Environment Capture and Simulation for UAV Cinematography Planning and Training

Stephen Boyle, Matt Newton, Fan Zhang and David R. Bull
Department of Electrical and Electronic Engineering
University of Bristol
Bristol, BS8 1UB, UK
{stephen.boyle,mn15359,fan.zhang,dave.bull}@bristol.ac.uk

***Abstract*—This paper presents a workflow for the generation of environmental models, which can be employed for training and planning of drone based shooting. This converts multiple 2D environmental and terrain images into 3D models and height maps, which are then imported into a simulation engine as environment assets. A subjective study has also been conducted to characterise the relationship between reconstruction quality and the number and location of input images for various environmental scenarios. Using this workflow, demonstration videos have been produced which combine extracted environments with an existing object model (cyclist) in simulation. These illustrate its utility in shot planning, rehearsal and training.**

***Index Terms*—UAV, simulation, 3D reconstruction**

I. Introduction

Due to their greater flexibility and lower cost, drones are becoming increasingly popular as camera platforms in film and broadcast production, replacing both dollies and helicopters. Providing multiple angles, flexible camera positioning and uninterrupted coverage, the use of drones in media production can significantly improve the viewing experience. Notable examples of drone use include the opening scene of *James Bond - Skyfall* (2012) [1], where drone-mounted cameras were employed to shoot a motorbike chase on the rooftops of a bazaar in Istanbul, and the broadcasting coverage using UAVs (unmanned aerial vehicles) for 2014 Winter Olympics in Sochi [2, 3] and for the Summer Olympics of 2016 in Rio, Brazil [4].

When drones are employed to cover live events such as sports, efficient operations and significant planning are essential if the viewing experience is to be maximised. This is complicated by the fact that directors, drone pilots and camera operators must react to unpredictable events. In this context, a flexible, reliable and realistic simulation tool would be of significant utility for planning, rehearsing, training and evaluating single or multiple drone operations for such scenarios. Such a tool would enhance productivity, improve safety, and ultimately increase the quality of the shots delivered to viewers. Forward planning of drone flights would also support building safety margins related to nearby buildings, crowds and other features, into drone operations. Camera shots could also be designed to provide suitable coverage of any landmarks which need to appear in-shot. Rehearsals could take account of the dynamics of the event, allowing exploration of multiple varied scenarios. Finally, for the case of autonomous drone operations, such a tool would enable the production of fight plans for the drone or drones deployed.

To achieve UAV flight planning, there are a number of commercial and royalty-free software packages available, including DJIFlightPlanner [5], Drone Harmony [6] and UgCS [7]. However, most of these packages cannot support 3D rendering of specific environments for realistic simulation. More recently, Google has developed a browser-based animation tool, Google Earth Studio [8], based on the massive 2D and 3D data of Google Earth, which can generate still and animated footage for actual environments at different viewing angles and positions. However this does not support the integration of object 3D models needed for simulating UAV shooting of dynamically changing scenarios and shot types.

In this paper, a workflow is proposed for capturing and simulating actual target environments. 3D assets are reconstructed from multiple input images and height maps are obtained from open-source terrain images. These are then imported into a simulation engine and integrated with object models for further simulation. To investigate the relationship between the number of input images and reconstruction quality, a subjective experiment was conducted on three different scenarios identifying the required input image numbers for acceptable reconstruction quality. Example environmental models have been generated using this workflow based on open source environmental and terrain images, which demonstrate its application in UAV based shoot planning and evaluation.

The rest of this paper is organised as follows. Section II presents the proposed workflow and its key steps for drone environment capture and simulation. The experiment conducted to evaluate the number of input images with respect to reconstruction quality is then described in Section III, while the experimental results and examples of two reconstructed environments are provided and discussed in Section IV. Finally, Section V concludes the paper and outlines future work.

II. Proposed workflow

The proposed workflow for generating background environments is shown in Fig. 1, in which there are four primary processing steps: (i) create environmental images, (ii) reconstruct 3D model, (iii) generate terrain data, and (iv) import into simulation engine.

