2D TO 3D CONVERSION OF SPORTS CONTENT USING PANORAMAS

Lars Schnyder, Oliver Wang, Aljoscha Smolic

Disney Research Zurich,
Zurich, Switzerland

ABSTRACT

Given video from a single camera, conversion to two-view stereoscopic 3D is a challenging problem. We present a system to automatically create high quality stereoscopic video from monoscopic footage of field-based sports by exploiting context-specific priors, such as the ground plane, player size and known background. Our main contribution is a novel technique that constructs per-shot panoramas to ensure temporally consistent stereoscopic depth in video reconstructions. Players are rendered as billboards at correct depths on the ground plane. Our method uses additional sports priors to disambiguate segmentation artifacts and produce synthesized 3D shots that are in most cases, indistinguishable from stereoscopic ground truth footage.

***Index Terms*—** Stereo vision, 2D to 3D conversion, Mosaicing, Segmentation, Sports Visualization, 3D Reconstruction

1. INTRODUCTION

While stereoscopic 3D movies have been hugely successful in theaters for some time, 3D-at-home has only recently began to gain traction. One bottleneck inhibiting its adoption is that there is not yet a sufficient amount of suitable 3D content available and few live broadcasts are viewable in 3D. This is because the creation of stereoscopic content is still a very expensive and difficult process. Filming in 3D requires highly trained stereographers, expensive stereo rigs, and a redesign of existing monoscopic content work-flows. As a result, techniques for converting 2D content into 3D are a very important alternative, both for new productions as well as conversion of existing legacy footage.

The general problem of creating a high quality stereo pair from monoscopic input is highly under-constrained. The typical conversion pipeline consists of estimating the depth for each pixel, projecting them into a new view, and then filling in holes that appear around object boundaries. Each of these steps is difficult and, in the general case, requires large amounts of manual input, making it unsuitable for live broadcast. Existing automatic methods cannot guarantee the quality and reliability that are necessary for TV broadcast applications.

We focus on a specific application of 2D to 3D conversion which allows us to use domain-specific priors to automate the conversion process. One area that has traditionally been at the forefront of advances in technology (such as the adoption of HDTV), is sports. Sports games are a prime candidate for stereoscopic viewing, as they are extremely popular, and can benefit from the increased realism that stereoscopic viewing provides.

Our method takes advantage of prior knowledge, such as known field geometry and appearance, player heights, and orientation. We create a temporally consistent depth impression by reconstructing a background panorama with depth for each shot (a series of sequential frames belonging to the same camera) and modelling players as billboards.

Our contribution is a rapid, automatic, temporally stable and robust 2D to 3D conversion method that can be used for far-back field-based shots, which dominate viewing time in many sports. For low-angle, close up action, a small number of real 3D cameras can be used in conjunction with our method to provide full 3D viewing of a sporting event at reduced cost. For validation, we use our solution to convert one view from ground-truth, professional-quality recorded stereo sports footage, and provide visual comparisons between the two. In most cases, our results are visually indistinguishable from the ground-truth stereo.

2. RELATED WORK

Stereo content creation requires the use of a view synthesis technique to generate a second view close to the first. This is a difficult problem, as it requires knowledge of scene depth. Conversion methods must therefore either use some form of manual input, such as user specified normals, creases and silhouettes [1], manually traced objects at key-frames in a video [2] or a priori scene knowledge.

We are interested in automatic methods that use scene specific assumptions, as manual approaches are not scalable for converting large video libraries or broadcast content. Two automatic techniques that use priors are Make3D [3] and Photo Pop-up [4]. Such methods operate by; building statistical models of 3D scenes, segmenting images, and applying a classifier to assign planar relationships between segments. These approaches are limited in the types of scenes that are supported, and provide no temporal stability for video input.

Fortunately, for 2D to 3D conversion, we do not require a full 3D representation of the world to be reconstructed, since the distance of virtual viewpoints to be rendered is very limited. Instead, a depth or disparity map for a given input image is sufficient.

