V4Design

Visual and textual content re-purposing FOR(4) architecture, Design and virtual reality games

H2020-779962

D4.1
Empirical study of visual content

Dissemination level:	Public
Contractual date of delivery:	Month 12, 31 December 2018
Actual date of delivery:	Month 12, 20 December 2018
Workpackage:	WP4: 3D model extraction from 2D visual content
Task:	T4.1 Empirical compilation and study of visual content in the light of 2D to 3D transformation
Type:	Report
Approval Status:	Final draft
Version:	1.0
Number of pages:	56
Filename:	D4.1_V4Design_EmpiricalStudyofVisualContent_20181220_v1.0.pdf

Abstract

The study of visual content, provided by partners in V4Design analyses the videos and images for use in video analysis and 3D reconstruction. It contains an overview of requirements that the visual content should adhere to and a set of guidelines for their selection. The largest part of this deliverable, however, consists of the analysis of provided data in the scope of 3D reconstruction, building and object localization, and aesthetics extraction.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	11/10/2018	ToC creation	Maarten Vergauwen (KUL), Jesper Wachtmeister (SLRS), Konstantinos Avgerinakis (CERTH)
0.2	26/11/2018	Input from content providers was incorporated	Jesper Wachtmeister (SLRS), Jolan Wuyts (EF), Stephan Gensch (DW)
0.3	27/11/2018	Input from CERTH was included	Kostas Avgerinakis (CERTH)
0.4	27/11/2018	Input from KUL was included	Maarten Vergauwen, Jens Derdaele (KUL)
0.5	28/11/2018	Input from SLRS was included	Jesper Wachtmeister (SLRS)
0.6	31/11/2018	Input from SLRS was included	Jesper Wachtmeister (SLRS)
0.7	03/12/2018	Input from CERTH	Kostas Avgerinakis (CERTH)
0.8	07/12/2018	Integration, introduction, conclusions and executive summary was included	Maarten Vergauwen (KUL), Jesper Wachtmeister (SLRS)
0.8.1	08/12/2018	Creation of 1st integrated version based on partner's contributions and circulation to the V4Design consortium	Maarten Vergauwen (KUL)
0.9	11/12/2018	Internal review by V4Design consortium provided comments and adaptations	Eleni Kamateri (CERTH)
1.0	15/12/2018	Creation of 2nd integrated version, rewriting parts according to internal reviewers' comments	Jesper Wachtmeister (SLRS), Maarten Vergauwen (KUL)

Author list

Organization	Name	Contact Information
SLRS	Jesper Wachtmeister	jesper@solarisfilms.se
KUL	Maarten Vergauwen	maarten.vergauwen@kuleuven.be
KUL	Jens Derdaele	jens.derdaele@kuleuven.be
CERTH	Elissavet Batziou	Batziou.el@iti.gr
CERTH	Spyridon Simeonidis	spyridons@iti.gr

CERTH	Konstantinos Avgerinakis	koafgeri@iti.gr
CERTH	Stefanos Vrochidis	stefanos@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr
EF	Jolan Wuyts	jolan.wuyts@europeana.eu
EF	Antoine Isaac	antoine.isaac@europeana.eu
DW	Eva Lopez	eva.lopez@dw.com
DW	Stephan Gensch	Stephan.gensch@dw.com
AF	Kriszta Doczy	kdoczy@artfilms-digital.com

Executive Summary

This deliverable has the goal to report on the activities done on the empirical study of visual data for use in WP4 of the V4Design project. The deliverable focuses on 2D to 3D reconstruction, building and object localization and aesthetics extraction.

We first describe the data compilation process performed by the data provider partners of the project. More specifically we focus on the procedure that was followed by the partners to select and compile relevant data for the tasks in WP4. Furthermore we give an overview of the compiled material, as well as a list of external content and relate this data to their relevant visual tasks.

We continue the deliverable with the results of the empirical study of the visual data for 2D to 3D conversion purposes. An extensive description of important aspects and limitations of 3D reconstruction methods is given, as well as factors that impact the accuracy of the results. Several examples are provided that illustrate these concepts. The content from every partner is analysed and findings are listed, once more illustrated by examples. It is found that the amount of data from content providers that is suitable for 3D reconstruction is limited but that recent evolutions (such as the use of drones and high-definition cameras) yield much more useable content. Furthermore, we describe the analysis of 3rd party content as well as possible alternative approaches for 3D extraction, such as the use of imagery of rotationally symmetrical objects.

The deliverable then describes the analysis of the 2D content for spatio-temporal building and object localization purposes. We list results of the study that was performed on the data, evaluating the usability of the different datasets to this end. It is found that the total number of images and videos from the interior datasets is certainly sufficient but the content is not so satisfying in some cases. For example there are many images which depict maps or text referring to a building, but not the building itself.

Finally we describe the analysis of the collected datasets for use in algorithms that aim at classifying content according to genre or style and can then be used to propose new texture and apply them to existing images. We see that external datasets provide a wealth of famous painting images that are annotated by school of art or creator. The datasets of the project's content providers are also extensive but in many cases lack such annotation.

Abbreviations and Acronyms

AE&TP	Aesthetics extraction and texture proposals
EC	External Content
IC	Internal Content
RSS	Rich Site Summary
SfM	Structure-from-motion
SoR	Surface of revolution
STBOL	Spatio-temporal building and object localization
ToC	Table of Content
WP	Work Package

Table of Contents

1 INTRODUCTION

The objective of the current deliverable is to perform the empirical study of visual content for 3D reconstruction, spatio-temporal building and object localization, and aesthetics extraction. The content that was provided by project partners and described in deliverable D2.1 "*Initial visual and textual dataset creation and legal and ethical requirements*" was analysed by the technical partners and evaluated in terms of its usefulness to cover research objectives.

The deliverable starts by giving a short overview of the data compilation process, which was explained in depth in D2.1. In contrast to D2.1, section 2 focuses on the properties by which the content providers were instructed to search their visual content. Instructions for scene and image selection include: camera motion, baseline, resolution and lighting conditions. A table of the visual data that V4Design partners selected and provided Internal Content (IC) to the consortium for visual analysis is presented here. Since the data compilation process was expanded to also include external content, a table of External Content (EC) collected for testing visual analysis components, is also presented.

Section 3 of the deliverable explains in depth the empirical study of the content for 3D reconstruction. Section 3.1 describes different methods of Multiview reconstruction (the extraction of 3D information from 2D sources) that are presented in depth by first outlining the fundamental physical and optical requirements to successfully achieve 2D to 3D reconstruction. Requirements for photogrammetric reconstruction from multiple images are explained together with factors that impact accuracy. We learn how some objects are hard to reconstruct based on detail, materials and complexity. It is discussed how factors like: Heterogeneity / Distinct features, Image resolution, Lighting homogeneity, Spatial resolution, Non-rigid scenes (Scenes with moving objects) and Complex objects all affect the reconstruction accuracy, and the possibility to achieve a successful 2D to 3D reconstruction.

In section 3.2 it is discussed how the material provided by the different partners (EF, AF, DW, SLRS) meet - and don't meet - the requirements for a successful 2D to 3D reconstruction. It becomes clear that the amount of data, provided by the project's content providers, that is suitable for 3D reconstruction using Multiview methods is limited.

Alternative approaches to 3D and information extraction are therefore discussed in 3.3. In section 3.3.1 YouTube and Online Image repositories are presented, tried and evaluated as visual data sources - and as potential providers of quality material with a higher success-rate in the 2D to 3D transformation process. In 3.3.2 symmetrical items and other well-captured objects are evaluated as optional ways to create 3D objects from single images - either using the assumed "known" symmetries in an object, or by means of recognition of features in an object based on metadata.

Next to 3D extraction from 2D data, two other tasks in V4Design comprise i) spatio-temporal building and object localization and ii) aesthetics extraction and texture proposals. The former task aims at finding buildings and specific objects in existing image and video datasets. An analysis was performed on the collected data, evaluating the usability of the different datasets to this end. Section 4 describes the results of this study.

Finally, section 5 concludes the document focusing on the analysis of the collected datasets for use in algorithms that aim at classifying content according to genre or style and at extracting specific aesthetic information from it. This information can then be used to propose new texture and apply them to existing images.

2 OVERVIEW OF DATA COMPILATION

Section 2 starts by describing the datasets that were delivered by the V4Design content providers (see also D2.1) and what visual properties the content providers were instructed to look for when searching their visual content. This is explained in 2.1. In section 2.2 the data from external sources are briefly described.

2.1 Internal content

The methods for collecting and delivering data and metadata differ between content provider partners. Some content providers, like DW and EF, have APIs that allow for easy extraction of datasets that are customisable in granularity and complexity. Other data stores, like those of AF and SLRS, require manual curation work to extract content that is relevant to V4Design. The general steps in the content collection and provision process are outlined below.

1. Definition of content-related user requirements by user partners, and content-related technical requirements by technical partners (cfr. Annex A of D2.1)
2. Creation of a keyword list (cfr. Annex C of D2.1) per PUC by technical partners
3. Extraction of content.

2.1.1 Data Compilation

A great deal of data collected for the initial dataset (delivered with D2.1. in M10) consisted of data for the purpose of 3D reconstruction and video analysis. The databases of content provider partners were queried to return relevant items that could be used for 3D reconstruction and video analysis tasks. These data from the existing databases of content provider partners are listed in *D2.1 Appendix B: List of components in the initial dataset.* Below is a list of the most important components of the initial dataset that were compiled from the existing databases of the content provider partners.

- Freely reusable images, texts and videos that were relevant to the PUC contents and were generally relevant to the architect and creative designer audience were selected and exported from the **EF** database
- **AF** exported a selection of their movies, along with metadata and timestamps of key frames.
- **SLRS** provided a selection of their movies, along with metadata and timestamps of key frames.
- **DW** provided a series of telenovela episodes, transcripts of those episodes, and interactive language exercises for PUC 3. To improve the scenario case of the Bauhaus experience for PUC4, DW provided several videos covering Bauhaus architecture and interiors.

In the following Table (Table 1) we can see the data that V4Design content providers provided to the consortium for visual and textual analysis. Each content is associated with the relevant visual task, namely Aesthetics Extraction and Texture Proposals (**AE&TP**), Spatio-Temporal Building and Object Localization (**STBOL**), 3D-reconstruction (**3D-recon**) and analysed in the following sections of this deliverable. The empirical study of the textual data was taken into account in D3.1 and will not be elaborated on the following sections.