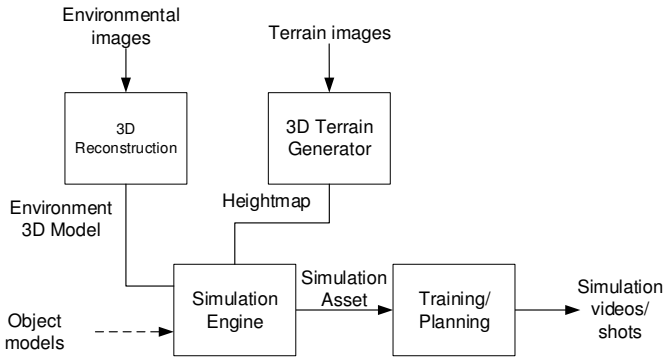


Fig. 1: Diagrammatic illustration of the proposed workflow.

A. Creation of environment images

Continuous and high quality environment images provide the basis for 3D reconstruction, and ultimately determine the quality of the reconstructed models. Environment images could, for example, be shot at the actual event site, possibly from an aircraft or a drone. This is however expensive and time consuming. Alternatively they can be generated by photo-scanning from different viewing points using software such as Google Earth [9]. This has the advantage of low cost, flexibility and time, and has been selected as the approach of choice in this work[1].

Fig. 2 shows examples of scanned environment images from Google Earth, which were shot at regular, short intervals (e.g. 0.25s) during navigation around a roundabout using a screen capture program, ScreenToGif [10]. Images were produced whilst orbiting the target background at a series of heights so that each part of the environment was captured from a number of different angles.

Fig. 2: Examples of manually scanned environment images from Google Earth for a roundabout scenario using ScreenToGif.

It should be noted that, in these scanned images, there are dynamic objects (such as cars and pedestrians), which may be present at various locations. This can cause significant distortions in reconstruction due to image alignment failure. The current solution in this work is to remove those 2D source images with dynamic objects. More advanced approaches, such as manually/automatically concealing them based on neighbouring background textures, will be the subject of future work.

[1]It should be noted that images from software such as Google Earth exhibits inconsistent quality for some locations, especially those in rural areas. In these cases, real environmental images may need to be captured to provide reliable sources with better reconstruction quality.

B. 3D environment reconstruction

Numerous 3D reconstruction software packages exist which can produce 3D models from 2D images. Notable ones, including Autodesk ReCap [11] and 3DF Zephyr [12], have been evaluated for this work. It was found that the former produced relatively poor results from Google Earth imagery. 3DF Zephyr Aerial Photogrammetry software was found to generate 3D models with improved quality, although it does require local graphic calculation capability for processing a large number of 2D images. For the current work, 3DF Zephyr was adopted as the reconstruction software to generate 3D environmental models.

During reconstruction, the default automated Zephyr workflow [13] was used, creating a dense point-cloud, then a mesh from the point-cloud and finally a textured mesh. To reduce the distorted area and the number of textures needed, these individual stages were run manually. After the dense point-cloud generation, the point-cloud was edited to remove areas with distortion, areas of low point density or areas which were not required (e.g far from the location of interest). The mesh and textured mesh generation stages were then run manually in turn. This resulted in a significantly smaller area for the final mesh and in many cases the model could be exported to a file having a single texture. The 3DF Zephyr program was then used to generate a 3D model and to export it as an FBX file for further integration.

C. Generating terrain data

There are three primary ways of obtaining terrain data: (i) satellite (and aerial) imagery, e.g. OpenTopography.org [14] and the USGS EarthExplorer [15], (ii) LIDAR (Light Detection and Ranging) [16] and (iii) 3D scanning techniques (e.g. 3D Laser Scanning).

In this work, based on the consideration of source data availability and cost, the tool available on the OpenTopography.org [14] website was used to select areas for certain locations. The SRTM (Shuttle Radar Topography Mission) GL1 (Global 1) data-set (having a 30m resolution) was employed. The data was output in GeoTiff (16 bit grey-scale TIFF) format and the option to generate hill-shade images from the DEM (digital elevation model) data was selected. The TIFF file was imported into Bundysoft L3DT software (Large 3D Terrain Generator) [17] as a height map. Using this software, the height map was resized and exported as a 16 bit grey-scale PNG format file.

D. Importing data into the simulation engine

In drone cinematography, simulation engines can be employed to design scenarios with specific camera/drone parameters, with the flexibility over the choice of background and foreground targets(s) and actions, providing a much lower cost solution compared to using real drone(s) at actual sites. There are many existing excellent simulation platforms including Unity [18], GameMaker [19] and Unreal Engine (UE4) [20].

