamma_r^k, \tag{1}$$

Where $(\mathbf{m}_r^k \mathbf{m}_i^k)$ are the correspondences known from the de-rectified stereo matching, H_Π^{ri} and $\mathbf{e}_i$ are obtained as described above and γ_r^k represents parallax. Its values can be obtained for each pixel of the reference frame by solving for γ_r^k in Eq. (1):

$$\gamma_r^k = \frac{(\mathbf{m}_i^k \times \mathbf{e}_i)^T (H_\Pi^{ri} \mathbf{m}_r^k \times \mathbf{m}_i^k)}{\left\|\mathbf{m}_i^k \times \mathbf{e}_i\right\|^2}. \tag{2}$$

It goes without saying that parallax can be computed only where correspondences can be established, hence in the area where the two images overlaps.

Let I_r be the *reference image*, i.e., the one for which we want to compute the parallax map. Let us consider a set of n auxiliary images I_i, where $i = 1 \dots n$, that depict a relevant portion of the same scene as I_r. Then we compute a parallax map for I_r by executing the steps described above for each pair (I_r, I_i). Each I_i will yield parallax values for a different portion of I_r.

In the sequences considered in this paper one image typically overlaps only with other two images, and these two does not overlap with each other, as represented in Fig. 3. Therefore, there is usually a central stripe in I_r where parallax cannot be computed.

This is why we introduced a hole-filling heuristic similar to the one presented in [27]. First we perform a local foreground/background segmentation: in the parallax map, for each unassigned pixel we compute the variance of its neighbourhood. A high variance indicates the presence of multiple depth layers, thus it is likely that the unassigned pixel is occluded by a foreground one in the conjugate image. Therefore, it is filled using the average of other background pixels in its neighbourhood. Otherwise, a low variance indicate a single depth layer and the pixel is filled accordingly using the average value of its neighbourhood.

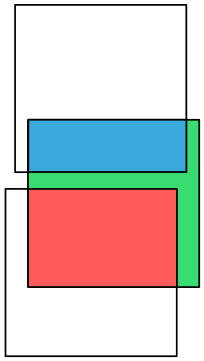

Fig. 3. Pictorial representation of the parallax integration. I_r is the image in the middle. The blue portion of the corresponding parallax map is computed when I_i is the image at the top, the red portion is computed when I_i is the image at the bottom, the green portion is computed with the hole-filling heuristic.

In principle, when sequences present a higher degree of overlap than those considered in this paper, it could be necessary to employ an integration procedure that allows to merge parallax values coming from different pair of images, as in [28].

2.2. View–Synthesis

Once a parallax map is computed, we generate the second element of the 3D stereo pair using the uncalibrated view-synthesis approach proposed in [29]. This section will only summarizes the main steps of the procedure; the actual method is more complex due to the need for addressing many practical aspects of image warping. Please refer to the original paper [29] for more details.

Trajectory generation. According to Eq. (1), the points transfer map from I_r to I_i is:

$$\mathbf{m}_i^k \simeq \begin{bmatrix} \mathbb{I} & 0 \end{bmatrix} D_{ri} \begin{bmatrix} \mathbf{m}_r^k \\ \gamma_r^k \end{bmatrix}, \tag{3}$$

where D_{ri} is the *uncalibrated rigid transformation matrix* between I_r and I_i which can be expressed as:

$$D_{ri} = \begin{bmatrix} H_\Pi^{ri} & \mathbf{e}_i \\ 0 & 1 \end{bmatrix}. \tag{4}$$

Arbitrary new images I_t can be synthesized by defining:

$$D_{rt} := e^{t \log D_{ri}} \quad t \in [0, 1]. \tag{5}$$

By continuously varying t a 1-parameter family of virtual cameras is obtained [21] which moves along a geodesic path interpolating the two reference cameras.