Table 1: Internal Content (IC) from V4Design content provider

#Content ID	Dataset name header	PUC#	Partner	Description	Relevant visual task#
IC001	Hans Scharoun files	1	EF	63 freely reusable objects, with metadata and image files focusing on Hans Scharoun	STBOL
IC002	M+ Kowloon files	2	EF	838 freely reusable objects, with metadata and image files focusing on: Hong Kong, Kowloon, Victoria Harbour	STBOL
IC005	Japanese Art in photos	2	AF	1200 photos focusing on Japanese art	STBOL
IC006	Japanese videos	2	AF	Films focusing on Japanese art	STBOL
IC007	Berlin Wall files freely reusable	4	EF	65 freely reusable objects, with metadata and image files focusing on: Berlin Wall, Berliner Mauer, Antifaschisticher Schutzwall, Wall of Shame, Checkpoint Charlie, Friedrichstrasse, Friedrichstraße, Glienicke Bridge, Glienicke Brücke, Oberbaumbrücke, Dreilinden, Drewitz	STBOL
IC008	General architecture design video dataset	1,2, 4	AF	Films on Chinese design and architecture, interviews and documentaries	STBOL
IC009	SLRS database	1,2, 4	SLRS	4 films from the Jesper Wachtmeister filmography (architecture documentaries) named: Microtopia, Great Expectations, Kochuu, Bruno is Back.	STBOL, 3D-recon
IC010	Nicos' Weg	3	DW	230 episodes of Nicos' Weg DW series including trailer	STBOL, 3D-recon
IC011	Chateau de Versailles	4	EF	402 freely reusable results, including media attachments focusing on Versailles and its premises.	STBOL
IC012	Bruno Mathsson	1	EF	11 images and 1 sound clip, not necessarily open for reuse or with attached media focusing on 'Bruno Mathsson'.	STBOL
IC013	Eglise de Sorbonne	4	EF	Small curated set of objects from DW clips via Euscreen, sketches, pictures, architectural drawings	STBOL
IC014	Notre Dame church in Dijon	4	EF	46 results, images and text, not necessarily freely reusable or with attached media focusing on Notre Dame Church in Dijon	STBOL

IC015	Hans Scharoun (architect)	1	EF	99 results, 97 images and 2 video clips, of Hans Scharoun, only in limited re-use or no re-use	STBOL
IC016	Jesper Wachtmeister films	1,2	SLRS	timestamps of possible interesting data, including timestamps, across 3 documentaries: 'Kochuu', 'Great Expectations', and 'Microtopia'	STBOL, 3D-recon
IC017	Artfilms	1	AF	Over 600 short clips, 2-5 minutes focusing on screeners for contemporary art	STBOL
IC019	PUC 1 Scenario 1 EF images	1	EF	920 CHOs with subject of ancient greek architecture with a focus on Delphi.	STBOL
IC020	PUC 1 Scenario 2 EF images	1	EF	5783 CHOs with subject of German pre-1950s architecture, focus on Berlin architecture.	STBOL
IC021	PUC 1 and PUC 4 AF video content	1, 4	AF	23 video objects with 9 architecture related videos, 9 Mediterranean focused videos, 5 videos on other topics	STBOL, 3D-recon
IC022	PUC 2 EF images	2	EF	2693 CHOs focusing on Japanese and Chinese material design.	STBOL
IC023	PUC 2 AF images	2	AF	600 CHOs with photos from Japan, China, and other images from the films AF distribute.	STBOL
IC024	PUC 2 AF video content	2	AF	8 video objects with 1 video object on Balinese art and design, 4 videos on Chinese architecture and art, 3 videos on Japanese architecture and art.	STBOL
IC025	PUC 3 DW content	3	DW	230 video objects of Nicos Weg, Web-based exercises for Nicos Weg, 2 screenplays (A1, B1), and four transcipts	STBOL
IC026	PUC 4 EF images	4	EF	7 CHOs focusing on the Gendarmenmarkt square.	STBOL, 3D-recon
IC027	PUC 3 Nico's Weg exercises	3	CERTH	440 webpages with textual information from Deutsche Welle webpages containing Nico's Weg exercises.	Text (irrelevant)
IC028	SLRS videos and timecodes	1, 2, 4	SLRS	3 video objects and associated timecodes	STBOL, 3D-recon
IC029	EF aesthetic extraction paintings	/	EF	~14.000 images and metadata focusing on collection of pre-1950 painting artworks in Europeana, for aesthetic extraction tasks.	AE&TP
IC030	EF CHO metadata	/	EF	~23.000 metadata text items the associated metadata for all provided CHOs, in JSON format, for	Text

	for textual analysis			textual analysis tasks.	
IC031	symmetrical vase dataset	2	EF	2262 CHOs with Subject of symmetrical small-scale objects like vases, amphoras, bowls.	3D-recon
IC032	Sum of all Paintings dataset	1-4	EF	~37.000 CHOs with Wikidata-extracted dataset from their SPARQL endpoint containing paintings with at least 1 associated genre or style tag.	AE&TP
IC033	Newspapers dataset	1-4	EF	approx. 6.694.985 digitised newspaper pages and metadata focusing on 592 different newspaper titles from across Europe, collected by European Library	Text (irrelevant)
IC034	drone footage	1, 4	AF	8 clips around ~10 minute duration, capturing from a drone: A ship, a building, cityscape and industrial port objects were filmed from every angle.	3D-recon

Additionally, specific needs for tasks in V4Design necessitated the creation of specific datasets for which the data was available in the databases of the content providers but the datasets themselves weren't collected and curated yet. Below is a list of the specific datasets needed for tasks that were created for V4Design only.

- For 3D reconstruction tasks from several images a dataset was needed consisting of multiple images of objects that were fairly symmetrical. EF constructed a dataset of a few thousand vases, jars, pots and amphoras for KUL to use in this task.
- For 3D reconstruction tasks from video footage some drone footage of selected buildings was needed. AF created a small set of drone footage for this purpose
- For the creation of a VR - reliving the date experience (PUC4), appropriate footage of the Gendarmenmarkt was needed. DW created video footage for this task. To help create further objects for the use case, DW has provided an extensive list of buildings and objects to be reconstructed in 3D.
- Additional shots and 3D scans of the apartment of Nico's Weg have been commissioned to be created in early December 2018 to provide interior details for PUC3.
- To provide input for testing and evaluating restyling of 3D objects, DW provided a 3D scan of the rooftop workshop of the August Macke Museum in Bonn alongside with texture maps. This model also provides input for the BIM extraction tasks.

2.2 External content

For aesthetic extraction tasks a dataset of homogeneously tagged paintings was needed with the creator and style attached to them. A dataset of this quality was created by extracting about 37 thousand paintings from Wikimedia commons using the wikidata query service and SPARQL. More information on this dataset can be found in Deliverable 2.1. Section 5: *Data collection from external providers.*

Table 2: External Content (EC) collected for testing visual analysis components

#Content ID	Dataset name header	PUC #	Partner	Description	Relevant visual task#
EC001	Wiki webpage scraping	1-4	CERTH	314 webpages with textual content and metadata collected from Wikipedia webpages (311 castles plus three pages addressing PUC4).	Text (irrelevant)
EC002	Wiki images scraping	1, 4	CERTH	663 images crawled from the aforementioned Wikipedia webpages.	STBOL
EC003	Twitter posts scraping	1-4	CERTH	40073 twitter posts from 21 user accounts.	Text
EC004	Flickr dataset	1, 4	CERTH	6209 images returned using 13 queries on the Flickr API: Eiffel tower, Volkswagen beetle, tour Eiffel, white tower, Λευκος Πύργος, Delphi oracle, Delphi temple, ancient Delphi, Gendarmenmarkt square, Franzosischer Dom, Deutscher Dom, Konzerthaus, Schiller Monument	3D-recon
EC005	youTube dataset	1, 4	CERTH	42 videos of famous buildings around the globe and 15 videos of famous statues	3D-recon
EC006	COCO dataset	1-4	CERTH	The Microsoft Common Object in Context (**COCO**) dataset contains more than 200,000 images that depict 91 categories that occur in both indoor and outdoor objects. The dataset also contains both bounding boxes and object segmentation output.	STBOL
EC007	Open Images	1-4	CERTH	**Open Images** is a dataset of ~ 9M images that have been annotated with image-level labels and object bounding boxes. The training set of V4 contains 14.6M bounding boxes for 600 object classes on 1.74M images, making it the largest existing dataset with object location annotations. The boxes have been largely manually drawn by professional annotators to ensure accuracy and consistency. The images are very diverse and often contain complex scenes with several objects (8.4 per image on average). Moreover, the dataset is annotated with image-level labels spanning thousands of classes	STBOL
EC008	ImageNet	1-4	CERTH	The **ImageNet** is a large visual database designed for use in visual object recognition software research. Over 14M images have been hand-annotated by ImageNet to indicate what objects are depicted; in at least one million of the images, bounding boxes are also provided. ImageNet contains over 20 thousand categories. Currently, bounding boxes for over 3000 popular synsets are	STBOL

				available. For each synset, there are on average 150 images with bounding boxes	
EC009	Places2	1-4	CERTH	**Places2** dataset contains more than 10M images comprising more than 400 unique scene categories.	STBOL
EC010	SUN397	1-4	CERTH	**SUN397** database contains 397 categories. The number of images varies across categories, but there are at least 100 images per category, and 108,754 images in total.	STBOL
EC011	Oxford Buildings	1, 4	CERTH	**Oxford Buildings** dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated.	STBOL
EC012	Pandora	1-4	CERTH	The **Pandora** paintings dataset has a collection of 18,720 paintings from many different resources. The image collection has been distributed among 18 style classes, having approximately 1,000 images. Engineers have ensured that only the relevant part of the images is shown and art experts also ensure that the artistic annotation is valid.	AE&TP
EC013	Wikiart	1-4	CERTH	The **Wikiart** paintings dataset is an image collection of 81,472 paintings images, from more than 1,000 artists. This dataset contains 27 different styles and 45 different genres. 81,446 paintings are used for style classification, while only 10 genres with more than 1,500 paintings are chosen for genre classification, with a total of around 64,995 samples. Similarly, only a subset of 23 artists with more than 500 paintings is chosen, with total amount of 19,051 images for artist classification.	AE&TP
EC014	Paintings -91	1-4	CERTH	**Paintings-91** dataset consists of paintings from 91 different painters. Three applications are foreseen where ground-truth annotation is available, namely artist categorization, style classification and saliency detection.	AE&TP

3 EMPIRICAL STUDY OF THE CONTENT FOR 3D RECONSTRUCTION

The following chapter describes the limitations of photogrammetric 3D reconstruction with regards to the input data. An initial analysis for Multiview 3D reconstruction was performed on the currently available V4Design dataset.