After comparing the features of various simulation engines, Unreal Engine (UE4) was selected for this work due to its

relative ease of use, its well maintained community support and its widespread use in the film industry. The control interface AirSim [21], which is based on UE4, can also be used for real-time interactive applications (e.g. training).

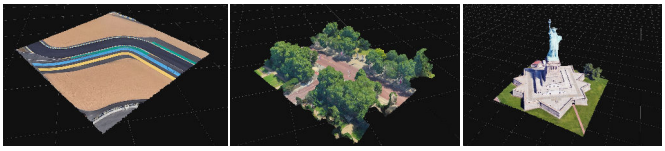
The reconstructed environment models, saved in standard 3D model format (.fbx), are imported into UE4 with corresponding height maps (imported through the Landscape Editor of UE4). The resulting environmental assets can then be combined with object models (e.g. cyclists in our demos) for simulation.

III. Subjective study on the number of input environment images

In order to reconstruct high quality 3D environmental models, many input 2D images, captured from different viewing angles, are needed. Reconstruction may therefore be time consuming, especially when the background area is large. To realise efficient data processing, it is important to understand how many images are "optimal" (to achieve a sufficiently high quality reconstruction) for a certain size of area.

A. Test content

In this study, three scenarios (source from Google Earth) of varying complexities were identified. All have the same test area size, 10000m^2 (100m$\times$100m), for simplicity and for easy tessellation in projects of a larger size. The three areas considered were London (The Mall), Le Mans (Section of Track) and New York (Statue of Liberty).



(a) Le Mans (b) London (c) New York

Fig. 3: Screen shots of the three reconstructed test scenarios when they were integrated into UE4 (without combining with height maps and 240 input images used).

Examples of the three test areas are shown in Fig. 3. Among these, Le Mans represents an area of low complexity with few obtruding objects, a relatively simple mesh, and basic texture. London is an example of a high complexity area with large amounts of foliage. This creates complex mesh structures, as well as obstructing interior objects from lower camera angles. The foliage and smaller structures also require a more complex texture to be created. Finally, New York is an example of a more object-based situation, similar to a scenario with a building or bridge as the focus. The small detailing in arm and crown regions create difficulty whilst the rest of the structure is relatively simple.

Reconstructions with six different numbers (240, 180, 120, 60, 30 and 15) of input images were built for each test scenario using the workflow described in section II. A twenty second HD (1920$\times$1080) free flying shot [22] was then generated for each model when it was imported into UE4. For the same test scenario, the flight path of camera (to capture test video clips) was kept constant. In order to benchmark to the source, for each scenario, the same shot was also captured in Google Earth using its Studio feature [8]. This results in a total number of 18 (6$\times$3) test videos for the reconstructed models and their three corresponding reference (from Google Earth).

B. Experimental methodology

The experiment was conducted in a darkened, living room style environment using a Samsung 28 inch UHD monitor (LU28E590DS), with screen size of 621$\times$341mm. The resolutions of the monitor were configured to 1920$\times$1080 (spatial) and 60Hz (temporal). The viewing distance was set to three times the height of the monitor, which is within the recommended range in ITU-R BT.500 [23].

A double stimulus continuous quality scale (DSCQS) methodology was used. In each trial, participants were shown sequence A and B twice, after which participants had unlimited time to respond to a question asked, *"Please score your viewing experience 1-5 for both videos (5=Excellent, 4=Good, 3=Fair, 2=Poor and 1=Bad)"*. Participants registered their answers by inserting a mark on a continuous scale providing any number between 1 and 5 on a continuous scale. All trials were randomly permutated at the beginning of test session for each viewer, as were the order of the reference and tested videos.

A total of 12 subjects (with an average age of 34) participated in this experiment. All were tested for normal or corrected-to-normal vision. Difference scores were then calculated from each trial and each participant by subtracting the quality score of the sequence for reconstructed model from its corresponding reference. Difference mean opinion score (DMOS) (18 in total) were obtained for each trial by taking the mean of the difference scores.

