MULTI-MODAL BIG-DATA MANAGEMENT FOR FILM PRODUCTION

Hansung Kim[1], Simon Pabst[2], Justin Sneddon[2], Ted Waine[2], Jeff Clifford[2] and Adrian Hilton[1]

[1]CVSSP, University of Surrey, Guildford, Surrey, UK
[2]Double Negative Ltd., London, UK
h.kim@surrey.ac.uk, sicp@dneg.com

ABSTRACT

Modern digital film production uses large quantities of data from videos, digital photographs, LIDAR scans, spherical photography and many other sources to create the final film frames. The processing and management of this massive amount of heterogeneous data consumes enormous resources. We propose an integrated pipeline for 2D/3D data registration for film production. We present the prototype application *Jigsaw*, which allows users to efficiently manage and process various data from digital photographs to 3D point clouds. A key requirement in the use of multi-modal 2D/3D data for content production is the registration into a common coordinate frame. 3D geometric information is reconstructed from 2D data and registered to the reference 3D models using 3D feature matching. We provide a public multi-modal database captured with a wide variety of devices in different environments to assist further research. An order of magnitude gain in efficiency is achieved with the proposed approach.

Index Terms— Big data management, Multi-modal data registration, Film production

1. INTRODUCTION

Digital film production creates final movie frames by combining data captured on the film set with additional elements created during post production. To be able to design these additional elements, it is necessary to capture a large number of assets on the set. For example, replacing something in the frame with a CG rendered object requires at least texture reference photos, information on the principal camera pose and lens, high dynamic range lighting data and LIDAR scans of the film set.

The amount of data captured on set to support digital post production is staggering. In 2014, the data for an average visual effect film Double Negative worked on consisted of several hundred terabytes of data in various file formats. Several years earlier, this number was an order of magnitude lower. A large visual effects facility usually works on several projects at once, making this an even more pressing concern. Table 1 shows an example from *Avengers: Age of Ultron*, which is currently in post production at Double Negative. While data storage is cheaper than ever, all of this data needs to be sorted, indexed and processed, which is a largely manual task keeping many artists busy for weeks during production.

Table 1. Example of data used in *Avengers 2*

Data	Format	Volume
Principal camera	DPX	250K frames/6.5TB
Witness cameras	MXF	17K files/2.5TB
Texture reference	CR2/NEF	580K files/14TB
LIDAR Scans	3D Point cloud	750GB

Smart tools are needed to make this processing more efficient and free up artist resources. In this paper we describe our approach to these issues, consisting of a new data registration pipeline targeted at digital film production and the prototype application *Jigsaw*, a flexible and powerful software platform for tasks revolving around data captured onset. It allows users to import vast amounts of data that can then be surveyed and grouped according to various tasks. It also allows users to start generic processing operations on the data. Historically, each type of data was usually processed on its own, e.g. LIDAR data was merged and converted into mesh representations, HDR photography was prepared for use in image based lighting further down the pipeline, and so on. Registration of assets into a single reference frame only took place later, which could lead to problems if there were errors in the raw data such as gaps in scene coverage.

The work presented in this paper takes a different approach in that it explicitly takes advantage of the multi-modal nature of the captured data. For example, reference photographs were not taken in arbitrary locations but in the same space that is covered by LIDAR data. Being able to register the reference footage in the same coordinate system as the LIDAR data has many benefits, e.g. being able to use the image for projection mapping. It also allows detection of errors in the raw data early in the processing pipeline, when there may still be time for correction.

Thanks to Aaron Carey for further software development. This research was supported by the European Commission, FP7 IMPART project (grant agreement No 316564)

1.1. Contributions

To the best of our knowledge, no fully integrated solution for the processing of large quantities of multi-modal 2D and 3D data has been published before. In this paper we propose a method for automatic registration of large heterogeneous multi-modal datasets into a common coordinate system. It is targeted at the digital film production pipeline and has been integrated in the prototype *Jigsaw* application.

