Research White Paper

WHP 197

June 2011

Tools for 3D-TV Programme Production

Oliver Grau, **BBC R&D**
Marcus Müller, **Fraunhofer HHI**
Josef Kluger, **KUK Filmproduktion GmbH**

BRITISH BROADCASTING CORPORATION

White Paper WHP 197

Tools for 3D-TV Progamme Production

Oliver Grau, BBC R&D
Marcus Müller, Fraunhofer HHI
Josef Kluger, KUK Filmproduktion GmbH

Abstract

This contribution discusses tools for the production of 3D-TV programmes as developed and tested in the 3D4YOU project. The project looked in particular into image-plus-depth based formats and their integration into a 3D-TV production chain. This contribution focuses on requirements and production approaches for selected programme genres and describes examples of on-set and post-production tools for capture and generation of depth information.

This document was originally presented at the 3DTV-Conference 2011 16-18 May 2011 Antalya, Turkey.

Additional key words: multimedia systems, stereo vision, tv broadcasting

White Papers are distributed freely on request.

Authorisation of the Head of External Relations is required for publication.

TOOLS FOR 3D-TV PROGRAMME PRODUCTION

Oliver Grau

BBC R & D
56 Wood Lane
London, UK

Marcus Müller

Fraunhofer HHI
Einsteinufer 37
Berlin, Germany

Josef Kluger

KUK Filmproduktion GmbH
Spitzwegstrae 6
München, Germany

ABSTRACT

This contribution discusses tools for the production of 3D-TV programmes as developed and tested in the 3D4YOU project. The project looked in particular into image-plus-depth based formats and their integration into a 3D-TV production chain. This contribution focuses on requirements and production approaches for selected programme genres and describes examples of on-set and post-production tools for capture and generation of depth information.

Index Terms— Multimedia systems, Stereo vision, TV broadcasting.

1. INTRODUCTION

The prime concept of 3D-TV is to add a stereoscopic sensation to the television experience. The technological way to implement this is to provide different views to each eye of the viewer. This can be achieved in many ways. The simplest concept is still based on principles of stereo photography developed in the 19th century, in which two views of the scene with two horizontally offset cameras are captured and then presented individually to each eye of the viewer. This two-view stereo forms the basis for current implementation of '3D' in cinema and recently emerging TV services. Although two-view stereo is likely to be the representation of choice of the industry for some time to come, it has a number of limitations. These arise mainly out of the fact that the parameters of a two-camera capture rig have to be fixed during the capture and can neither be changed in post-production nor at the end-device. This limitation led to a number of research initiatives that looked into topics including depth-based and model-based stereo representations, e.g. [1]. These representations allow for adjustments in post-production and at the user side. Moreover, they enable usage of more advanced display techniques, in particular auto-stereoscopic and holographic displays.

This contribution describes new tools and production approaches for 3D-TV programme making. Further, it discusses some production requirements and findings of some production tests carried out in the 3D4YOU project [2].

Research in 3D-TV has a long history. A key requirement for 3D-TV was the change from analogue to digital TV services. Projects like DISTIMA [3] and MIRAGE [4] made some of the first experiments with electronic production of 3D content. However, almost all recent 3D-TV activities rely on the straightforward concept of an end-to-end stereoscopic video chain of two separate video streams, one for the left and one for the right eye. Due to these restrictions, stereo capture had to fit to the display geometry and vice versa. Display properties and viewing conditions as well as stereoscopic 3D production rules have to be taken into account by the production staff from the beginning during shooting.

A more flexible representation is to enrich video images with depth maps providing a Z-value for each pixel. This data representation is often called video-plus-depth. The capture and representation of a scene in video-plus-depth brings a number of challenges: When a scene is captured by a two-camera stereoscopic rig, it is impossible to compute an absolutely correct depth map for each view, because of occlusions and other model violation (like the lack of texture). Different capture and depth reconstruction techniques have been suggested. The ATTEST project investigated the use of an active time-of-flight sensor co-axial to one main camera [1]. Another configuration is to use a central, principal camera and two satellite cameras to reduce the effects of occlusions. Several capture configurations for 3D-TV productions have been investigated recently by the 3D4YOU project[2].