3.1 Multiview reconstruction

In this section we will briefly describe important aspects of the Multiview 3D reconstruction algorithms used in V4Design and their impact on the suitability of image and video content provided by the partners.

3.1.1 Photogrammetric reconstruction from multiple images and its limitations

Photogrammetric Multiview 3D reconstruction, also known as **structure-from-motion (SfM),** is a technique for estimating three-dimensional structures from two-dimensional imagery. The perspective projection that lies at the basis of the formation of 2D imagery gives rise to the loss of one dimension: the depth of the scene, i.e. the distance to the cameras. This depth information can be calculated from two-dimensional data sources using common information that is present in overlapping parts between different images or video stills. In order to do so, the reconstruction method computes both the calibration of the cameras and correspondences between images.

Not all 2D image or video data is equally suitable for 3D reconstruction purposes. In the following subsections we will list some important constraints that limit the applicability of the method.

Camera motion

Similarly to how humans perceive depth in a scene by walking around or viewing it from multiple angles one can mathematically calculate this depth from image sequences. This brings us to the main requirement for image-based 3D reconstruction: the existence of camera motion which gives rise to parallax in different images. For this reason it is impossible to calculate depth information from images, captured without changing camera positions in between recordings. Usage of a tripod camera is very common in media productions such as movies, news broadcasts or documentaries: steady shots are guaranteed and it is very cost-effective compared to a steadily moving camera. This limits the camera to a 'panning', 'tilting' and zooming function (Figure 1) making 3D reconstruction impossible.

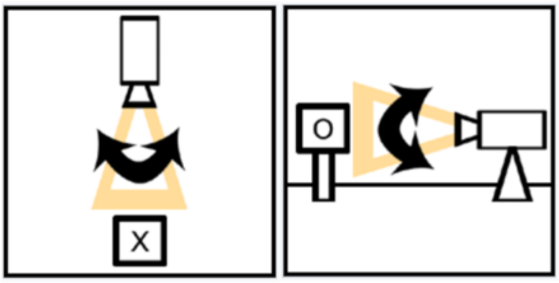


Figure 1. Left: panning camera motion. Right: tilting camera motion. Video shots containing panning and/or tilting motions only contain insufficient information for depth reconstruction.

Baseline

The baseline is defined by the relative distance between consecutive images in a scene. Shots taken from a camera limited to panning and tilting have no baseline. Depth information received from high baseline images will be more accurate than those received from a low baseline (Figure 2). Similarly when one considers a camera in a forward-only motion, despite moving a long distance, the angle of intersection will remain low (Figure 3).

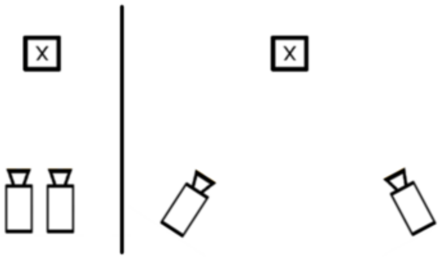


Figure 2. Left: low intersection angle, depth reconstruction will be less accurate. Right: large intersection angle, depth reconstruction will be more accurate.

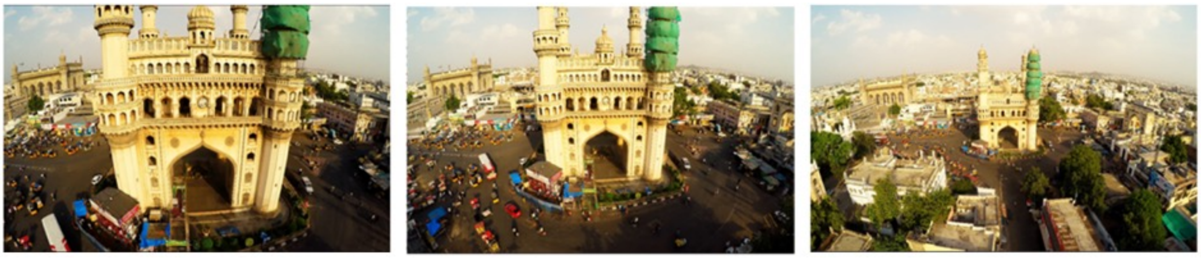


Figure 3: Drone footage using a forward-only motion. This results in a low intersection angle in-between all video frames. Subsequently, a photogrammetric reconstruction from this castle will either fail or be of bad quality.

On the other hand, a very high baseline results in very different projections of the scene into the images. This makes the task of matching different features in-between images more complex and in many cases impossible. A good example of a wide-baseline recording is shown in Figure 4.

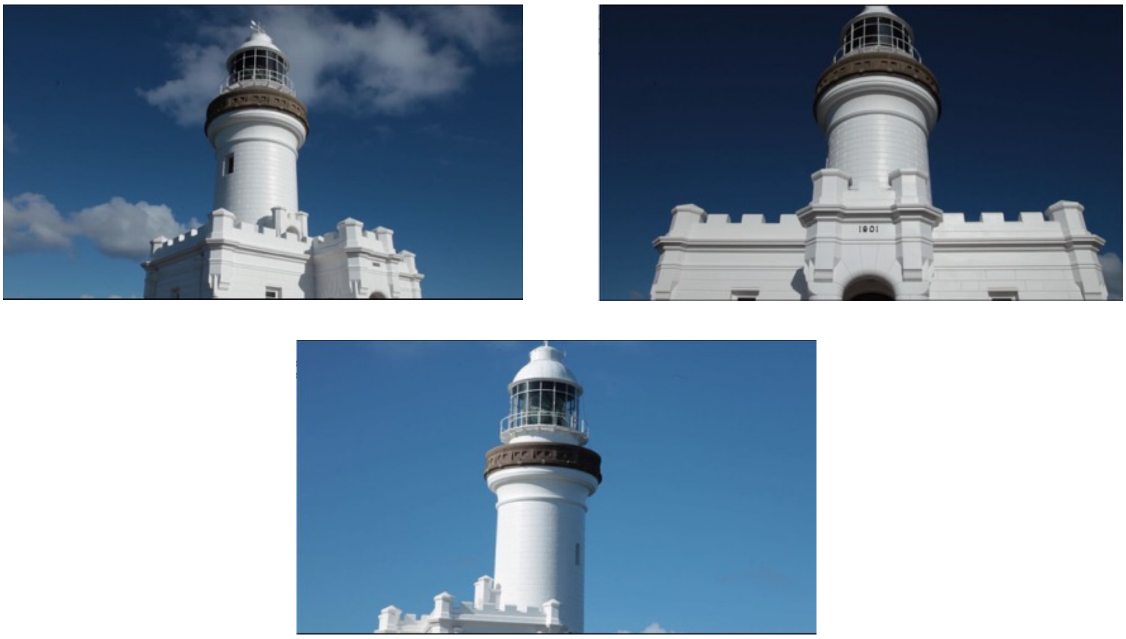


Figure 4: Images provided by AF containing a very high baseline

Original footage

Input footage should not be edited during post-production in a way that interferes with the mathematical model of the used camera. In general any editing that displaces the 2D pixel coordinates or asymmetrically resizes the resolution may render reconstruction impossible. Any other edits such as color correction do not impact reconstruction capability. Some common post-productions techniques include:

(Asymmetrical) cropping:

> Cropping means cutting away sections from the video as seen in Figure 5. Therefore original pixel coordinates change. This may cause a large error during the camera calibration process of a SfM pipeline, eventually causing the reconstruction to fail or produce undesirable results.

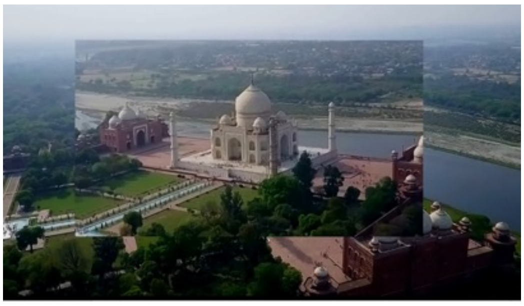


Figure 5: Cropped and/or resized video

Video compilation:

> In some videos we encountered multiple video sequences that are shown side-by-side as depicted in Figure 6. Once again pixel coordinates change, rendering mathematical interpretation of these frames useless.

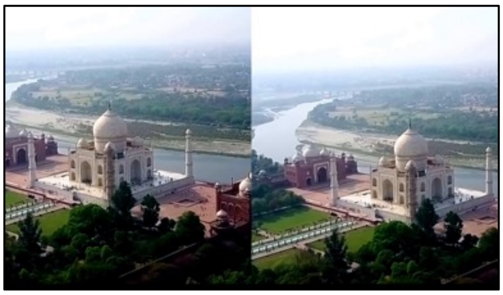


Figure 6: Edited drone footage from Taj Mahal. The video displays 2 drone shots next to each other. An automated algorithm cannot detect this.

3.1.2 Accuracy impacting factors

While some 2D content may meet the minimum requirements for Multiview 3D reconstruction, additional prerequisites arise if one aims to obtain high quality and geometrically correct results. For an overall high quality model the following components will be judged:

- Reconstruction accuracy
- Model integrality
- Texture quality

In order to obtain a high quality model, the input data should score well on each of these components.

Reconstruction accuracy

The reconstruction accuracy defines how well digitally reconstructed areas match the real world objects that were depicted in the imagery: the error between a potential ground-truth model and the reconstructed parts. This is mainly determined by following input data factors:

Heterogeneity / Distinct features

Multiview images are stitched/aligned together using points of similarity between a set of input images. Therefor if a scene or object has a large number of similar features or large homogenous areas, such alignment may contain small to large errors or even fail. These errors in turn may have a large impact on the depth calculation.

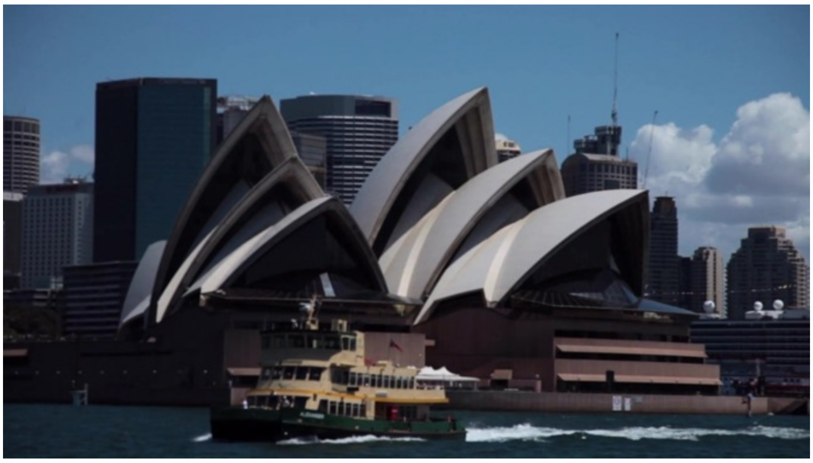


Figure 7: Video of opera house in Sidney (AF) containing many homogenous surfaces: The roof for instance is plain white and the walls are brown.