Jigsaw has been used and evaluated in several digital film productions and significantly reduced the time and work required to manage and process onset data. Previously errors in data capture were only detected in post-production often weeks after production when it was too late to correct. Some tasks that were traditionally only performed after principal photography had concluded can now be carried out onset, which allows operators to verify the data they captured is of high enough quality and identify errors such as missing data.

1.2. Related Works

Registration of multi-modal 2D and 3D datasets acquired using different sensors is a challenging task. The datasets exist in different domains with different formats and characteristics such as sampling resolution, accuracy and colour. There has been some prior research into 2D/3D data matching and registration [1, 2], but they assume only a single modality case. In our previous work, we tested the performance of existing 3D feature descriptors for multi-model data registration [3] and proposed a way to combine descriptors in different domains for more robust feature matching [4].

2. WORKFLOW

The *Jigsaw* application offers a flexible user interface for organising massive collections of files originating from a directory structure. It provides a flexible set of tools to organise, register, annotate, visualise and process data. It supports a wide range of file formats, including raw digital stills and common image/video formats, Spherical HDR images, LIDAR point clouds, GPS data, etc., in an essentially infinitely sized hierarchical workspace as illustrated in Fig. 1.

Fig. 2 shows the overall process for multi-modal data registration. We assume that the 3D point cloud obtained from LIDAR scans is the target reference to register other modalities. LIDAR provides the most accurate and complete scene information and also respects the absolute scene scale [5]. Only points visible from the scanner position are recorded, thus several scans from different locations are usually necessary to achieve good scene coverage. 3D data from active range sensors is directly registered to LIDAR through 3D feature detection and matching. 2D footage is registered via 3D reconstruction such as stereo matching or Structure from Motion (SfM) techniques. For example, high dynamic

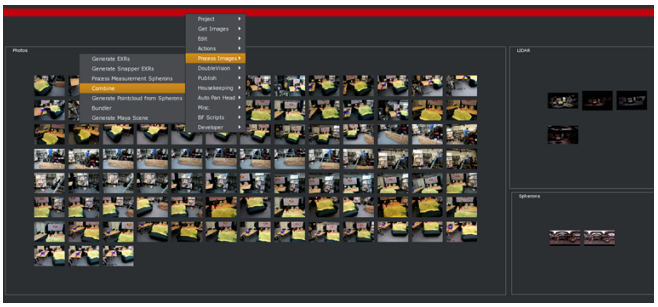


Fig. 1. Multi-modal footage (still photos, LIDAR, stereo Spherons) loaded in a processing group in *Jigsaw*

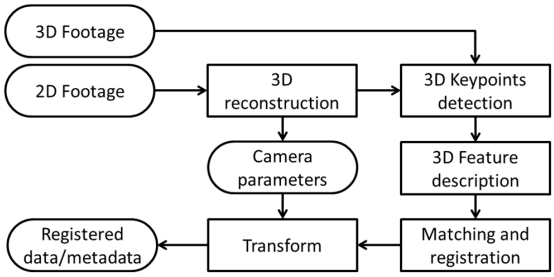


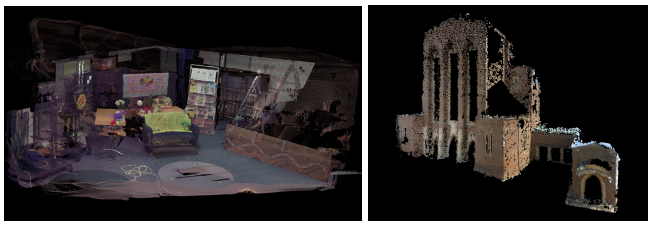
Fig. 2. Workflow for registration

range spherical imaging is often used to capture high resolution environment maps and lighting conditions on set. 3D geometry from spherical images can be recovered by vertical stereo matching as proposed in [6]. Video streams and still photos are widely used to capture additional reference information. 3D geometry and camera poses can be reconstructed by SfM and multi-view stereo methods [7, 8]. Camera poses extracted by 3D reconstruction are in arbitrary coordinate systems and scales, but can be automatically aligned to the reference LIDAR data through the transform matrix resulting from 3D feature matching and registration.