Virtual camera orientation. In the 3DS conversion application, on the contrary, the virtual camera position is alongside the actual one, i.e., outside its trajectory. The required parameters H_Π^{rv} and $\mathbf{e}_v$ can be specified as follows. Since there is no rotation between the reference image and the virtual one (images are coplanar), the infinite plane homography H_Π^{rv} is the identity matrix. As for the epipole, $\mathbf{e}_v = \begin{bmatrix} t & 0 & 0 \end{bmatrix}^T$ with $t \in \mathbb{R}^+$, defines a virtual viewpoint displaced along the direction of the rows of the image (assuming that the motion occurs roughly along the columns direction). Thus the orientation of the virtual camera can be computed with:

$$D_{rv} = \begin{bmatrix} 1 & 0 & 0 & t \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, t \in \mathbb{R}^+. \tag{6}$$

Forward mapping of parallax maps. Points in the original image are mapped forward, to the virtual image, according to

$$\mathbf{m}_v^k \simeq H_\Pi^{rv} \mathbf{m}_r^k + \mathbf{e}_v \gamma_v^k. \tag{7}$$

Using the new set of correspondences (m_r^k, m_v^k) we are able to compute parallax values γ_v^k for the virtual image, using Eq. (2) again.

Backward mapping of colour. This time the pixel grid of the virtual image is used as a reference to determine corresponding point in the original images through the value of γ_v^k. Points in the virtual image are mapped (backward), to points in the reference images to get a colour assigned using:

$$\mathbf{m}_r \simeq H_\Pi^{vr} \mathbf{m}_v + \mathbf{e}_r \gamma. \tag{8}$$

The formula is applied pixels-wise using the reference image as a source for pixel values.

3. RESULTS

In this section we report some experimental results obtained with our method applied to images taken from the AFN dataset that depict an aerial view of Pisa, Italy, during World War II (February 1944).

In Fig. 4 we show the output obtained with the scheme described in Section 2 applied to an image triplet. We can observe that the rendered image is geometrically correct, the illumination is the same as the reference one and it is visually plausible thanks to a very limited presence of artefacts. In Fig. 5 we report some details of the same output, in order to better appreciate the good quality of the synthetic image.

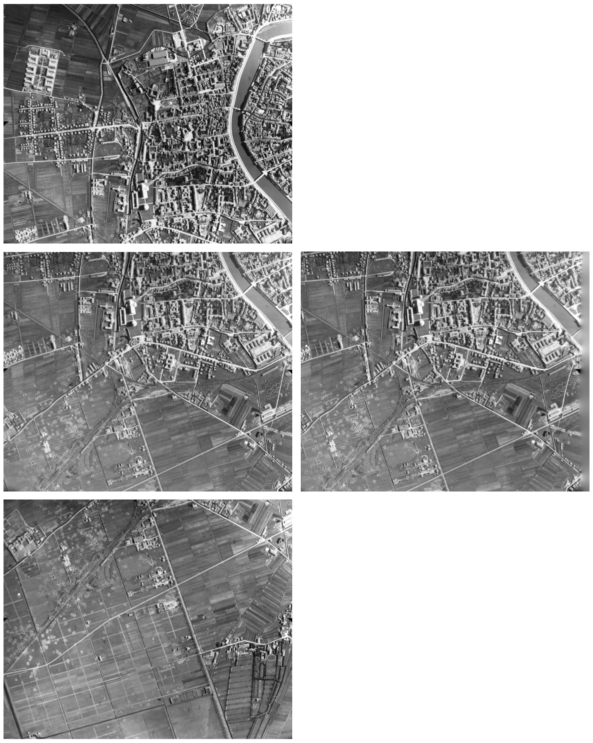

Fig. 4. On the left the three input images (4200×4900) , reference image is the one in the middle; on the right the rendered (synthetic) stereo image.

4. CONCLUSIONS

We presented a method for the conversion to 3DS of historical aerial photographs. The proposed solution overcomes some potential limitations of other techniques that use actual exposures for 3DS by rendering virtual images in an unconstrained fashion with respect to the flight trajectory.

The results are promising, however some issues are still to be addressed, such as stabilizing in time the virtual camera position, and upsampling the sequence in order to be able to play the 3DS video at a reasonable speed.

Acknowledgements

All the aerial images are property © of MiBAC-ICCD, Aerofototeca Nazionale, fondo RAF.

5. REFERENCES

[1] M. Brown and D. G. Lowe, “Automatic panoramic image stitching using invariant features,” *Int. Journal Comp. Vis.*, vol. 74, no. 1, pp. 59–73, 2007.

[2] J. Bourgeois and M. Meganck, *Aerial Photography and Archaeology 2003: A Century of Information*, Archaeological Reports. Academia Press, 2005.