Image resolution

Image resolution also has a large impact on the reconstruction quality. Many of the data provided by the content partners has some historical value or was captured some time ago. Full HD resolution (1920x1080 pixels) was adopted in the mid to late 2000s. Before this, common formats were the European PAL format (720x576) and NTSC (854x486) mainly used outside Europe. Additionally many of these NTSC and PAL format cameras were recorded or stored on video tape. These resolution formats were largely unchanged with the introduction of widescreen televisions (16:9 aspect ratio), instead the aspect ratio of individual pixels were changed to a rectangular shape from a square one. Based on 3D reconstructions tested by KUL a full HD resolution provides much higher quality 3d reconstruction results. Currently, in 2018, most productions and cameras support a 4K resolution (3840 x 2160 pixels) further improving reconstruction quality.

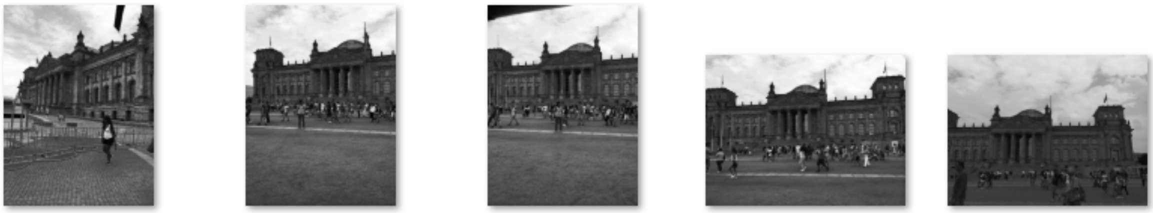


Figure 8: Low resolution images.

Model integrality/completeness

The resulting 3D model should be as complete as possible. This means that no holes or missing parts are present. For this a video or image sequence must show the objects and scenes from all sides: a 360° view, since areas that are not visible in an image sequence cannot be reconstructed.

Texture quality

The final phase of a reconstruction process is the texture mapping. After dense 3D information has been successfully extracted from video and image input material, the resulting meshes need to be colorized. To this end patches from the original images will be copied onto the resulting mesh. In order to get a high quality mesh the following factors are important:

Lighting homogeneity:

Since texture patches from several images will be combined on a mesh, large lighting differences (e.g. combination of morning/midday/evening/night) in the scene may cause a poor texture blend for the resulting mesh or even ‘night’ and ‘day’ areas in a single reconstruction. For this reason all images ideally should have been shot under the same lighting conditions. Figure 9 shows two images of the Gendarmenmarkt from content provider DW. The same building is shown but the lighting conditions are clearly different, which results in different colouring of the scene. The lighting conditions in Figure 10 are so bad that the image is close to unusable for 3D reconstruction. In general, and this is the case for all photogrammetric techniques, the lighting conditions and resulting colour changes will be ‘baked’ into the resulting mesh.

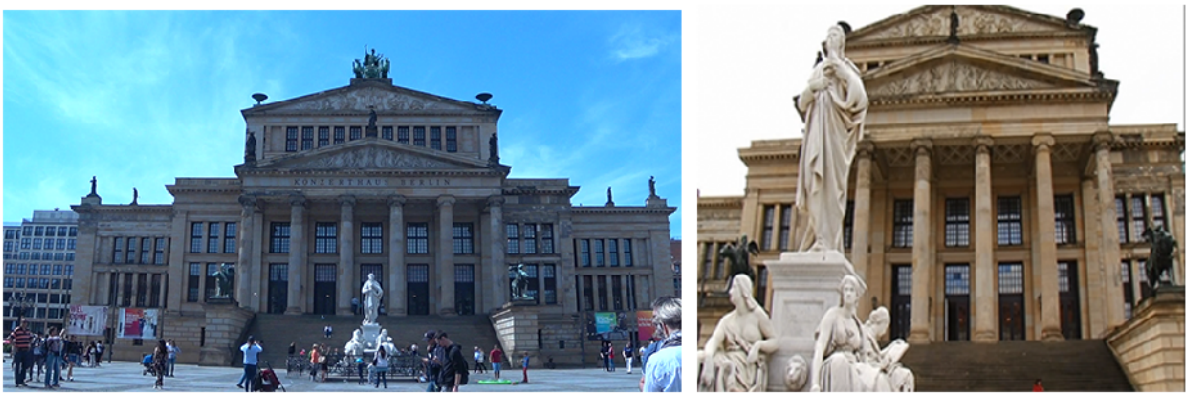


Figure 9: Two images from the Gendarmenmarkt dataset, taken with different lighting conditions

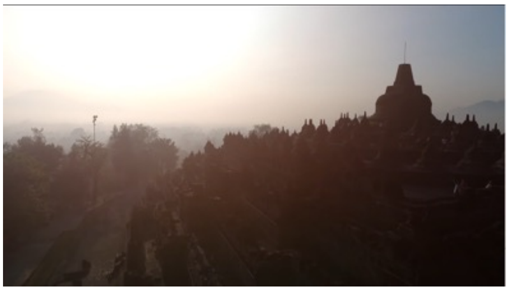


Figure 10: Bad lighting conditions.

Spacial resolution:

The special resolution defines the size of a pixel in relation to the area that pixel maps on the specific model. Individual pixels in an image taken from a high distance will map a large area on the resulting model.

In geographic information systems (GISs), spatial resolution is measured by the ground sample distance (GSD) of an image, the pixel space on the Earth's surface.

3.1.3 Hard to reconstruct objects

Some objects may not be suitable for reconstruction, even from imagery that complies with stated input limitations listed in the previous sections. These objects can be divided into the following categories:

Reflective structures

If the surface of an object is very reflective, it may look very different from varying viewing angles, especially in bright lighting conditions. As a result, computer vision algorithms are unable to determine corresponding feature points between different images. This is one of the main reasons why the reconstruction of cars is not feasible or will result in poor or unusable meshes.

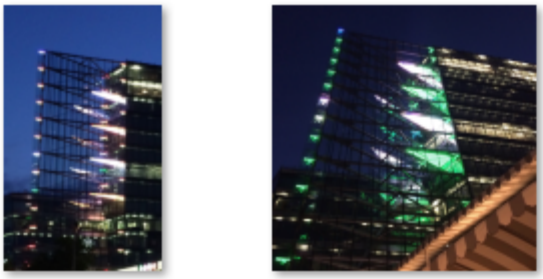


Figure 11: High reflective footage (AF)

Non-rigid scenes

Since photogrammetric 3D reconstruction recovers depth from multiple images, it is important the observed scene is rigid. Objects that change shape or position during successive footage capture cannot be reconstructed. Examples of these are humans, animals, plants and trees. This also includes smaller objects that may have changed position in a scene. Examples of such are opening doors or cars. The restriction to static scenes is important because it eliminates a large percentage of existing movies, which typically focus on people (see example in Figure 12)

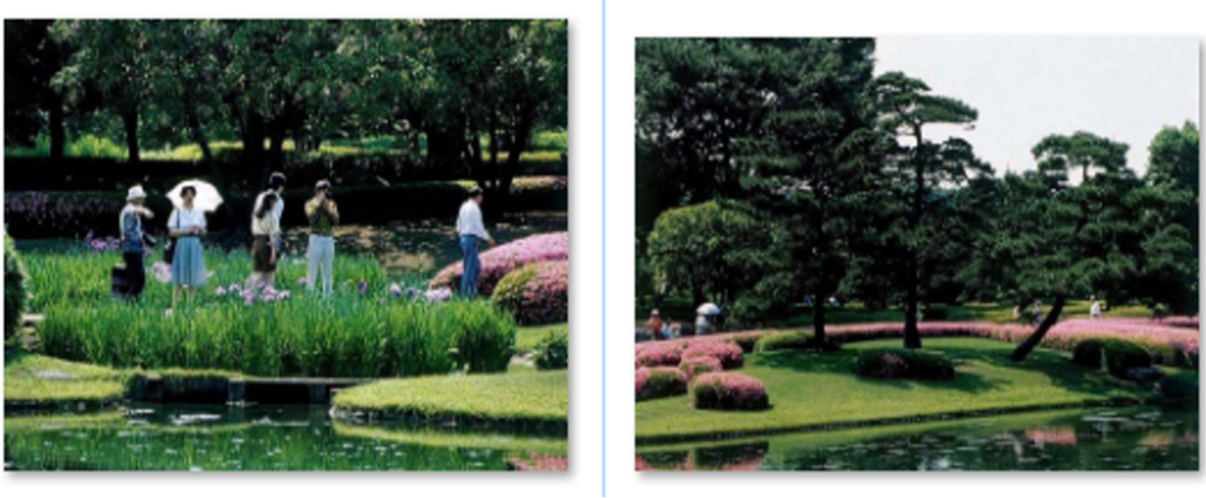


Figure 12: Data provided by AF showing moving people (left) but also plants and trees (right)

Complex objects

The dense matching stage of the Multiview 3D reconstruction pipeline searches for correspondences for all pixels of an image in other images. This is only possible by employing some regularisation term in the object function that needs to be minimized. This term typically imposes a certain amount of smoothness on the reconstructed depth. Such algorithms, however, yield poor results for scenes with many occlusions or depth changes. A good example, shown in Figure 13, is a bike that contains narrow parts, such as the frame and the spokes in the wheels. A second example in Figure 14 shows a piece of art made up from a complex metal structure

Figure 13: Bike (EF)

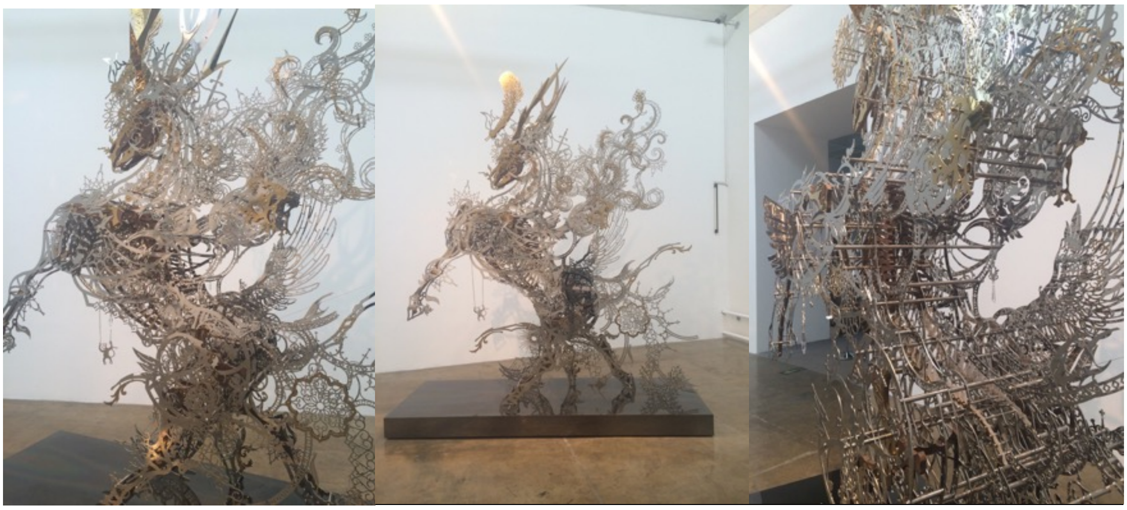


Figure 14: Complex artwork (AF). Images captured from different viewpoints but too complex for reconstruction.