2.1. 3D registration process

The processing pipeline starts with the user importing the raw files into *Jigsaw*. The individual files can then be inspected and annotated by adding shot data from slates or by importing GPS data taken on-set. Files and folders can be grouped in a hierarchical structure independent of their location on the file system as shown as Fig. 1.

The first processing step is 3D reconstruction for 2D data sets as described in Section 2 and shown in Fig. 3 (b). The most popular registration method for 3D data sets is the Iterative Closest Point (ICP) algorithm [9], which requires a rough initial alignment to avoid local minima. Therefore feature matching and initial registration is performed as a prior step to estimate an initial alignment for ICP registration refinement. Details of the registration are given in Section 2.2. If the automatic feature matching and registration fail, cor-



(a) Studio-Spherical (b) Cathedral-Photos

Fig. 3. Visualisation of the reconstructed point clouds

responding points can be manually selected for initial alignment. Using the transform matrices resulting from the registration, all original footage and metadata are transformed into the target reference coordinate frame.

Depending on the number of input datasets and the computation environment (local or distributed to a farm), this registration process for a typical scene takes between several minutes and one hour. The resulting scene contains the merged input 3D datasets in a common (LIDAR) coordinate frame. The format can easily be converted into different representations to facilitate importing into other applications. Transformed camera positions for all input still images and videos are automatically exported into a common format readable by other 3D tools.

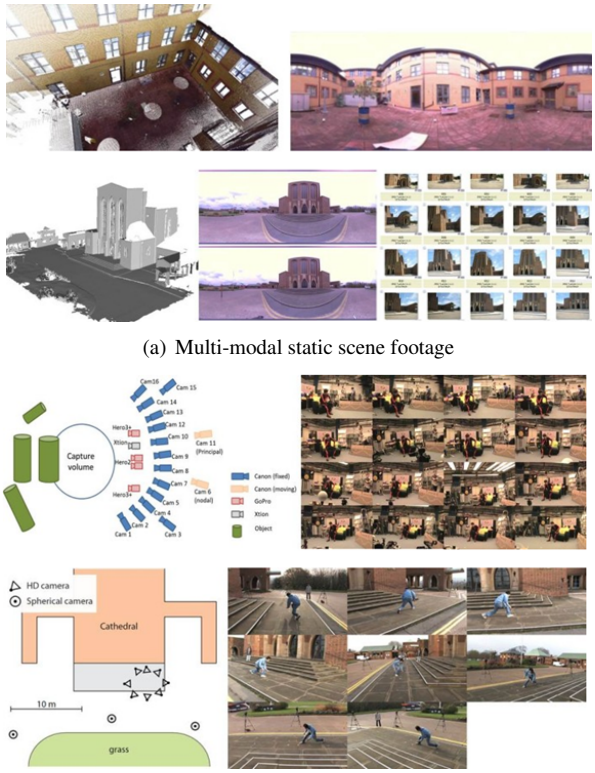
2.2. 3D feature detection and matching

3D keypoints are computed using the 3D extension of Kanade-Tomasi detector [10], which extracts a large number of evenly distributed feature points. 3D descriptors are then calculated for the detected keypoints. We have tested many 3D descriptors in different domains for multi-modal data registration and concluded that the Hybrid Fast Point Feature Histograms (HFPFH)[4] combining local and keypoint domains performs best for general outdoor scenes, and Colour Signature of Histograms of Orientations (CSHOT) [11] works best in a stable ambient lighting environment such as studio capture. HFPFH extracts two separate FPFH descriptors for the same keypoint set in the local point cloud domain and the keypoint domain. CSHOT combines local shape and colour (CIELab) information in one SHOT descriptor.