However, for conversion of video, the temporal stability of depth maps is very important in order to avoid noticeable temporal flickering artifacts. We achieve temporal stability through our novel concept of first constructing a background panorama for a given shot, and then creating a depth map for the entire sequence. Temporally stable depth maps for each frame are then reprojected from the panorama. We use a state-of-the-art video mosaicing approach to create our background panoramas.

Other existing methods of automatic stereo video conversion from single view video typically work by reconstructing a dense depth map using parallax between frames [5], or structure from motion [6]. However, these methods require static scenes and specific camera paths, and in cases where parallax does not exist in a video sequence, such as with a rotating camera, these methods would not work. Our method produces high quality automatic stereo conversion without assuming static content.

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

3. METHOD

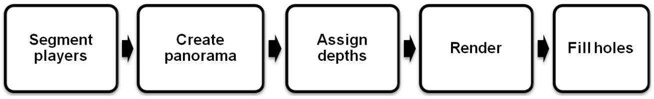


Fig. 1. Overview of our processing pipeline.

The basic idea of our approach is to separate static and dynamic parts of the scene and process them each using specific algorithms [7]. Our method works for wide field shots and exploits assumptions about the image content. The overview of the pipeline is shown Figure 1. First, each input image is segmented into static background and moving players. Then, a background panorama is constructed from the whole shot using a classical mosaicking approach, assuming a fixed rotating camera. From this, a depth map is created for the whole panorama using assumptions about the planar structure of the field, and a heuristic, but sufficiently accurate model for the background. Background depth maps for each frame can then be computed by an inverse projection from the panorama depth map using the previous homography. By design, these depth maps are temporally stable and consistent throughout the shot. Then, segmentation of the players is improved considering the background panorama, and each segmented player is represented as billboard with depth derived from its location on the background model. Ambiguities in segmentation are corrected so as to not cause noticeable artifacts, giving us a final depth map for each input image. Finally, stereo views are rendered with disocclusions inpainted from known background pixels. In the following sections we describe each of these steps in more detail.

3.1. Player segmentation

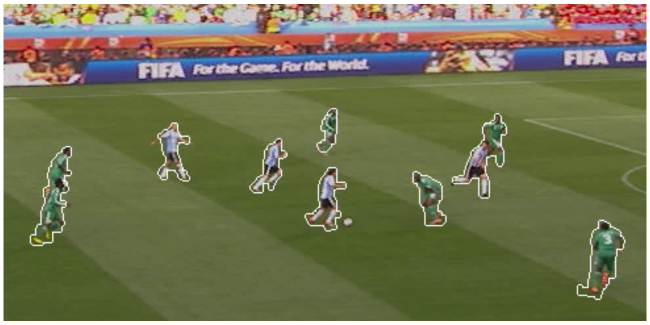


Fig. 2. Player segmentation. White lines are drawn around automatically detected player regions.

The first step of our process is segmentation into static and dynamic parts (Figure 2). For that, we use a standard support vector machine (SVM) [8] that is trained on a small database of field and player appearances. A vector of RGB colors in a small window around each pixel is used as a descriptor for the SVM. After the background panorama is created (see next section), player segmentation is refined, exploiting background information from the panorama. This approach worked well in our experiments, however, any more sophisticated segmentation may be substituted at this stage, as multiple object segmentation is an area of ongoing research, and development of segmentation was not the focus of this paper.

3.2. Background modeling using a video panorama

Temporal stability and consistency is one of the most important conditions for convincing stereo content generation. We therefore introduce a novel approach using a panorama to address these issues. Figure 3 illustrates the approach.

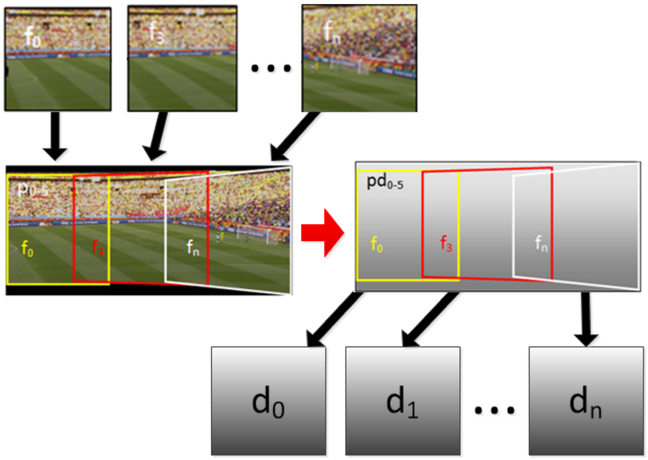


Fig. 3. Temporally consistent depth map generation. A panorama is created using video mosaicing. A depth map is created for the panorama yielding a complete background model. Depth maps for each frame are then extracted from the panorama by inverse homography projection.