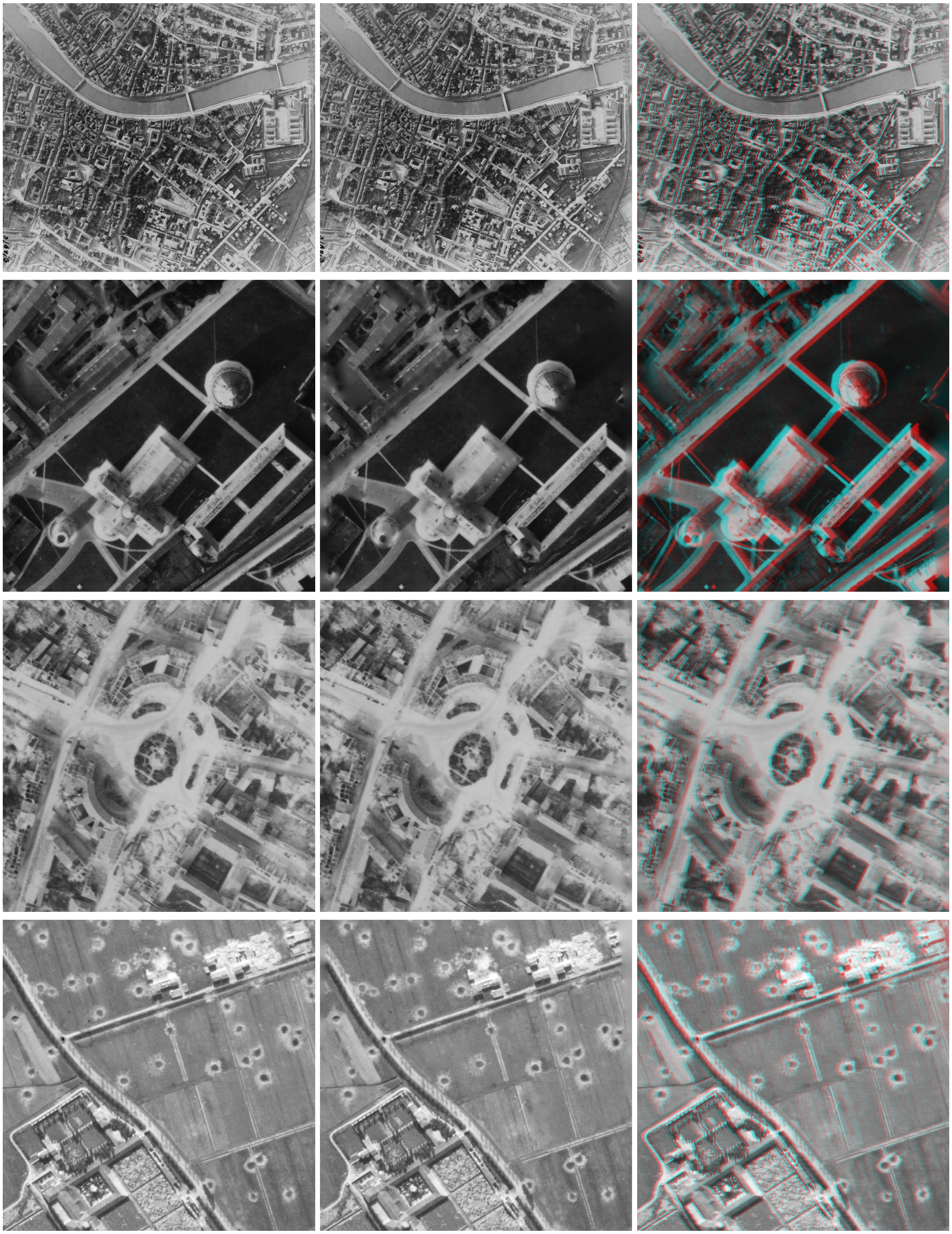

Fig. 5. A more detailed view of the results. From left to right: reference image, synthetic image, red-cyan anaglyph (to be viewed in colour with suitable glasses). The second row depicts *Piazza dei Miracoli* with the famous leaning tower.

[3] K. Brophy and D. Cowley, "From the air: Understanding aerial archaeology," *Scottish Archaeological Journal*, vol. 28, no. 2, pp. 159–160, 2006.

[4] D. N. Riley, *Air Photography and archaelogy*, University of Pennsylvania Press, 1987.

[5] D. R. Wilson, *Air Photo Interpretation for Archaeologists*, Tempus, 2000.

