Scene Reconstruction for Storytelling in 360° Videos

Gonçalo Pinheiro[1(✉)], Nelson Alves[1], Luis Magalhães[2], Luís Agrellos[3], and Miguel Guevara[1]

[1] Centro de Computação Gráfica,
Campus de Azurém, Edifício 14, 4800-058 Guimarães, Portugal
goncalo.pinheiro@ccg.pt
[2] University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
[3] GMK, Cais das Pedras n°08, 4050-465 Porto, Portugal

Abstract. In immersive and interactive contents like 360-degrees videos the user has the control of the camera, which poses a challenge to the content producer since the user may look to where he wants. This paper presents the concept and first steps towards the development of a framework that provides a workflow for storytelling in 360-degrees videos. With the proposed framework it will be possible to connect a sound to a source and taking advantage of binaural audio it will help to redirect the user attention to where the content producer wants. To present this kind of audio, the scenario must be mapped/reconstructed so as to understand how the objects contained in it interfere with the sound waves propagation. The proposed system is capable of reconstructing the scenario from a stereoscopic, still or motion 360-degrees video when provided in an equirectangular projection. The system also incorporates a module that detects and tracks people, mapping their motion from the real world to the 3D world. In this document we describe all the technical decisions and implementations of the system. To the best of our knowledge, this system is the only that has shown the capability to reconstruct scenarios in a large variety of 360 footage and allows for the creation of binaural audio from that reconstruction.

Keywords: 360 videos · Storytelling · Scene Reconstruction · Binaural sound · Computer vision · Computer graphics · 3D reconstruction · People detection · People tracking

1 Introduction

With the rising popularity and accessibility of 360-degrees cameras, 360-degrees movies are bound to become increasingly common. Contrary to regular footage films, the director has no control as to where the spectator is looking. The action can occur while the viewer is unaware, as such a solution that redirects the spectator's attention without explicitly telling him where to look and maintaining the

P. Cortez et al. (Eds.): INTETAIN 2018, LNICST 273, pp. 117–126, 2019.
https://doi.org/10.1007/978-3-030-16447-8_12

in-scenario abstraction must be developed. To address this problem we propose the creation of binaural audio from conventional audio sources.

The use of binaural sound is described as good solution [10] but in order to be able to use it we must reconstruct all planes contained in the scene for modelling the room acoustics.

Focusing in this exact problem is the S3A spatial audio team proposing [14] where they present a block world reconstruction, from 360-degree stereo images, proving that planes and their materials are enough to render accurate spatial audio [15]. We follow the previously presented approach and simplify the process by estimating only the planes presented in the scene.

Persons are one important sound source on videos and usually are in motion. Since we intend to reconstruct the scenario and use binaural sound it makes sense to track persons in order to get their trajectory in the 360° video. Thus, we aim for a recording trajectories, which provides the system with the ability to associate a sound source to a person. This allows to take their motion into consideration when rendering the audio.

Our system's end goal is to provide the editor with the tools necessary to create binaural sound for arbitrary 360° videos. This toolkit will integrate a video editing software, as an add-on or pluggin, to take advantage from the editor's experience in their preferred software.

In this paper existing solutions for 3D reconstruction and people detection and tracking are reviewed and we present the initial thoughts and rough sketches of our system.

2 Related Work

Related to our work are the approaches that reconstruct scenarios, detect and track persons from videos or images. As such, a brief overview of the state of the art in 3D reconstruction, person detection and tracking is presented.

2.1 3D Scene Reconstruction

The 3D reconstruction of real objects or scenes is a topic that falls in the computer vision field. For that purpose, there are several algorithms and techniques in the literature that are highly dependent on the specific scenario of application. In the context of this work, it is important to understand the type of footage that can be fed to the system: (1) Single vs Stereo Videos - single perspective vs two records with a fixed baseline between the cameras and (2) Camera Motion: the camera may be still or in motion;

Stereo Reconstruction. A system capable of generating accurate dense 3d reconstructions from stereo sequences was developed by Geiger et al. [9]. This reconstruction pipeline combines a sparse feature matcher in conjunction with a robust visual odometry algorithm with efficient stereo matching and a multi-view linking scheme for generating consistent 3d point clouds. Kim and Hilton

[14] propose a block world reconstruction from spherical stereo image pairs. Before reconstructing, the spherical image is converted to a cubic projection for an easier facade alignment. Regions are also segmented in order to identify and reconstruct planes.