We define $T_{i,j}$ as the homography transform (represented by a 3x3 matrix) that projects frame i onto the plane of frame j. Our process can then be summarized as follows:

1. Compute homographies $T_{i,i-1}$ using a method by Shi and Tomasi [9] and KLT feature tracking.
2. Create a panorama using the accumulated homography $T_{i,0} = T_{i-1,0} * T_{i,i-1}$, $T_{0,0} = I$. This is used to warp all images i onto the first image plane. We denote this homography $T_{i,0}$ as T_i.
3. Given this panorama, compute a consistent sequence-wide depth map (see below).
4. For a specific frame, transform the sequence depth map into its local coordinates using inverted homographies T_i^{-1}, giving us the background depth map for each frame.

We note that although our approach uses only frame-to-frame information (which leads to a small accumulated error over the whole sequence), it achieves sufficient quality for our application. This is because the panorama image is only used for the construction of per-frame temporally stable depth maps. We therefore found this method to be an optimal trade-off between running time and accuracy.

In step 3 of the process, a depth model of the panorama is created. This can be done with sophisticated automatic systems like Make3d, prior knowledge about stadium geometry, or even created by hand for each camera since they are usually stationary. However, in our experiments we choose a simple heuristic that produces perceptually high quality results. A linear depth ramp is assigned to the panorama which, in terms of geometry, is essentially approximating the model of the stadium background as a smooth upwards-curve, and is projected to the per-frame depth maps using T_i^{-1}. As alternative we experimented with modeling the exact plane location with user input, but found no perceivable difference in our results. Depth assignment to the panoramas also enables shot-wide restrictions to maximal (and minimal) disparity. This ensures that overall disparities remain in a comfortable range [10], which is harder to achieve with a frame-by-frame depth prediction approach. If camera zooms are applied during the sequence, our approach results in scene magni- and minification since we keep the assigned depths.

3.3. Player Billboards

Next, the depth values for segmented foreground players have to be assigned. This is done by assuming that the camera is vertically aligned and that players are in close contact with the ground. Players are then modeled as billboards whose depth is assigned from the per-frame depth map at the lowest point (in image space) of the segmented region.

This method makes the assumption that objects do not fly, which is violated by jumping players and (usually fast) moving balls. However, since we assume a far-field panoramic view, jumping players only affect the final disparity in a sub-pixel range which is hardly noticeable. Even more important for the moving balls, motion cues often override the disparity depth perception, which then results in a realistic depth impression for far field shots.

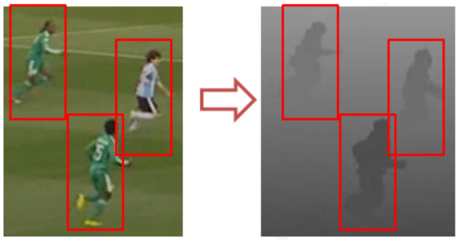


Fig. 4. Player Billboards. Each depth is assigned from per-frame depth maps.

In the optimal case, each player is segmented into its own region. However, multi-object segmentation is a difficult problem in the face of occlusions and changing object appearances. In addition, in our use case it is common for multiple players to group tightly together and move in similar directions, further confusing segmentation techniques. These errors cause the 3D artifact that players who are in such clusters and are higher up in the image plane, will have the appearance of floating over the players below them, as their assigned billboard depth does not correspond to their on-field location.