Many techniques for image-based (passive) depth reconstruction have been suggested in literature. In particular for stereo matching a huge body of work exists. The interested reader might be referred to [5]. This paper describes two methods, which were applied in the 3D4YOU project: A narrow-baseline stereoscopic depth estimation is outlined in section 3.2. Further, a wide-baseline method that is based on a visual hull computation is briefly described in section 3.3.

The remainder of this paper is organized in three main parts: The next section gives a brief outline of production requirements. Section 3 describes production tools. Section 4 describes some experiences made in test productions. The

This work has been supported by the FP7 3D4YOU and TSB iview.

paper finishes with some concluding remarks.

2. PRODUCTION REQUIREMENTS

Today the production of (stereoscopic) 3D (or S3D) is dominated by the movie industry with an increasing number of S3D movies made every year. The take-up in broadcast is still slow. The main reason lies in the required investments: At the production side new expensive production facilities are needed and households need new 3D-TV sets. Therefore, 3D-TV will be most likely be implemented first for high-budget genres, like premium football league and high-end drama or documentary.

In difference to feature films sport productions have usually a 'live' character. Even if not broadcasted live, the production times are very quick. For example summaries of events will be usually broadcasted on the same day. There are a number of other TV programme genres that are also produced 'live', for example news and a number of live-shows. The 'live-character' of these programmes would not allow methods that need, for example a long offline conversion process or manual intervention as typical in movie industry.

Another difference of TV production is that budgets are usually a number of magnitudes lower than movie productions. This applies to most TV-productions, like daily news, children, educational and most entertainment programmes. The lower budget implies that production methods must not involve expensive post-production work. In the past that meant that many TV productions were done mainly 'live' in order to minimize the post-production time and costs. Nowadays it is recognized that TV production is 'file-based' and therefore post-production are more widely adopted. Still expensive manual post-production processes can usually not be afforded. The implication of these requirements is that there is a demand for tools that enable cost-effective production and mainly automated (and therefore affordable) post-production.

Fig. 1. Two different rigs used for 'Peregrine Production', mirror rig (left) and mini rig (right) from KuK.

Different programme genres require also different technical set-ups. An obvious example is whether a production is done in the controlled environment of a studio or out-doors. Figure 1 shows two scenarios of a test production for a natural history programme. The image on the left shows the filming of a peregrine falcon in an out-door location. For these scenes a mirror rig with two full-size broadcast HD-cameras was used. The image on the right shows a studio set-up. In this scene a close-up of the head of the peregrine falcon was filmed. In this case the subject needed to be reproduced at a different scale, which basically requires a shorter base-line of the stereo camera. For this purpose a mini-rig was used.

3. PRODUCTION TOOLS FOR 3D-TV

3.1. On-Set Stereo Assistance Systems

To prevent eye strain and visual fatigue a number of rules have to be followed during a stereoscopic production. Main causes for visual discomfort are improper stereo geometry and retinal rivalry. Retinal rivalry can be caused by photometric distortions (e.g. different photometric properties of the cameras, difference in sharpness, brightness or contrast), geometric distortions (e.g. lens distortion, vertically misaligned left and right images, different focal lengths) or perception conflicts (e.g. stereoscopic window violation) [6].

Stereo geometry refers mainly to the inter-axial distance of the cameras and the position of the convergence plane. Both have to be adjusted carefully with respect to a given scene-structure and the target viewing conditions. Due to the accommodation-convergence conflict there is only a limited depth space close to the screen where comfortable 3D viewing is possible. Since this depth space is limited compared to the real world the stereographer has to 'to bring the whole real world inside this virtual space called the comfort zone'. Two parameters control the mapping from real world to the comfort zone: The inter-axial distance and the convergence plane. The inter-axial distance controls the depth volume of the reproduced scene while the convergence plane defines the position of this depth volume with respect to the screen plane.

Fig. 2. STAN as demonstrated at NAB 2009

Camera assistance systems have been developed to assist the stereographer on set to adjust all important parameters. One example is the stereoscopic analyzer (STAN) [6]. The core of STAN is a stereo analysis module where highly distinctive feature points are detected and matched in real-time between the two views. Most of the functionality is based on these feature points: First of all they are used to estimate the geometric relations between the two cameras. This information is used to give feedback about the actual rigging and to correct camera orientation and lens settings directly on set. Further, the camera parameters are used to remove any remaining vertical disparities or directly together with other

parameters on set to rectify the images on the fly for broadcast. Also the detection of near and far point which are needed to calculate the optimal inter-axial distance is based on these points and will be visualized in real-time for the set-up of cameras or can be directly used to control motorized rigs, as shown in figure 2.