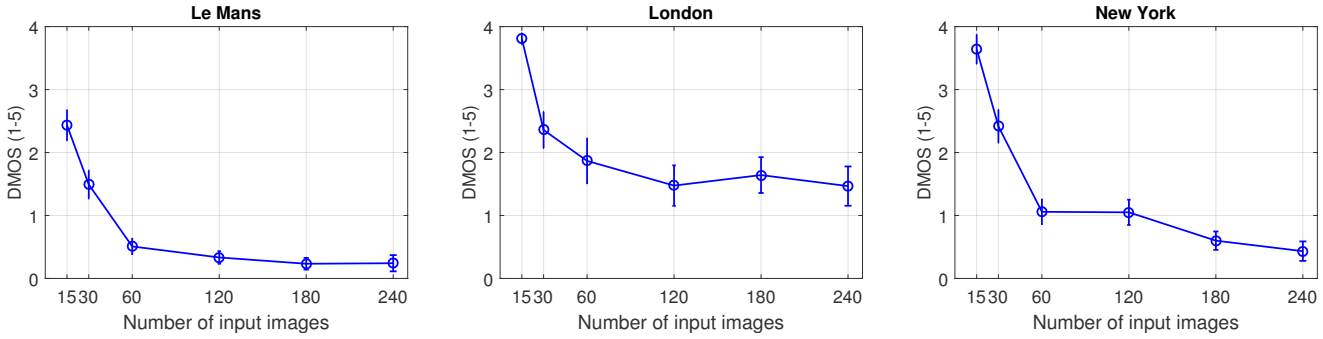
IV. Results and discussions

This section presents the experimental results on the optimal number of input images for actual environment reconstruction, and reports example images of two additional reconstructed 3D models when they were integrated into UE4 for simulation.

A. Subjective results on optimal number of input images

Fig. 4 shows the subjective evaluation results with the number of input images used plotted against the average subjective quality. This indicates that the reconstruction quality for different input image numbers is highly influenced by the environment complexity.

It is observed that the complexity in the London scenario causes a higher average DMOS score (lower quality) for all tested input image numbers, even with high input image count. The fall off in quality towards lower image counts may be caused by inconsistencies and holes in the mesh, a result of camera obstruction. The Le Mans scenario maintains a relatively low average DMOS until very low image counts (lower than 60) due to the lack of any obscuring objects. The drop off in quality is mainly caused by the surface becoming



(a) Le Mans (b) London (c) New York

Fig. 4: Results of experiment testing the optimal number of input images. Here the error bar represents the 95% confidence interval.

bumpy as the depth maps became less accurate. This results in a significant decrease of the perceived quality due to a non flat road surface. The high detail areas in the New York reconstruction like the arm and crown deteriorate quickly with decreasing image counts (lower than 120). This impacts the score heavily despite the rest of the reconstruction being of a relatively acceptable quality, as these areas of the mesh are vital to the image (region of interest).

It can be concluded from the results, in general, that to achieve a relatively high reconstruction quality, for an object based scenario like New York in this work, more than 180 input images are needed per 10000m^2. For other background environments like Le Mans and London, it generally requires at least 120 input images for an area of 10000m^2. It must be noted that, in practice, the "optimal" input image numbers are also highly influenced by the flight path and source data used.

B. Reconstructed environments

Additional to the test scenarios presented above, Fig. 5 and 6 present screen shots for another two 3D environmental models, reconstructed from Google Earth scanned images, when they were imported into the employed simulation engine UE4. The former shows a Roundabout scenario (location 51°31'20" N, 2°35'31" W), which is at north Bristol in the United Kingdom, with the size of the reconstructed area of around 300×130m^2. The latter illustrates a complex, object-based scenario, reconstructed for a iconic building in Bristol, the Wills Memorial Tower [24] (location 51°27'22" N, 2°36'16" W). The size of the reconstructed area is approximately 110×90m^2, and the height of the building is 65.5m.

Based on the reconstructed 3D environmental models, it is possible to plan and simulate different shot types together with various object models, as demonstrated in Fig. 5.(b), where a Cyclist asset (obtained from UE4 marketplace) [25] was integrated into the Roundabout environment for a flyby shot (the definition of different shot types can be found in [22]). Demo videos of the all reconstructed environmental models presented in this paper can be accessed through *https://drive.google.com/drive/u/1/folders/1WJ95kamcG-oWWTQ-ACfoknJJj9MlCdFS*.

(a) Without object. (b) With a cyclist in the scene.