Once the 3D descriptor sets are computed, all datasets are registered to the target (LIDAR) point cloud using RANSAC-based descriptor matching to find an optimal 3D rigid transform matrix between two point clouds. For HFPFH descriptors, Hybrid SAC-IA [4] is used to refer to both local and keypoint descriptors for optimisation. Finally the ICP algorithm is used to refine the registration by minimising the distance between two point clouds.

3. PUBLIC MULTI-MODAL DATASETS

To support researches in this area, we provide 10 TB (uncompressed) of multi-modal film production dataset captured in various indoor and outdoor locations. It includes raw captured footage and 3D reconstructions for various indoor/outdoor static scenes and multiple synchronised video capture for dynamic actions in the scene. Various capture devices such as LIDAR, spherical cameras, DSLR still cameras, HD (1920×1080) video cameras, HD 2.7 K cameras and RGBD cameras were used as illustrated in Fig. 4. The dataset is available in compressed format at:
`http://cvssp.org/impart/`
Details can be found in the capture notes provided on the website [12].



(a) Multi-modal static scene footage

(b) Multi-view video sequences in the same environments

Fig. 4. Public multi-modal dataset

4. RESULTS

The processing pipeline was tested on three representative datasets from the database: Studio, Patio and Cathedral. All scenes were scanned and captured by a LIDAR scanner, spherical, DSLR and RGBD cameras. The Studio scene was captured in a 3D production studio equipped with an array of KinoFlo fluorescent tubes on the ceiling, with flickerless operation and a consistent colour spectrum. The Patio scene was captured in a courtyard surrounded by buildings. The main area is shaded and the background has repetitive geometry and texture patterns from bricks and windows. The Cathedral scene was captured in an open area with direct sunlight which caused drastic changes in the colour spectrum depending on the capture time.

Table 2. Initial registration result to LIDAR

Data set	HFPFH	CSHOT
Studio-Photos	Success	Success
Studio-Spherical	Success	Success
Studio-RGBD	Success	Success
Patio-Photo set 1	Success	Failure
Patio-Photo set 2	Success	Success
Patio-Spherical	Success	Failure
Patio-RGBD	Failure	Success
Patio-Witness	Failure	Failure
Cath-Photos	Success	Failure
Cath-Spherical	Success	Success

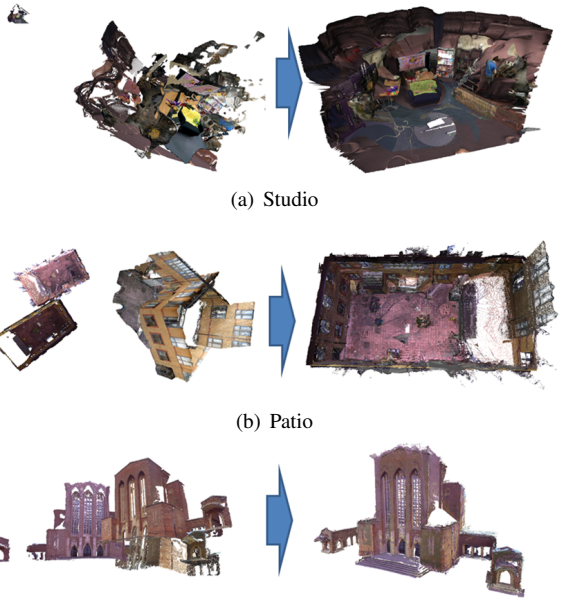
Table 2 shows the performance of the two descriptors for registration of other modalities to LIDAR. The HFPFH descriptor shows stable performance over most cases, but fails for the Patio-RGBD data. The reason for this is that the cylinder-shaped green objects in the scene do not have well-defined corners but only a distinctive colour. The CSHOT descriptor fails for some of the outdoor scenes but can successfully register the Patio-RGBD data. This shows that the two descriptors are complementary in registration. Both descriptors fail for the Patio-Witness video cameras because the reconstruction is too sparse. Therefore, four corresponding points are manually selected for initial alignment. Fig. 5 shows examples of registration results. At the start of the process, the datasets all exist in their own coordinate systems, which are then transformed and aligned with the LIDAR data using the proposed pipeline. The camera poses in the LIDAR coordinate system can be retrieved as illustrated in Fig. 6 since the registration process transforms not only geometry but also camera parameters.