3.2 Study of content partner dataset for Multiview reconstruction

Taking into account the aspects and constraints of the previous sections, we conducted an empirical study of the content that was made available by the content partners. We will discuss the data from EF, AF, DW and SLRS. in the following subsections.

3.2.1 Europeana Foundation

EF provides both image and video material, generally of historical significance. Works of art, artefacts, furniture of specific era's in time etc. Part of this data was originally provided by museums and as such it is well documented (year, type, material, measurements). However, most of this data does not meet the necessary requirements for use in a 3D reconstruction pipeline.

Digitized museums inventory

Historical objects from museums around the world are digitized using high quality photos. For reconstruction purposes either a minimum of two viewing angles or prior 3D information is required. For the vast majority of the elements in this EF dataset, multiple images are not available and therefore actual 3D information is not feasible by means of traditional photogrammetric methods. Cases where multiple images are available, are cropped and/or lack any baseline (Figure 15). Section 3.3.2 discusses a research topic within the V4Design project that aims to make use of a portion of this data for single view 3D reconstruction. Well-structured metadata make it possible to group some images for one particular object. This is very helpful for batching images together and initiating a reconstruction process on them.

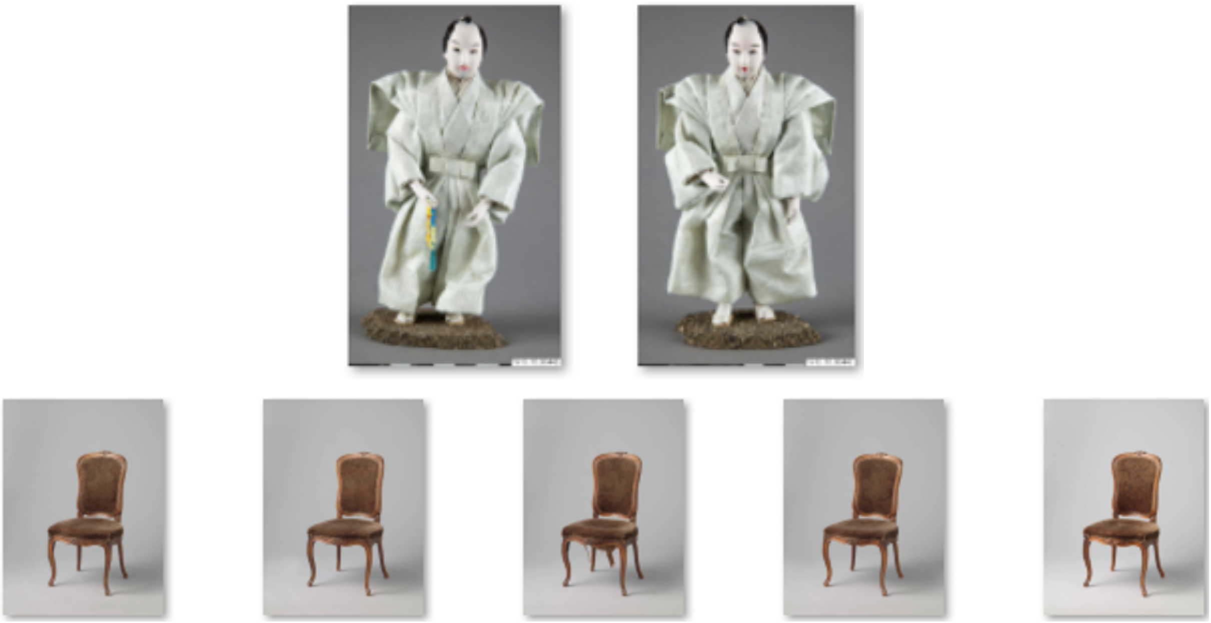


Figure 15: Some image datasets from EF containing the same object

Historical video footage

A limited amount of data from EF also features historical video footage. The majority of this originates from news agencies (tv news broadcasts). For this data no content was deemed suitable for 3D reconstruction due to poor video quality and limited/no camera movement (Figure 16).

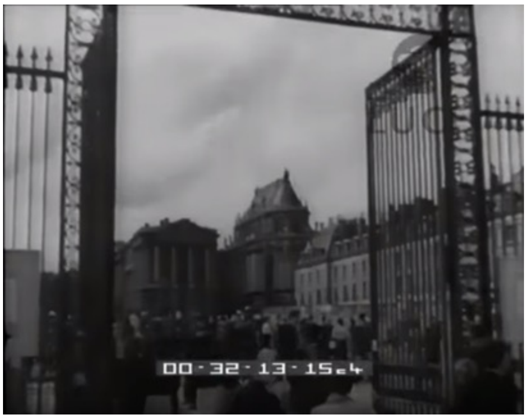


Figure 16: Historical news footage (EF)

Digitized (old) photography

This type of data consists of digitally scanned photos. These images were initially captured on photographic film and subsequently digitally scanned. Both greyscale and colour images are available. Since the images were scanned later as seen in Figure 18, no prior knowledge of the camera calibration, such as the focal length is available, making the reconstruction process more difficult. In some cases no scan was taken but instead a photo of the developed film (photo of a photo), which makes 3D information extraction impossible with current techniques.

The general low quality footage and limited Multiview content rules out reconstruction using only this data. However, thanks to the available metadata, provided by EF, this imagery could be combined with other sources. This was done specifically for the Gendarmenmarkt square (Figure 17) in pilot use case 4 (further described in section 3.2.3).

This data, and in extent a majority of all data by EF, is well annotated for 3D reconstruction purposes. Figure 19 shows the filename provided for the images, usually consisting of an ID and ending in an incremental numbering format. This incrementing number indicates the photo number from one specific scene or batch. This tremendously helps the matching strategy algorithm for 3D reconstruction.

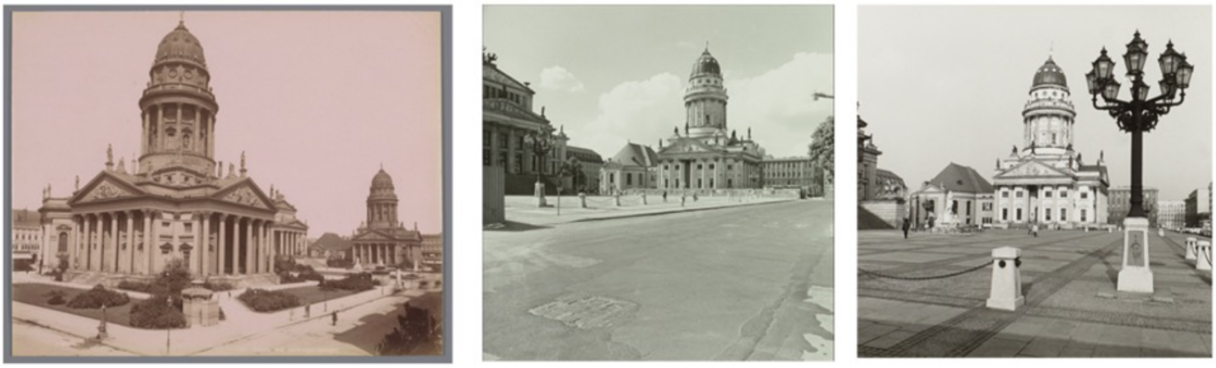


Figure 17: Historical images from the Gendarmenmarkt provided by EF

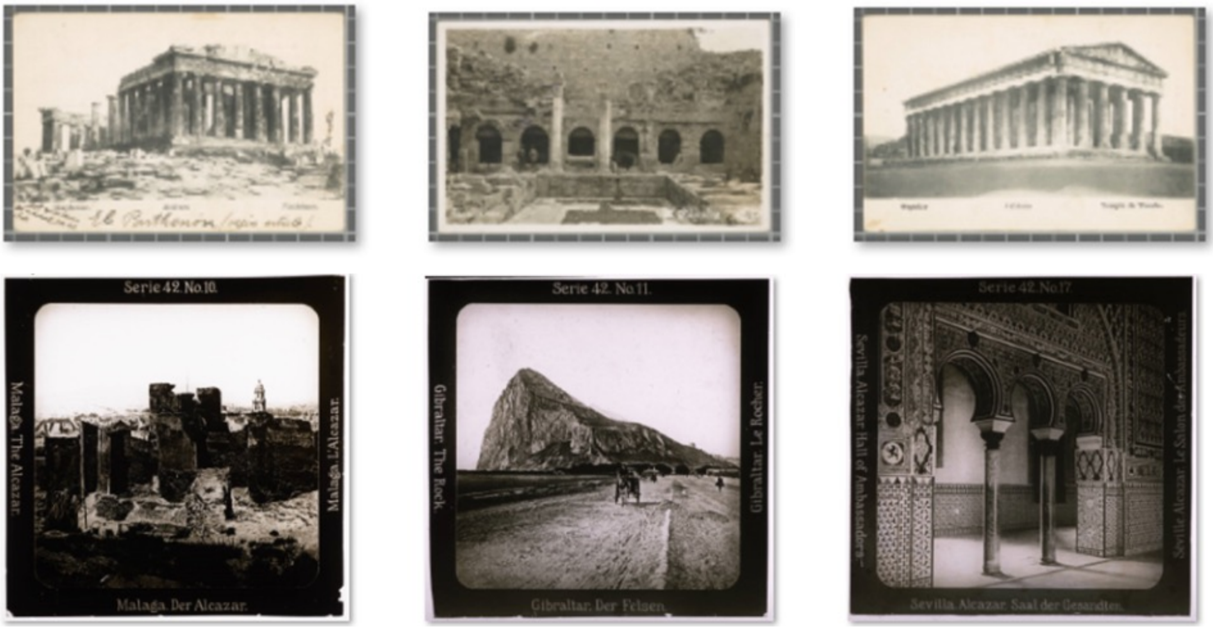


Figure 18: Scanned photos taken several decennia ago. Naturally many of them are of (very) low quality.