Fig. 5: Screen shots of the reconstructed 3D model for the Roundabout scenario in UE4.

Fig. 6: A screen shot of the reconstructed 3D model for the Wills Memorial Tower in UE4.

V. Conclusion

In this paper, a workflow was presented for UAV-based cinematographic training and planning, based on the simulation of the actual target environments. This generates 3D background models from 2D environmental images, and imports them and their corresponding height maps into a simulation

engine, Unreal Engine 4. To investigate the optimal number of input images for various environment types, a subjective study was conducted on three reconstructed scenarios and six different input image numbers. The proposed workflow will be useful in the application of UAV shot planning, rehearsal and training, especially for the coverage of live events. Future work should focus on the comparison between using open source environmental images and real aerial footage, and the use of realistic control tool, such as Microsoft AirSim [21] for UAV shoot training and planning.

ACKNOWLEDGMENT

The authors acknowledge funding from the European Union's Horizon 2020 (MULTIDRONE No. 731667), and the EPSRC (The Centre for Doctoral Training In Communications at University of Bristol).

REFERENCES

[1] S. Mendes (Director), M. G. Wilson, and B. Broccoli (Producers), "Skyfall," Eon Productions, 2012.

[2] A. Charlton, "Sochi drone shooting olympic tv, not terrorists," Associated Press, February 2014. [Online]. Available: http://wintergames.ap.org/article/sochi-drone-shooting-olympic-tv-not-terrorists

[3] Live Production, "Review sochi 2014: Broadcasting the magic of the games across the world," February 2014. [Online]. Available: http://www.live-production.tv/news/sports/review-sochi-2014-broadcasting-magic-games-across-world.html

[4] K. Gallagher, "How drones powered Rio's olympic coverage," The Simulyze Blog, August 2016. [Online]. Available: http://www.simulyze.com/blog/how-drones-powered-rios-olympic-coverage

[5] DJI, "Djiflightplanner," 2018. [Online]. Available: https://www.djiflightplanner.com/

[6] D. Harmony, "Drone harmony," 2018. [Online]. Available: http://droneharmony.com/

[7] SPH Engineering, "UgCS," 2018. [Online]. Available: www.ugcs.com

[8] Google, "Google earth studio," 2018. [Online]. Available: https://earth.google.com/studio/docs/

[12] 3Dflow, "3DF Zephyr Website," 2018. [Online]. Available: https://www.3dflow.net/

[9] ——, "Google earth," 2018. [Online]. Available: https://www.google.com/earth/

[10] N. Manarin, "ScreenToGif," 2018. [Online]. Available: https://www.screentogif.com

[11] AutoDesk, "ReCap: Reality Capture and 3D Scanning Software," 2018. [Online]. Available: https://www.autodesk.com/products/recap/overview

[13] 3Dflow, *3Dflow 3DF Zephyr User Manual*. Verona, Italy: 3Dflow s.r.l., 2013.

[14] OpenTopography.org, "OpenTopography Website," 2018. [Online]. Available: https://www.opentopography.org/

[15] USGS, "EarthExplorer Website," 2018. [Online]. Available: https://earthexplorer.usgs.gov/

[16] V. Verma, R. Kumar, and S. Hsu, "3D Building Detection and Modeling from Aerial LIDAR Data," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. New York, NY, USA: IEEE Conference Publications, 2006, pp. 2213–2220.

[17] BundySoft, "L3DT: Large 3D Terrain Generator," 2018. [Online]. Available: http://www.bundysoft.com/L3DT/

[18] Unity Technologies, "Unity." [Online]. Available: https://unity3d.com/

[19] YoYo Games , "Gamemaker." [Online]. Available: https://www.yoyogames.com/gamemaker

[20] Epic Games, "Unreal engine." [Online]. Available: https://www. unrealengine. com

[21] Microsoft AI & Research, "Microsoft AirSim Software," 2018. [Online]. Available: https://github.com/Microsoft/AirSim

[22] C. Smith, *The Photographer's Guide to Drones*. Rocky Nook, Inc., 2016.

[23] *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Std. Recommendation ITU-R BT.500-13, 2012.

[24] S. Whittingham, *Wills Memorial Building*. University of Bristol, 2003.

[25] Leedo Studio, "Cycling." [Online]. Available: https://www.unrealengine.com/marketplace/en-US/slug/cycling