Structure from Motion. Several methods have been presented using structure from motion in tasks of reconstruction from moving videos [19,24,25,28]. Structure from motion has also been applied to 360 videos [29]. Some approaches refine the structure from motion using bundle adjustment for an accurate reconstruction [21,27]. Other processes also filter the reconstruction with a priori information and/or geographic references [3,21,30].

Simultaneous localization and mapping (SLAM) has also been used on reconstruction tasks in motion footage [13,32] and proved to work well in populated environments [20].

Depth Estimation from Single Image. When trying to reconstruct the scenario of a video with no camera motion the challenge is in the depth estimation task since it is impossible to triangulate the position of each pixel from different perspectives.

Using conventional computer vision techniques some methods were developed to estimate depth from single images or static videos and a posteriori reconstruct the scenario. Liu et al. [17] performed a semantic segmentation of the scene. Given the semantic context of the scene, depth is estimated considering a pixel or super-pixel at a time. Zhuo et al. [33] developed a method to estimate the structure of an indoor scenario from a single image. Local depth is estimated creating super-pixels for an easier extraction of uniform planes.

Some approaches have been made using deep learning and convolutional neural networks. Eigen et al. [7] presented a system capable of generate a depth map from a single image using two CNN's. The first estimates depth at general level and the second one does the estimation locally. Ewerth et al. [8] combined monocular depth clues and feature extraction feeding it to a ranking model. Chen et al. [6] used a RGB-D dataset and created a new dataset with the closest and farthest planes labelled to train an auto-encoder CNN to generate a depth map.

2.2 Person Detection and Tracking

The ability of detecting and tracking persons in videos, has long been of interest for the computer vision community specially driven by the automatic visual surveillance goal. Following, are some of the more relevant approaches described in the literature.

Using Conventional Computer Vision. Andriluka et al. [4] firstly detect pedestrians in real-world scenes using an object detection model that doesn't

consider any temporal constraints. To provide hypotheses for the position they propose a kinematic limb model. This grants the system expressiveness and robustness which reduces the number of false positives and facilitates detecting people in crowded scenes. Breuers et al. [5] developed three modules for addressing this problem in RGB-D images. The detection is made based on depth templates of upper bodies while the tracking is made using the MDL-tracker described by [12]. The third module is responsible for analyzing the head orientation and skeleton pose.

Using Deep Learning. Lin et al. [16] developed RetinaNet, a CNN for object and person detection. RetinaNet is composed by two sub networks, one responsible for feature extraction and one responsible for classification and drawing the bounding box. Tome et al. [26] reconstruct the 3D human pose from a single RGB image and the 2D joint estimation through a multi-stage CNN architecture. Guler et al. [11] mapped all pixels belonging to a person in a video using CNNs. Firstly, the whole body is estimated using classification and regression. After each body part is then located and a mesh drawn in all body pixels. Stewart et al. [23] present a new loss function which is applied to a classifier in order to identify the candidate bounding boxes of extracted representation generated by a CNN layer. Spinnelo et al. [22] uses AdaBoost for people classification training on acquired data considering that the appearance of people is highly variable. Instead of trying to segment the body in the different parts the detected body is segmented by height creating independent classifiers for each-one. Zhang et al. [31] presents a network for joint human detection and head pose estimation from rgb-d videos.

3 System Architecture

The system can be summed into five main modules presented in Fig. 1. Through analyzing an input video file, our system will allow for an editor to associate different sounds with their respective sources.

The first module is responsible for video pre-processing. The goal is to prepare the video to the subsequent processing steps. An example of a pre-processing method is to undistort the 360 video, from an equirectangular, fisheye or dual-fisheye format, to provide the next module with a standard video in an interpretable format for the computer vision methods that will be applied. This module is also capable of determining if the camera is moving or still which will determine the method used for reconstructing the footage.