[6] M. Abrate, C. Bacciu, A. Hast, et al., "Geomemories: A platform for visualizing historical, environmental and geospatial changes in the italian landscape," *ISPRS International Journal of Geo-Information*, vol. 2, no. 2, pp. 432, 2013.

[7] A. Hast and A. Marchetti, "Towards automatic stereo pair extraction for 3D visualisation of historical aerial photographs," in *3D Imaging (IC3D), 2014 International Conference on*, 2014, pp. 1–8.

[8] A. Hast and A. Marchetti, "Stereo visualisation of historical aerial photos - a valuable digital heritage research tool," in *Digital Heritage*, 2015, pp. 1–4, Short Paper.

[9] G. Zhang, W. Hua, X. Qin, T.-T. Wong, and H. Bao, "Stereoscopic video synthesis from a monocular video," *IEEE Trans. on Vis. and Comp. Graph.*, vol. 13, no. 4, pp. 686–696, 2007.

[10] Y. J. Jung, A. Baik, J. Kim, and D. Park, "A novel 2D-to-3D conversion technique based on relative height-depth cue," in *IS&T/SPIE Electronic Imaging*, 2009, pp. 72371U–72371U.

[11] C.-C. Cheng, C.-T. Li, P.-S. Huang, et al., "A block-based 2D-to-3D conversion system with bilateral filter," in *Int. Conf. on Consumer Electronics*, 2009, pp. 1–2.

[12] L. Zhang, C. Vázquez, and S. Knorr, "3D-TV content creation: automatic 2D-to-3D video conversion," *IEEE Trans. on Broadcasting*, vol. 57, no. 2, pp. 372–383, 2011.

[13] T. Gurdan, M. R. Oswald, D. Gurdan, and D. Cremers, "Spatial and temporal interpolation of multi-view image sequences," in *Pattern Recognition*, pp. 305–316. 2014.

[14] B.-T. Choi, S.-H. Lee, and S.-J. Ko, "New frame rate up-conversion using bi-directional motion estimation," *IEEE Trans. on Consumer Electronics*, vol. 46, no. 3, pp. 603–609, 2000.

[15] S.-H. Lee, O. Kwon, and R.-H. Park, "Weighted-adaptive motion-compensated frame rate up-conversion," *IEEE Trans. on Consumer Electronics*, vol. 49, no. 3, pp. 485–492, 2003.

[16] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, *3D-TV System with Depth-image-based Rendering*, 2014.

[17] S. Laveau and O. Faugeras, "3-D scene representation as a collection of images and foundamental matrices," in *Proc. Int. Conf. Patt. Rec.*, 1994, vol. 1, pp. 689–691.

[18] S. Avidan and A. Shashua, "Novel view synthesis by cascading trilinear tensors," *IEEE Trans Vis. and Comp. Graph.*, vol. 4, no. 4, pp. 293–306, 1998.

[19] M. Irani and P. Anandan, "Parallax geometry of pairs of points for 3d scene analysis," in *Europ. Conf. Comp. Vis.*, pp. 17–30. 1996.

[20] A. Shashua and N. Navab, "Relative affine structure: Canonical model for 3D from 2D geometry and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 873–883, 1996.

[21] A. Fusiello, "Specifying virtual cameras in uncalibrated view synthesis," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 604–611, 2007.

[22] P. Fragneto, A. Fusiello, L. Magri, B. Rossi, and M. Ruffini, "Uncalibrated view synthesis with homography interpolation," in 2^{nd} *Joint 3DIM/3DPVT Conf.*, 2012, pp. 270–277.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[25] A. Fusiello and L. Irsara, "Quasi-euclidean epipolar rectification of uncalibrated images," *Machine Vis. and Appl.*, vol. 22, no. 4, pp. 663 – 670, 2011.

[26] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 328–341, 2008.

[27] G. Ramachandran and M. Rupp, "Multiview synthesis from stereo views," in *Int. Workshop. on Systems, Signals and Image Proc.*, 2012, pp. 341–345.

[28] F. Malapelle, A. Fusiello, B. Rossi, E. Piccinelli, and P. Fragneto, "Uncalibrated dynamic stereo using parallax," in *Proceedings of the 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2013, pp. 224–229.

[29] F. Malapelle, A. Fusiello, B. Rossi, and P. Fragneto, "Novel view-synthesis from multiple sources for conversion to 3DS," in *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, 2015, pp. 456–467.