Artists at Double Negative have evaluated the proposed pipeline integrated into the *Jigsaw* software package. The feedback was very positive and the quality of the results achievable in a short amount of time and with minimal user input allowed for a much improved throughput. Processing of data that previously took weeks could now get done within several days representing an order of magnitude efficiency gain. Being able to run some of the processing on set using laptops instills increased confidence in the onset data captured. Errors in the captured data or insufficiently sampled areas can be rapidly detected before it is too late to correct.

5. CONCLUSION AND FUTURE WORKS

In this work, we proposed a pipeline for big multi-modal data registration and described the prototype application *Jigsaw* for efficient 2D/3D data management in film production. Our approach works with data acquired from a wide variety of capture devices used in current digital media production. *Jigsaw* lets users import 2D and 3D datasets into a hierarchical structure, carry out the proposed automatic registration computation, display results and convert them into various output formats for further processing. Trial use of the application in production on 10TB multi-modal datasets achieves an order of magnitude reduction in time required for data processing and allows verification of data quality and identification of errors during production. The multimodal 2D+3D film production dataset used in this work have been released publicly for research at: `http://cvssp.org/impart/`.

Future work is needed to develop a more robust feature matching algorithm to automatically register the principal and witness (sparse) cameras. We are also looking into speeding up the algorithms further to make it possible to process even more data directly on set. Preliminary experiments with GPU acceleration look promising.

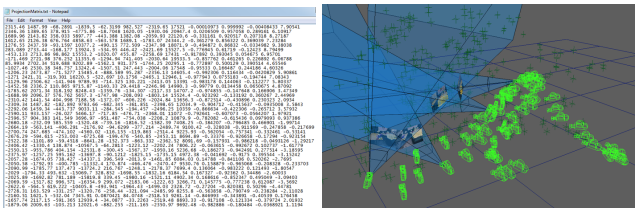


(a) Studio

(b) Patio

(c) Cathedral

Fig. 5. Registration results (Left: Original, Right: Registered)



(a) Camera parameters (b) Visualised camera poses

Fig. 6. Registered digital stills to the LIDAR coordinate

6. REFERENCES

[1] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *Proc. ECCV*, 2012.

[2] M. Restrepo and J. Mundy, "An evaluation of local

shape descriptors in probabilistic volumetric scenes," in *Proc. BMVC*, 2012, pp. 46.1–46.11.

[3] H. Kim and A. Hilton, "Evaluation of 3d feature descriptors for multi-modal data registration," in *Proc. 3DV*, 2013, pp. 119–126.

[4] H. Kim and A. Hilton, "Hybrid 3d feature description and matching for multi-modal data registration," in *Proc. ICIP*, 2014, pp. 3493–3497.

[5] M. Lemmens, "Airborne lidar sensor," *GIM International*, vol. 21, no. 2, 2007.

[6] H. Kim and A. Hilton, "3d scene reconstruction from multiple spherical stereo pairs," *International Journal of Computer Vision*, vol. 104, no. 1, pp. 94–116, 2013.

[7] N. Snavely, S.M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," in *Proc. ACM SIGGRAPH*, 2006, pp. 835–846.

[8] Autodesk, "Recap360," `http://recap360.autodesk.com/`.

[9] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[10] C. Tomasi and T. Kanade, "Detection and tracking of point features," *Pattern Recognition*, vol. 37, pp. 165–168, 2004.

[11] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *Proc. ICIP*, 2011, pp. 809–812.

[12] H. Kim and A. Hilton, "Impart multi-modal/multi-view datasets," `http://cvssp.org/impart/`, DOI: 10.15126/surreydata.00807707.