

ESTIMATING 3D CAMERA MOTION FOR RENDERING AUDIO IN VIRTUAL SCENES

Werner Bailer*, Pau Arumí†, Toni Mateos†, Adan Garriga†, Jaume Durany†, David García†

* JOANNEUM RESEARCH, Steyrergasse 17, 8010 Graz, Austria

† BARCELONA MEDIA Innovation Centre, Diagonal 177, 08018 Barcelona, Spain

Keywords: digital cinema, calibration, tracking, ray-tracing.

Abstract

In the production phase of digital cinema content it is important to allow the director not only to *preview* the final rendered scene early in the shooting process, but also to *prehear* the 5.1 surround and HRTF-based binaural versions, thus enabling visual and auditive artistic decisions to be taken at the shooting stage. In many cases camera tracking data is not available for all cameras on the set (e.g. handheld ones) and thus the motion of the camera needs to be estimated. In this paper we describe an approach for estimation of 3D camera motion and its use for real-time audio rendering.

1 Audio Rendering

A real-time 3D-audio system was successfully implemented and tested in the IP-RACINE (<http://www.ipracine.org>) project. The system uses the positions of the actors/sound sources and the position and orientation of the camera/listener to accomplish two goals. First, it automatically dresses the dry audio with the acoustics of the 3D scene where it takes place, producing an exhibition-independent intermediate signal with the right reverberation. Second, it decodes and spatialises the resulting audio for playback in any given exhibition setup, so that sources appear to emit sound in accordance with their location in the virtual world.

The system runs in two phases. First, as soon as the 3D scene is known (well before the shooting stage), it performs an off-line intensive ray-tracing computation of the impulse responses (IR) between pairs of points on a grid that discretises the 3D scene. Each IR consists of four separate first order Ambisonics channels [3].

The second phase, which happens in real-time, uses the camera and source position parameters, either from data coming from mechanical tracking devices or from motion estimation as described below. The system, implemented on a series of CLAM [1] real-time data-flow networks, retrieves IR's corresponding to the sources and target positions, performs low-latency convolutions with the incoming audio, and smoothes IR transitions with suitably designed cross-fades. The use of Ambisonics technology provides the flexibility of fully exploiting any given surround or 3D exhibition system, from 3D HRTF-based binaural audio [2] to 5.1 or 22.2 surround setups. Multiple moving sound sources and listeners can be processed in real time in a normal CPU.

2 Camera Motion Estimation

We assume that the internal parameters of the camera are known with the exception of the focal length, which might change due to zooming. The estimation of dimensions in the scene can only be done up to scale and only the relative position of the camera can be estimated.

The task is thus to estimate the sequence of external parameters and the focal length of the camera. In contrast to structure from motion approaches our main focus is not a precise reconstruction of scene geometry, but a smooth coherent sequence of camera parameters. We use a Lucas-Kanade tracker to get point correspondences in subsequent frames. We apply the algorithm presented in [4] to the frames in a time window. If the time window is short we can regard local object motion as noise and the calibration method is sufficiently robust. If we perform the calibration for overlapping time windows we obtain different results for the projection matrix of a single frame as the reference coordinate system in each of the reconstructions is arbitrary. As the reconstructions are metric the different projection matrices for one frame are connected by affine transformations.

In order to chain the camera parameters for the individual frames we use the Expectation-Maximisation (EM) algorithm in a larger time window. The E-step consists of estimating affine transformations of the estimated projection matrices to the reference coordinate system, the M-step of finding new estimates for the sequence of parameters that are smooth and a maximum likelihood approximation of the reconstructions.

Acknowledgements

This work has been partially funded by the EC, contract FP7 215475 "2020 3D Media" (<http://www.20203dmedia.eu>).

References

- [1] X. Amatriain, P. Arumí, D. Garcia. A framework for efficient and rapid development of cross-platform audio applications, *ACM Multimedia Systems Journal*, (2007).
- [2] W. G. Gardner, K. D. Martin, HRTF measurements of a KEMAR, *JAES* 97(6), pp. 3907-3908, (1995).
- [3] M. Gerzon. Periphony: With-Height Sound Reproduction, *JAES* 21(1), pp. 2-10, (1973).
- [4] T. Svoboda, D. Martinec, T. Pajdla. A convenient multi-camera self calibration for virtual environments, *PRESENCE*, 14(4), pp. 407-422, (2005).