3.2. Depth generation from stereo and multi-view

The main challenge of stereo and multi-view depth generation is to find the corresponding points in two images. Since this correspondence problem is under-constrained all approaches impose constraints, like uniqueness, smoothness and colour-compatibility. They vary by the way these constraints are exploited: *Local methods* use them only for a small neighborhood around the pixel of interest whereas *global methods* define the constraints in form of a cost function for the whole image, which is minimized by algorithms like Graph Cuts or Belief-Propagation (for an overview see [5]).

Although global methods outperform local methods for still images their computational complexity and their lack of temporal consistency make local methods - so far - the choice for 3D video processing. For example local window-based algorithms that use recursive matching in horizontal, vertical and temporal direction implicitly apply spatial and temporal smoothness constraints while enabling real-time performance. Typical errors of such window-based methods like foreground fattening and mismatches due to occlusion and low- or periodic texture are usually corrected during a post-processing step. These methods aim to align depth discontinuities to object borders, remove noise and miss-matches and to fill occlusions. Most of them either apply color segmentation and plane fitting [7] or use special filters to refine initial depth estimates at object borders and to smooth in homogeneous areas. The filter coefficients are calculated using the color image but applied to the depth image, a process often called cross-bilateral filtering. The depth image is smoothed in areas of constant colour while depth discontinuities are preserved. Figure 3 shows an initial depth map (2^{nd} row) and one refined by cross-bilateral filtering. Details of this processing can be found in [7].

Obviously there are occluded areas in the stereo-images that cannot be matched. With only two views these areas can only be filled by heuristics e.g. by extrapolating the background depth. Attaching satellite cameras to the rig enables the computation of depths in the occluded parts by stereo matching and provides additional geometric constraints like trifocal and quadrifocal consistency to resolve ambiguities.

Figure 3 shows some results of a multi-view approach that was used to process the data captured during a four camera test production within the 3D4YOU project. Only the results for camera 2 are shown, using adjacent cameras 1, 2 and 3. Camera 3 would be processed accordingly, using cameras 2, 3 and 4. First a pair-wise stereo-matching is performed for the neighboring cameras, yielding disparity map D_{21} pointing from camera 2 to camera 1 and D_{23} pointing from camera 2 to camera 3 (2^{nd} row of figure 4). After removing all inconsistent matches (3^{rd} row) the holes are mutually filled (bottom row left) before the disparity map is finally filtered with respect to the colour image of camera 2 using a median-variant of cross-bilateral filtering (bottom row right).

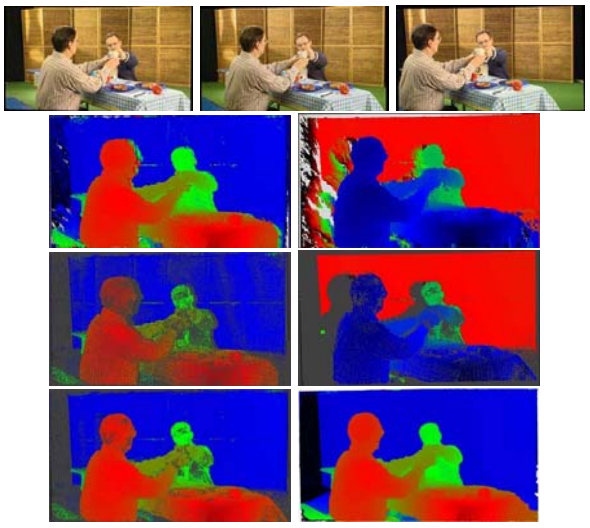

Fig. 3. Top: original images (left, middle, right), 2^{nd} row: original disparities (D_{21} left, D_{23} right), 3^{rd} row: consistent disparities (D_{21} left, D_{23} Right), bottom row: fused disparities D_{21} and D_{23} (left) and filtered disparities D_{21} (right)

3.3. Depth generation from wide baseline camera data

This work looks into real-time 3D modeling from wide-baseline multi-camera configurations for broadcast applications, including stereoscopic 3D and free-viewpoint video. It presents an alternative production method, which derives 3D stereoscopic content from the normal broadcast coverage cameras, plus optional locked-off cameras. The approach makes use of multi-camera 3D reconstruction techniques that have been previously developed for post-match analysis of sports scenes. These techniques perform an automatic camera calibration and segmentation of the foreground action. From this information a 3D model of the foreground action is computed. This data is then converted into a 3D-TV format, either as image-plus-depth or as a conventional S3D image pair for the left and right eye with a fixed inter-ocular distance.