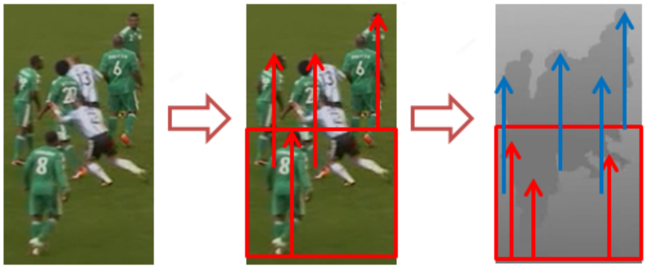


Fig. 5. An example of multiple players in the same billboard. Depth is modified for upper parts of the regions (blue arrows) taking average player height into account.

It is possible to alleviate these artifacts by again using application-specific priors (Figure 5). We first compute a per-frame estimation of player size by finding the average segmented region size in the reconstructed background panorama, and projecting this into each frame using T_i^{-1}.

Billboards bigger than this threshold are initially assigned the depth of the bottom player in the group, leading to the aforementioned players “floating-on-heads” effect. To solve this, we modify the billboard depth above the average player height. It is assumed that parts higher than this threshold belong to players further back. For these players a corresponding virtual foot position is computed according to the player size estimation (blue arrows in Figure 5). While players in front remain at the original assigned depth value, players behind are smoothly blended into the depth computed by their virtual foot positions.

Such billboard rendering is sufficient in our case given the limited distance of virtual views to be rendered and the limited player size. A more sophisticated approach is presented in [11] which allows for wide range free viewpoint navigation.

3.4. Stereo rendering

In order to render the images, we need to convert the final corrected depth values into pixel displacements.

One major advantage that our approach has over real stereo filming, is that we have total control over the parameters such as virtual interaxial camera distance and convergence for our synthesized stereoscopic content. This means that producers can easily optimize stereo parameters to minimize visual fatigue across scene cuts, create desired stereo effects for specific scenes, and place on-screen graphics at appropriate depth locations. Furthermore, stereoscopic errors that are hard to compensate for during live filming, such as objects breaking screen borders (causing stereo framing violations), can be completely avoided.

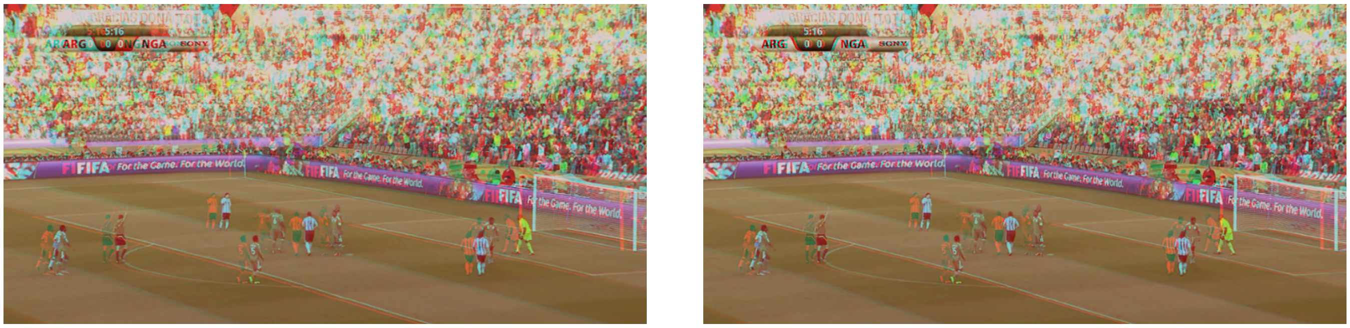
In our experiments we adjusted our system such that results best matched the ground truth footage. Once we decide on the desired virtual interaxial and convergence, stereo image creation follows standard depth image based rendering [12]. We project the single view into two views at each side so as to reduce the size of disoccluded holes in any one image.

To correctly render occluded regions we draw our images in depth-order. Disocclusions on the other hand lead to holes in the resulting virtual images. These holes can either be filled by background extrapolation around the billboard or from precomputed background information as described in [12]. Precomputed background would be available from our panorama. However, since in our case we do have very small disocclusions of few pixels for typical camera distances to players, these holes usually get very small/thin. Therefore they can be filled with simple background extrapolation (Figure 6 (c)).