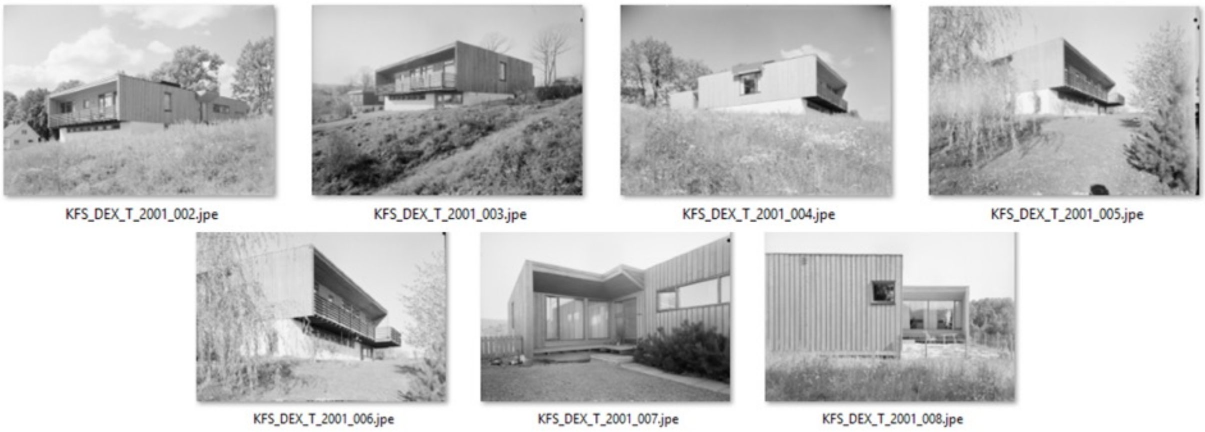


Figure 19: Images from 1 scene are incrementally annotated in the filename. This is seen throughout data provided by EF. Image baseline is too high for 3d reconstruction for this case.

3.2.2 ArtFilms Ltd (AF)

The V4Design content provider AF has a collection of art films and documentaries that were made available for 3D reconstruction evaluation. The content consisted of both video footage and image collections.

Video footage

AF has presented standard video data, as well as video data from drones, an example of which is shown in Figure 20. They have indicated that the use of high quality drone footage will increase in future productions. Most of the other footage provided by AF did not contain any parallax (moving cameras). A single exception to this was the footage from the Sydney opera house taken from a boat (Figure 7).

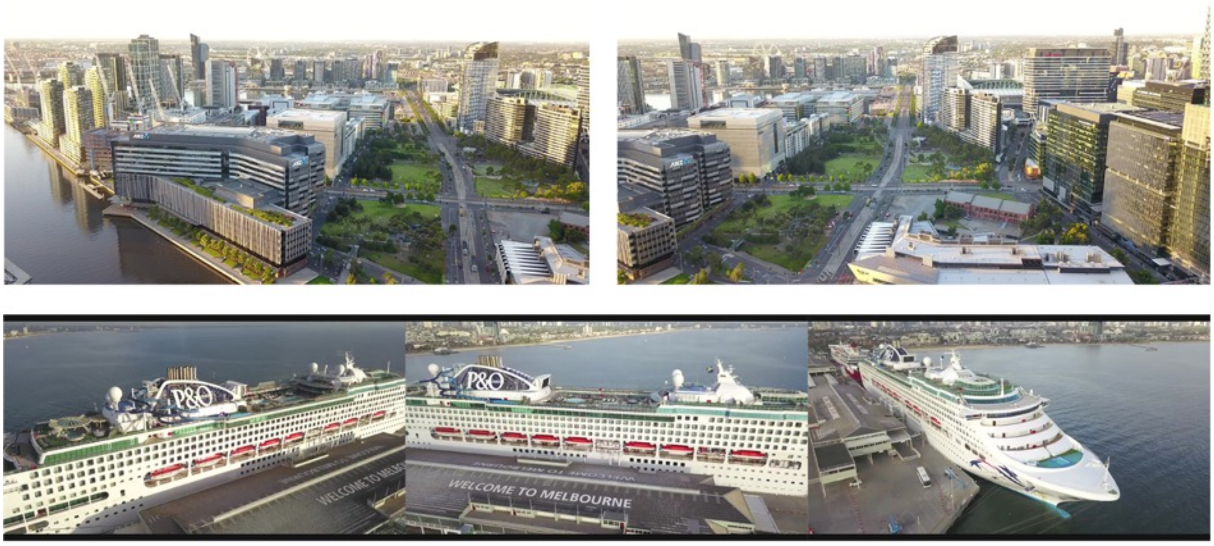


Figure 20: Drone footage from our content providers (AF)

Image collections

AF additionally provided a dataset of several photos. These images are provided in groups/batches of 10-1000. The majority of these images are not suitable for 3D reconstruction since only 1 image per scene is available (no stereo- or multi-view). Cases where multiple viewpoints from a single scene were available suffer from a generally low baseline (see *3.1.2Accuracy impacting factors*) as seen in Figure 21.

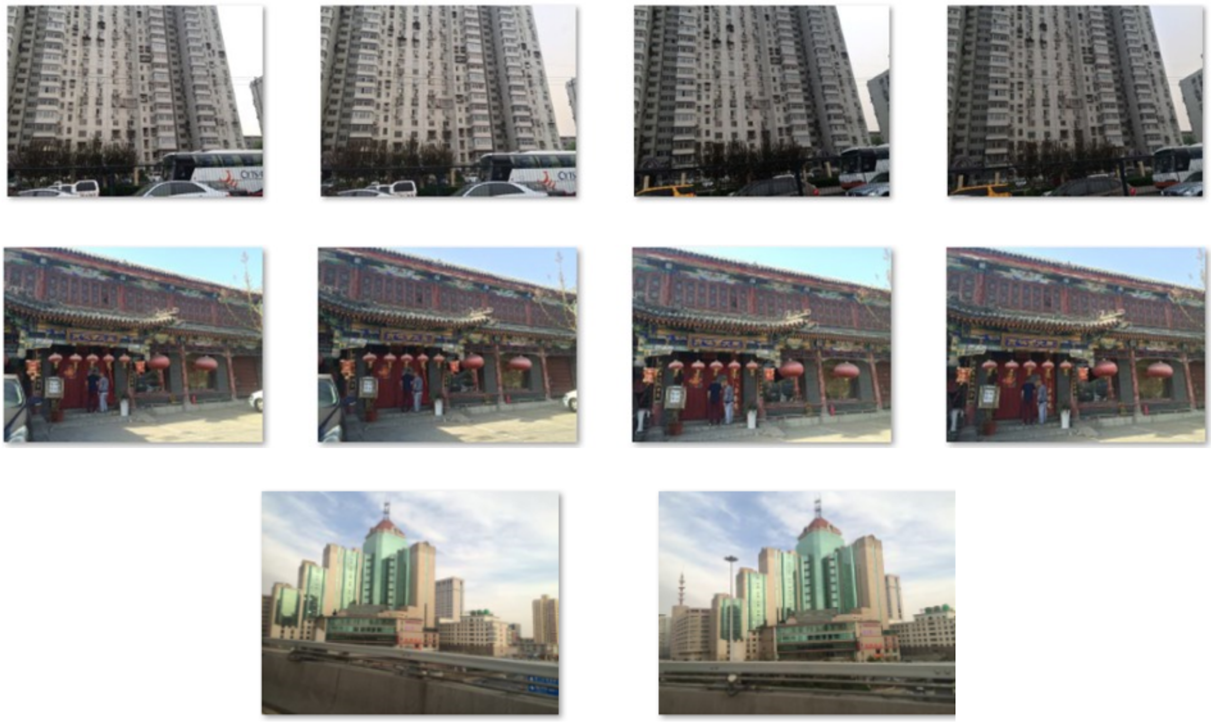


Figure 21: Some images provided by AF containing images from multiple viewpoints

3.2.3 Deutsche Welle (DW)

DW has been the main content provider for PUC 4: reconstruction of the Gendarmenmarkt (Figure 22), as well as PUC 3: Nico's Weg.

The Gendarmenmarkt

DW has been the main provider for data concerning the Berlin Gendarmenmarkt. This data was extensively used in several dissemination activities ('Digital assembly 2018' – Sofia, 'MADVR 2018' - Munich and 'Gamescom 2018' – Köln). Here a VR representation was created from the resulting 3D reconstruction of the Berlin Gendarmenmarkt. The Gendarmenmarkt square was successfully reconstructed from a variety of image sources provided by different providers (Figure 22), around 90% of the input data originated from DW. Additional data was provided by EF and CERTH, which respectively consists of several historical images (see Figure 17) and re-usable video material from online repositories.