The objective of the second module is to reconstruct the 3D model of a scene. For this purpose were identified three scenarios. If the camera is moving, it is possible to construct the point cloud using Structure from Motion techniques. In case of a still camera, depth has to be inferred before constructing the point cloud. In the stereoscopy case, depth of each feature point can be estimated due to the different perspectives of the scenario. In any case, a point cloud is created from which planes are extracted and the mesh of the scenario is generated.

Working in parallel with the 3D reconstruction module, the detection and tracking module is responsible for mapping the motion of the tracked persons into the generated spatial reconstruction. The trajectory will be represented by equally spaced points generating an approximation of the movement. In order to free the system from heavy processing the sound will only be mapped from a selected subset of all points identified. The goal is to automatically associate speech or sound to the detected person and regardless of their movement, the sound will always seem to follow its source.

The plane segmentation module will identify the main flat surfaces in the point cloud, generated by the reconstruction, which are the most significant for simulating acoustics.

The storyteller module will be responsible for the user interface, since it will have access to the results of all previous stages. Additionally, it is at this stage that the binaural sound is effectively rendered.

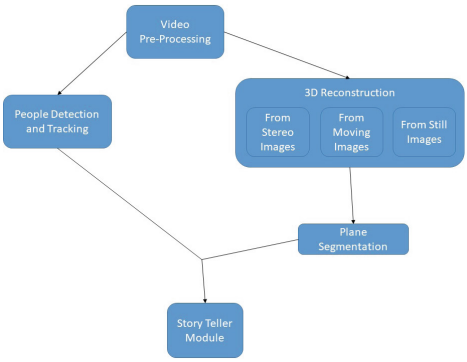

Fig. 1. System architecture

In following sections we specify all technical decisions made to date while developing the system.

3.1 Pre-processing

Motion Detection. Before reconstructing the scenario contained in a video it is necessary to understand the camera motion. This will decide which reconstruction method it's possible to apply.

For this task the Lucas-Kanade method for optical flow estimation was considered. This motion estimation is calculated frame by frame. Features are detected and distances are calculated between them in consecutive frames. Movement is assumed if the distance of a particular pair of features is bigger than one pixel. We consider that a whole frame is moving if most of the features are moving too.

360 Unwrapping. In order to make the video interpretable for the next steps, more specifically to the detection and tracking module it's necessary to remove the distortion contained in each frame.

From the 360-degree video, frame views are rendered at a specific vertical and horizontal field of view and resolution from an image in equirectangular projection using the software provided by [1].

3.2 Reconstruction

Motion. From videos where the camera is moving it is possible to reconstruct the scene by applying structure from motion. VisualSFM is being used to accomplish this task. Due to the fact that we are not performing keyframe selection, the computation is naturally heavier.

Still Camera. To reconstruct the scenario from a video with no camera motion we are using the PlaneNet DNN proposed by Liu et al. [18]. This system is capable of reconstructing planes from a single image which can be compared to reconstructing from a frame of a video filmed with a still camera.

Stereo. Using stereo footage we have two perspectives of the same scenario, so it is possible to infer the depth of each pixel by creating a disparity map.

Disparity maps were created using block matching algorithm, the Semi-Global Block Matching Stereo Correspondence Algorithm, the Stereo Belief Propagation and the Stereo Constant Space Belief Propagation. Visually, the Stereo Constant Space Belief Propagation algorithm performed better when compared to the others methods.

Having the disparity map and the camera's intrinsic parameters, it is possible to re-project that disparity to 3D space using the OpenCV library.

3.3 Person Detection and Tracking

Detection. The detection phase has the goal of giving the initial bounding box to the tracking algorithm as we try to automate the whole detection and tracking process. This process is made using the RetinaNet proposed by Lin et al. [16]. For each detected person a bounding box is generated.

Tracking. Having the RetinaNet bounding box we resize it to 20% of the original size to guarantee that the pixels inside of the bounding box belong exclusively to the tracked person. To accomplish this task we tested several algorithms like BOOSTING, Multiple Instance Learning (MIL), Kernelized Correlation Filters, tracking, learning and detection, MEDIANFLOW, GOTURN, MOSSE and the Discriminative Correlation Filter with Channel and Spatial Reliability (DCF-CSR). In our testing samples the DCF-CSR algorithm was the most robust under different light conditions and re-tracking occluded persons.