4. RESULTS

We compare a stereo image created by our method from a single view, to the corresponding ground truth stereo image. These are shown as red-cyan anaglyph images in Figure 7. We note that while the anaglyph images are not identical, the depth impression is very similar. In addition we provide video results in comparison to ground truth stereo on our web-page[1], and strongly encourage readers to view our video results on a quality 3D display where possible.

[1] Supplementary material: http://zurich.disneyresearch.com/videodata/icip/



(a)

(b)

Fig. 6. We show our approach (a) using a linear panorama depth map converted from a single view of a ground truth stereo video (b).

While it is possible to notice some differences between footage, such as the graphic overlay, most viewers were not able to distinguish the two videos.

Our implementation computes multiple passes to create homographies, panorama and stereo frames. Running unoptimized research-quality code on a standard personal computer, we achieve per frame computations of 6-8 seconds. We note that this does not depend on the total video length, and it would be possible to run our algorithm on streaming footage given a short delay and increased processing power.

5. CONCLUSIONS AND FUTURE WORK

Our method is simple, robust, and produces convincing results, but still has some remaining limitations. For one, the segmentation method that we use currently operates mainly on a single frame, and does not use tracking information across sequences. While this step is easily replaceable, it is the main source of artifacts in our results, and a major topic of future work.

In addition, our depth assignment assumptions are specific to stadiums with a flat field and rising stadium seats. Other methods for depth map construction would have to be used for different terrain, such as golf courses. One area of future work could be to combine our method with other software or some minimum amount of manual interaction to generate depth maps appropriate to different background structure.

However, despite its simplicity, we have found that our method has sufficient accuracy for many cases. This is partially thanks to robustness in human stereo perception, where other cues, such as motion parallax, lighting, and known object size compensate for slight stereoscopic inaccuracies.

In conclusion, we have presented a method for creating stereoscopic footage from monoscopic input of wide field sports scenes. Static background and moving players are treated separately. For static background we propose a novel depth map from panorama approach to ensure temporal stability (which was the leading cause of visual artifacts in our studies). Moving players are treated as billboards. Our method creates high quality conversion results that are in most cases indistinguishable from ground truth stereo footage, and could provide significant cost reduction in the creation of stereoscopic 3D sports content for home viewing.

6. REFERENCES

[1] L. Zhang, G. Dugas-Phocion, J.-S. Samson, and S. M. Seitz, "Single view modeling of free-form scenes," in *CVPR (1)*, 2001, pp. 990–997.

[2] A. van den Hengel, A. R. Dick, T. Thormählen, B. Ward, and P. H. S. Torr, "Videotrace: rapid interactive scene modelling from video," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 86, 2007.

[3] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3-d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 824–840, May 2009.

[4] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM SIGGRAPH*, August 2005.

[5] G. Zhang, W. Hua, X. Qin, T.-T. Wong, and H. Bao, "Stereoscopic video synthesis from a monocular video," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 686–696, July-August 2007.

[6] S. Knorr, M. Kunter, and T. Sikora, "Super-resolution stereo- and multi-view synthesis from monocular video sequences," *Sixth International Conference on 3-D Digital Imaging and Modeling*, pp. 55–64, August 2007.

[7] K. Müller, A. Smolic, M. Drose, P. Voigt, and T. Wiegand, "3-d reconstruction of a dynamic environment with a fully calibrated background for traffic scenes," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 15, no. 4, pp. 538–549, 2005.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] J. Shi and C. Tomasi, "Good features to track," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1994, pp. 593 –600.

[10] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross, "Nonlinear disparity mapping for stereoscopic 3d," *ACM Trans. Graph.*, vol. 29, no. 3, pp. 10, 2010.

[11] M. Germann, A. Hornung, R. Keiser, R. Ziegler, S. Würmlin, and M. H. Gross, "Articulated billboards for video-based rendering," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 585–594, 2010.

[12] C. Fehn, "Depth-image-based rendering (dibr), compression and transmission for a new approach on 3d-tv," in *SPIE Stereoscopic Displays and Virtual Reality Systems XI*, January 2004, pp. 93–104.