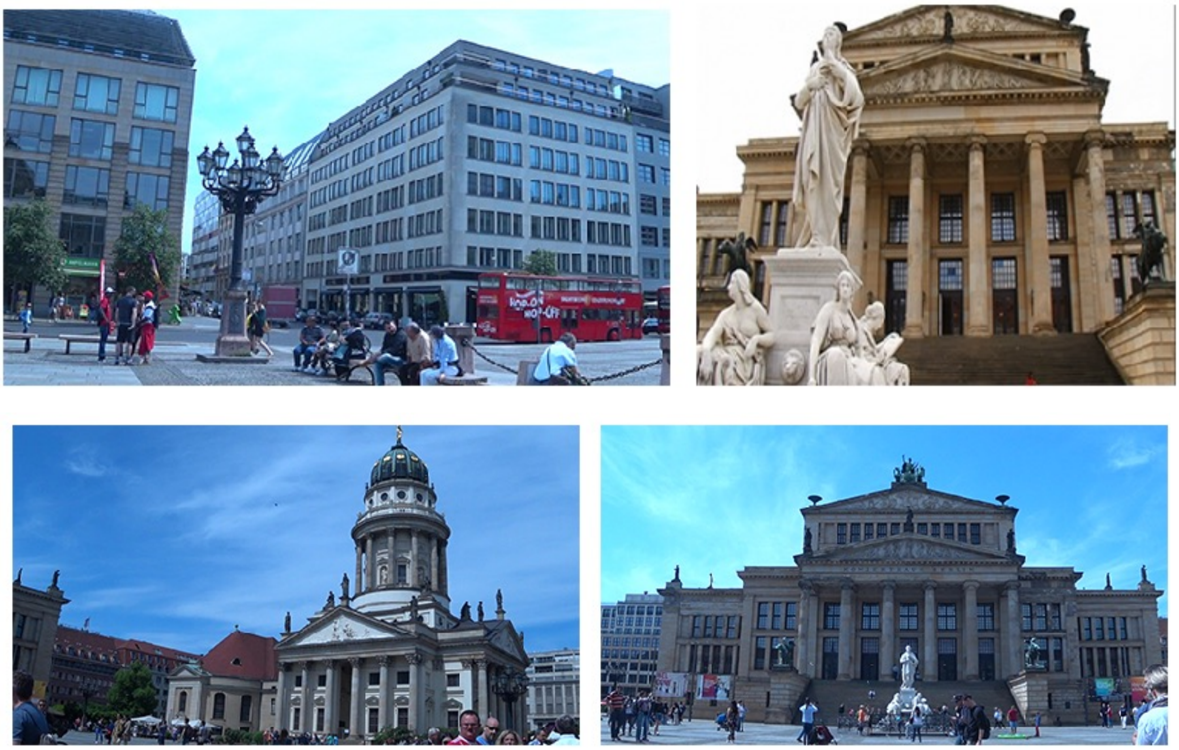


Figure 22: Footage provided from the Gendarmenmarkt square in Berlin. This contributed to our work in PUC4 and was presented in several dissemination activities.

Nico’s Weg

Nico’s Weg is a German language course organized around a series of video telling the story of Nico who recently moved to Germany. Each episode is a couple of minutes long. The method of filming is similar to a typical sitcom. Being a language course the main camera target are the actors and their conversations, filmed in an ‘over the shoulder’ type of way. Here the camera is generally in close proximity to the actors, causing areas behind them to be blurry (due to lens focus) as seen in image Figure 23. For this reason Multiview reconstruction techniques are not applicable. No exceptions have been found throughout the series where reconstruction is possible.

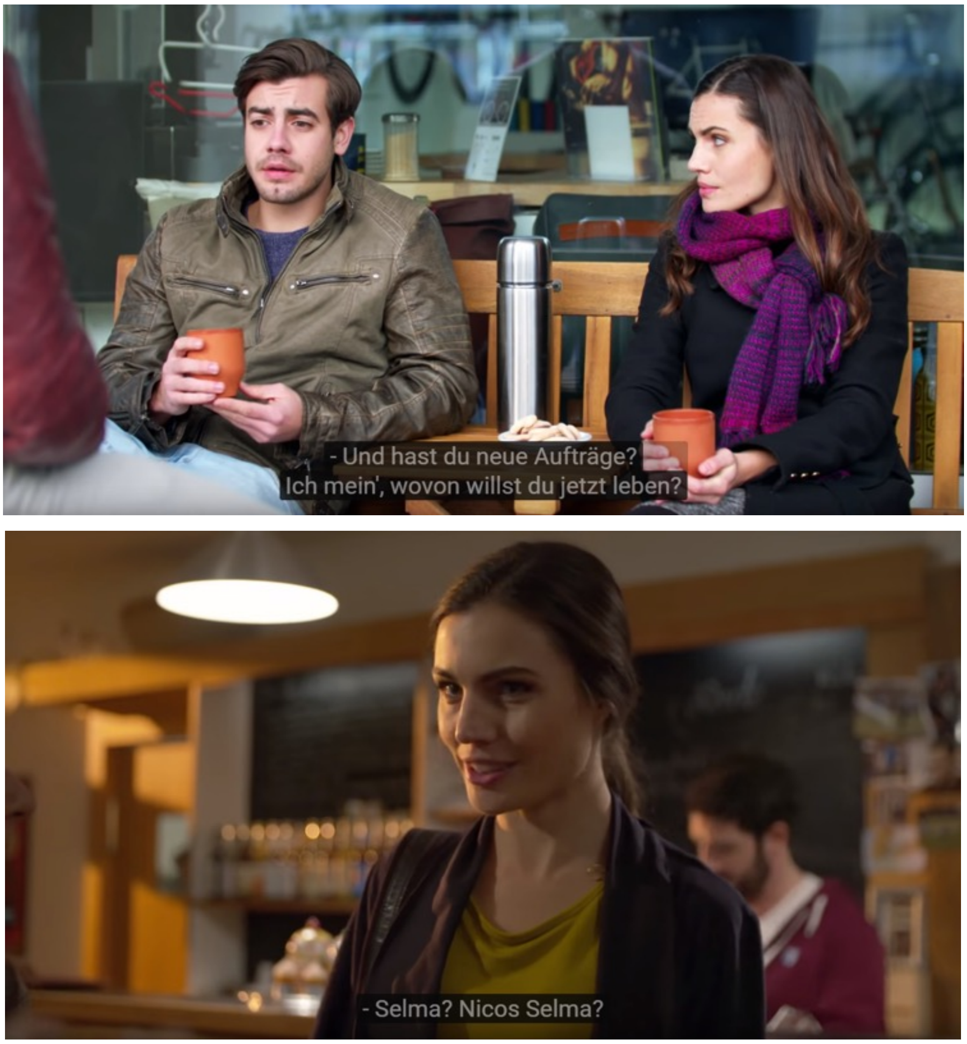

- Und hast du neue Aufträge?
Ich mein', wovon willst du jetzt leben?
- Selma? Nicos Selma?

Figure 23: Nico's Weg. Top: typical ‘over the shoulder’ filming. Bottom: camera focus on actors.

Other data

As a news agency DW continuously records and provides new material. One of their series called #DailyDrone is of particular interest. This series is a daily bird's-eye view of Germany (Figure 24). (Nearly) every day a different exciting location is recorded by the viewfinder of a drone camera, mainly focusing on famous sights in Berlin, Cologne, Hamburg and Munich.

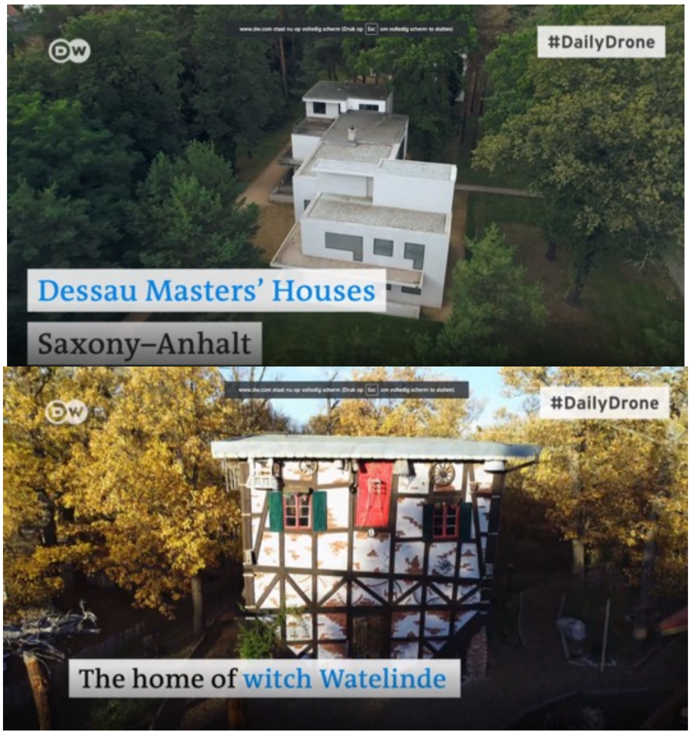


Figure 24: #DailyDrone series providing a bird's-eye view of different interesting locations, recorded by a drone

Finally, DW has provided some, very limited, 3D model data (Figure 25) from interior buildings as well. Such data could be useful for KUL's objectives in task 4.4.

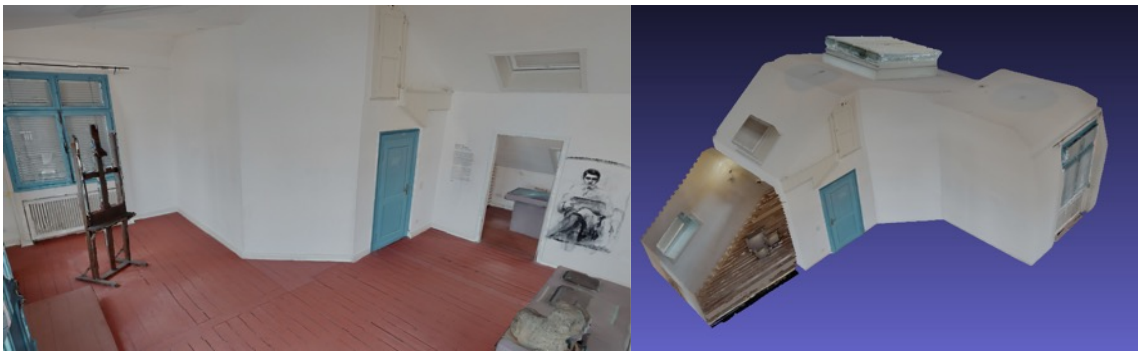


Figure 25: Small scale 3D model provided by DW

3.2.4 SLRS Multimedia (SLRS)

SLRS provided several full length documentaries to the V4Design dataset. However, due to the nature of how these documentaries are filmed, nearly all scenes are captured by a tripod camera. As discussed in section 3.1.1, no 3D information can be extracted from these scenes. Two scenes with a moving camera were found. For the *Microtopia* documentary two shots were taken from a moving train (Figure 26), providing enough camera movement for 3D reconstruction. These shots however do contain several obstructions such as passing trees and railroad infrastructure. Though this does not prohibit depth computation it may have a negative impact on the resulting texture, where some of these unwanted objects may get projected on the reconstructed facades.

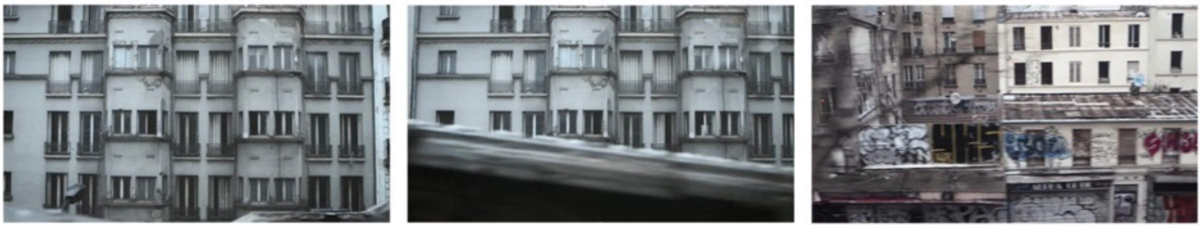


Figure 26: Train shot from Microtopia

A second documentary *Kochuu* contains a moving camera scene showing a Japanese ceiling (Figure 27) and an architectural building taken from a forest. Production of this documentary took place during 2003, tailored for traditional PAL and NTSC video formats. For this reason the documentary was provided with a low 774x480 resolution, making edges blurry (Figure 28).