4 Preliminary Results

In this section we present the results of the work realized to date. All the results described were measured qualitatively with direct observation.

The system is already able to detect if there is any camera motion, showing some instability when performing this task on a video with a moving crowd. In addition, vertical and horizontal stereoscopy are also identified.

The 360 unwarpping proved to work well when removing the distortion from a frame. In Fig. 2 it is presented the full 360° frame. The Fig. 3 is rendered at the center of the original frame and the Fig. 4 at 90°. It is possible to observe that both figures present no distortion.

Fig. 2. 360-degree frame

Fig. 3. Image rendered at 0°

Fig. 4. Image rendered at 90°

From the pair of stereo Figs. 5 and 6 it was possible to reconstruct the dense point cloud presented in the Fig. 7.

Fig. 5. Left image

Fig. 6. Right image

Fig. 7. Reconstructed mesh

As described previously, the detection performed well in different light conditions as we can note in Figs. 8 and 9.

Fig. 8. Test 1

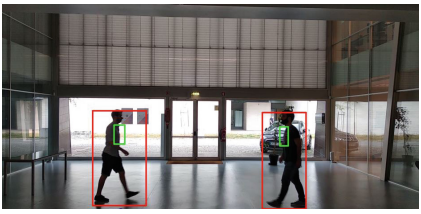

Fig. 9. Test 2

5 Conclusions and Future Work

In this paper, we have proposed an architecture for an automatic system, which provides the necessary scene information for rendering binaural audio. This is the second system of its kind to be documented to date, after the one proposed by S3A spatial audio [2]. Despite the previously presented approach, that reconstructs only from stereo footage, our system has shown the capability to reconstruct from a variety of 360-degree videos. To complement the process, a person detection and tracking method is integrated for associating speech to its source.

For future work, the first task to address is the selection of keyframes in order to reduce the processing needed for the reconstruction. The implementation of our own Structure from Motion system is also considered.

In order to give more robustness to the tracking algorithm, the inference of the inner bounding box must be refined. In addition, a module for plane segmentation must be developed. The final task is the integration of all these modules into one consolidated software, so as to provide binaural audio for storytelling.

Acknowledgments. This article is a result of the project CHIC - Cooperative Holistic view on Internet and Content (project n° 24498), supported by the European Regional Development Fund (ERDF), through the Competitiveness and Internationalization Operational Program (COMPETE 2020) under the PORTUGAL 2020 Partnership Agreement.

References

1. 360-degree projection. https://github.com/bingsyslab/360projection. Accessed 11 June 2018
2. S3A spatial audio. http://www.s3a-spatialaudio.org. Accessed 24 July 2018
3. Akbarzadeh, A., et al.: Towards urban 3D reconstruction from video. In: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT 2006). IEEE Computer Society (2006)
4. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2008. https://doi.org/10.1109/CVPR.2008.4587583