The system consists of three main component blocks: The capture system is fully integrated into the production infrastructure and performs a live ingest of the multi-camera video streams. For our tests we implemented the capture system with standard IT server hardware equipped with HD-SDI frame-grabber cards.

The processing block automatically computes 3D infor-

mation from the multi-camera streams. The format conversion converts the image- and 3D-data into a 3DTV delivery format. The system has been applied to studio scenes and sport events. The processing performs a foreground segmentation, that separates the foreground action from the background, i.e. the pitch and other background objects like stands. In the studio cameras are usually locked-off and can be calibrated once before use. At sports events moving match cameras (pan, tilt and zoom) are used, which need to be calibrated every frame. Finally a 3D model is generated using a visual hull computation. Figure 4 illustrates an example scene and the generated depth map. More information on the processing can be found in [8].

Fig. 4. Rugby game (left) and synthesised depth (right).

4. PRODUCTION EXPERIMENTS

Within the 3D4YOU project a number of test productions have been made. The results shown in Figure 3 were part of a test in a studio environment. Scenes of different complexities and different depth were designed and captured with a specially designed rig, which combined a stereo camera pair, two side cameras and a time-of-flight sensor. The captured data then was used in the project to develop algorithms for the entire 3D-TV chain, including post-production conversion to image-plus-depth, coding and transmission and is partially made available to the public.

Another 3D4YOU production has been made alongside a BBC production. The 'story' was set around Peregrine falcons in London. The tests were made in a real production environment and various production stereo-camera rigs were used for the capture (see Figure 1). The challenge of the production was that a number of shots were aiming to capture the falcon in motion. Further, many scenes although filmed in a number of locations in the UK had then be combined with footage filmed in London. This required an intense post-production phase.

Many of the scenes use chroma key backgrounds behind the flying bird. Intensive chroma keying and generation of alpha mattes was necessary. For the depth map creation the generated alpha matte was a helpful tool to create occlusion layers and to properly separate foreground and background objects. For several shots that have already been produced in 2D by the BBC team a 2D-3D conversion was needed. Figure 5 shows an example of a scene and a composition with background.

Fig. 5. Original source material shot in full HD quality (left) and composited foreground and background (right).

5. CONCLUSIONS

This contribution discussed some tools for the production of 3D-TV programmes as developed and tested in the 3D4YOU project. Many programme genres in broadcast are produced live and with tight budgets. We have demonstrated a number of approaches that work completely automated or help the production team during the capture and improve efficiency.

The experience of test productions targeted to different programme genres have shown that it is impossible to develop a single capture system that can be used in all possible scenarios. The 3D4YOU project developed and tested a capture rig that can be used in the studio. For out-doors use conventional stereo-rigs and a wide-baseline set-up was demonstrated.

6. REFERENCES

[1] A. Redert and et al., "Attest advanced three-dimensional television systems technologies," in *Proc. of 3DPVT*, Padova, Italy, June 2002, pp. 313–319.

[2] Bogumil Bartczak and et al., "Display-independent 3d-tv production and delivery using the layered depth video format," *accepted for IEEE Trans. on Broadcasting*, 2011.

[3] M. Ziegler, "Digital stereoscopic imaging and applications. the race ii project distima," in *Proc. of IEE Colloquium on Stereoscopic Television*, 1992.

[4] C. Girdwood and P. Chiwy, "Mirage: An acts project in virtual production and stereoscopy," in *IBC Conference Publication*, Sept. 1996, number 428, pp. 155–160.

[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1, pp. 7–42, 2002.

[6] F. Zilly, M. Müller, and P. Kauff, "The stereoscopic analyzer - an image-based assistance tool for stereo shooting and 3d production," in *Proc. of ICIP 2010*, 2010.

[7] M. Müller, F. Zilly, and P. Kauff, "Adaptive cross-trilateral depth map filtering," in *Proceedings of the 3D-TV Conference*, Tampere, June 2010.

[8] Oliver Grau and et al., "A robust free-viewpoint video system for sport scenes," in *Proc. of 3DTV-Conference*, 2007.