5. Breuers, S., Beyer, L., Rafi, U., Leibe, B.: Detection-tracking for efficient person analysis: the DetTA pipeline. CoRR abs/1804.10134 (2018). http://arxiv.org/abs/1804.10134
6. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. CoRR abs/1604.03901 (2016). http://arxiv.org/abs/1604.03901
7. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. CoRR abs/1406.2283 (2014). http://arxiv.org/abs/1406.2283
8. Ewerth, R., et al.: Estimating relative depth in single images via rankboost. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 919–924, July 2017. https://doi.org/10.1109/ICME.2017.8019434
9. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3D reconstruction in real-time. In: 2011 IEEE Intelligent Vehicles Symposium (IV), pp. 963–968, June 2011. https://doi.org/10.1109/IVS.2011.5940405
10. Grani, F., et al.: Audio-visual attractors for capturing attention to the screens when walking in cave systems. In: 2014 IEEE VR Workshop: Sonic Interaction in Virtual Environments (SIVE), pp. 3–6, March 2014. https://doi.org/10.1109/SIVE.2014.7006282
11. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: dense human pose estimation in the wild. CoRR abs/1802.00434 (2018). http://arxiv.org/abs/1802.00434
12. Jafari, O.H., Mitzel, D., Leibe, B.: Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 5636–5643, May 2014. https://doi.org/10.1109/ICRA.2014.6907688
13. Kim, A., Eustice, R.M.: Active visual slam for robotic area coverage: theory and experiment. Int. J. Robot. Res. **34**(4–5), 457–475 (2015). https://doi.org/10.1177/0278364914547893
14. Kim, H., Hilton, A.: Block world reconstruction from spherical stereo image pairs. Comput. Vis. Image Underst. **139**, 104–121 (2015). https://doi.org/10.1016/j.cviu.2015.04.001. http://www.sciencedirect.com/science/article/pii/S1077314215000831
15. Kim, H., et al.: Acoustic room modelling using a spherical camera for reverberant spatial audio objects. In: Audio Engineering Society Convention 142, May 2017. http://www.aes.org/e-lib/browse.cfm?elib=18583
16. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR abs/1708.02002 (2017). http://arxiv.org/abs/1708.02002
17. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1253–1260, June 2010. https://doi.org/10.1109/CVPR.2010.5539823
18. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: PlaneNet: piece-wise planar reconstruction from a single RGB image. CoRR abs/1804.06278 (2018). http://arxiv.org/abs/1804.06278
19. Polic, M., Förstner, W., Pajdla, T.: Fast and accurate camera covariance computation for large 3D reconstruction (2018)
20. Riazuelo, L., Montano, L., Montiel, J.M.M.: Semantic visual SLAM in populated environments. In: 2017 European Conference on Mobile Robots (ECMR), pp. 1–7, Sept 2017. https://doi.org/10.1109/ECMR.2017.8098697
21. Saurer, O., Pollefeys, M., Hee Lee, G.: Sparse to dense 3D reconstruction from rolling shutter images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3337–3345 (2016)

22. Spinello, L., Arras, K.O., Triebel, R., Siegwart, R.: A layered approach to people detection in 3D range data. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, pp. 1625–1630. AAAI Press (2010). http://dl.acm.org/citation.cfm?id=2898607.2898866
23. Stewart, R., Andriluka, M., Ng, A.Y.: End-to-end people detection in crowded scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
24. Sturm, P., Triggs, B.: A factorization based algorithm for multi-image projective structure and motion. In: Buxton, B., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1065, pp. 709–720. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61123-1_183
25. Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A.: Hierarchical structure-and-motion recovery from uncalibrated images. Comput. Vis. Image Underst. **140**, 127–143 (2015). https://doi.org/10.1016/j.cviu.2015.05.011. http://www.sciencedirect.com/science/article/pii/S1077314215001228
26. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: convolutional 3D pose estimation from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
27. Wong, K.H., Chang, M.M.Y.: 3D model reconstruction by constrained bundle adjustment. In: 2004 Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 902–905, Aug 2004. https://doi.org/10.1109/ICPR.2004.1334674
28. Yu, R., Russell, C., Campbell, N.D.F., Agapito, L.: Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video. In: The IEEE International Conference on Computer Vision (ICCV), December 2015
29. Yu, S., Lhuillier, M.: Incremental reconstruction of manifold surface from sparse visual mapping. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission, pp. 293–300, October 2012. https://doi.org/10.1109/3DIMPVT.2012.11
30. Zakharov, A.A., Barinov, A.E.: An algorithm for 3D-object reconstruction from video using stereo correspondences. Pattern Recogn. Image Anal. **25**(1), 117–121 (2015). https://doi.org/10.1134/S1054661815010228
31. Zhang, G., Liu, J., Li, H., Chen, Y.Q., Davis, L.S.: Joint human detection and head pose estimation via multistream networks for RGB-D videos. IEEE Signal Process. Lett. **24**(11), 1666–1670 (2017). https://doi.org/10.1109/LSP.2017.2731952
32. Zhou, H., Zou, D., Pei, L., Ying, R., Liu, P., Yu, W.: StructSLAM: visual SLAM with building structure lines. IEEE Trans. Veh. Technol. **64**(4), 1364–1375 (2015). https://doi.org/10.1109/TVT.2015.2388780
33. Zhuo, W., Salzmann, M., He, X., Liu, M.: Indoor scene structure analysis for single image depth estimation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 614–622, June 2015. https://doi.org/10.1109/CVPR.2015.7298660