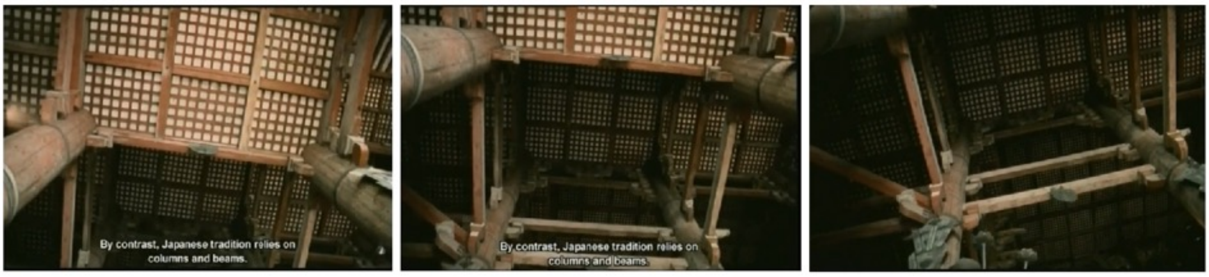

By contrast, Japanese tradition relies on columns and beams.
By contrast, Japanese tradition relies on columns and beams.

Figure 27: Multiview scene in 'Kochuu' documentary

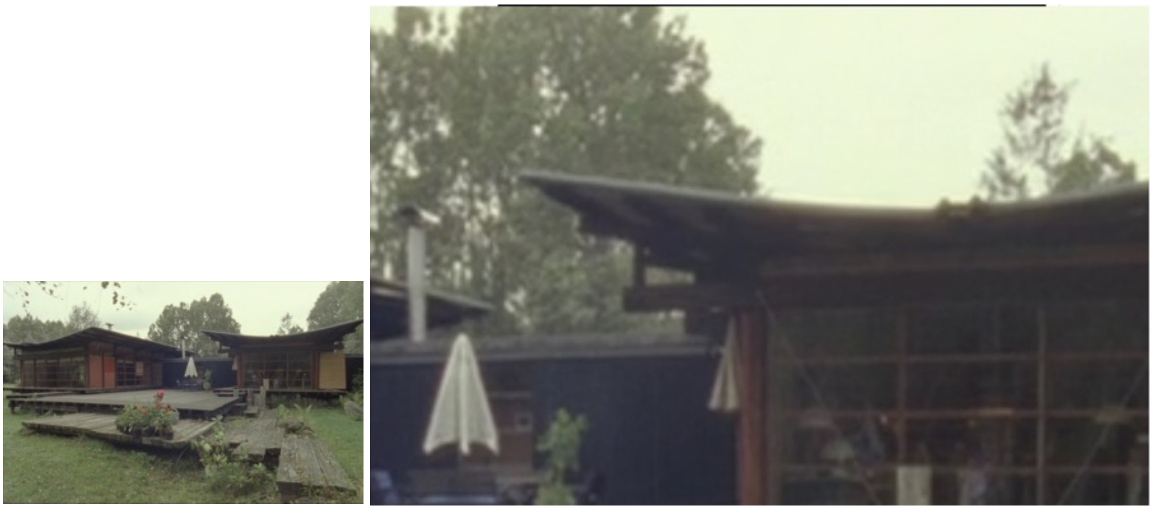


Figure 28: frame from 'Kochuu' measuring 744x480 pixels. Right: Zoomed in section. Edges have no sharpness.

Other documentaries such as ‘Great Expectations’ (Figure 29) lack any moving cameras.

Figure 29: Overview of ‘Great expectations’ documentary shots. These shots were taken from a tripod and thus provide no depth information.

From this initial empirical study we conclude that current data provided by SLRS may produce no more than a handful of 3D reconstructions. It remains to see whether these resulting reconstructions will be of high enough quality for the user due to the limited viewpoints (section of a ceiling, single façade, etc.).

3.3 Alternative approaches to 3D and information extraction

From the discussion in sections 3.1 and 3.2 it becomes clear that the amount of data, provided by the project’s content providers that is suitable for 3D reconstruction using Multiview methods is limited. The V4Design technical partners will investigate different solutions to this problem and this chapter will discuss the empirical analysis of two such solutions: the analysis of party from 3rd parties and the analysis of imagery of partners for other, single-view reconstruction methods.

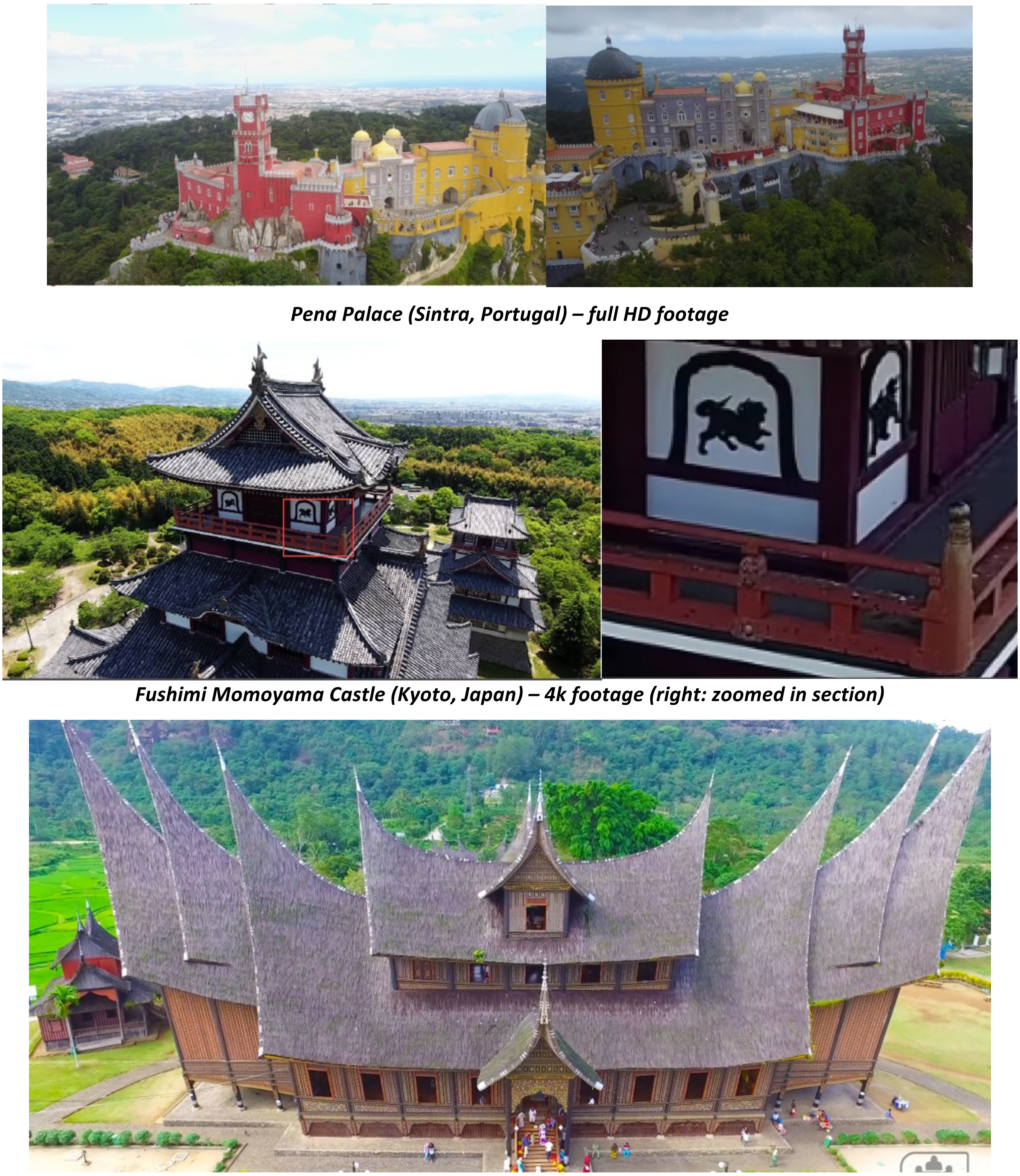
3.3.1 3rd party data analysis

In addition to content provided by V4Design content providers, material from 3rd party video repositories was also adopted in the initial content dataset. Potential copyright liabilities regarding this data are discussed in D2.1 (chapter 6).

YouTube data

The first iteration mainly focused on YouTube and specifically searched for video material from drone recordings. YouTube currently provides the biggest online video repository. Users around the world are free to upload videos in very high quality up to 8K resolution (7680×4320 pixels), which is very interesting for image-based 3D reconstruction. A second advantage of such online repositories and their user friendly way of sharing video content is the high amount of unedited material. For a 3D reconstruction such input data that has undergone no alterations, such as cropping is key. The reason for focusing on drone footage is that it may be queried very easily and generally complies with the input requirements for successful Multiview 3D reconstruction (section 3.1.1). Drone data is typically very useful for large scale 3D reconstruction such as building, cities, ruins, boats ... The consumer drone market has risen significantly in recent years. Economically priced drones currently record footage in high definition or 4K resolution. Therefore such high quality data will become even more accessible in the coming years.

For the reasons above CERTH initiated an initial list of 43 available drone videos on YouTube. Initial tests deemed the majority of this data (37 videos) very suitable towards 3D reconstruction. The majority of this data is high quality (Figure 30) and will likely result in good quality models. Several videos consist of raw 4k drone footage (Fushimi Momoyama Castle, Istana Pagaruyung) which is ideal for high quality models.



Pena Palace (Sintra, Portugal) – full HD footage

Fushimi Momoyama Castle (Kyoto, Japan) – 4k footage (right: zoomed in section)

Pagaruyung Palace (Batusangkar, Indonesia) – 4k footage

Figure 30: Drone footage from online repositories from many well-known locations.

In some cases, the video material did not suffice:

Image compilation video (2 cases)

These videos are an edited compilation of several photos. This should not necessarily be an issue if unedited imagery is used. However, in the case of Figure 31, zoom and move effects were added to the photos. For this reason, original internal camera parameters cannot be calculated.

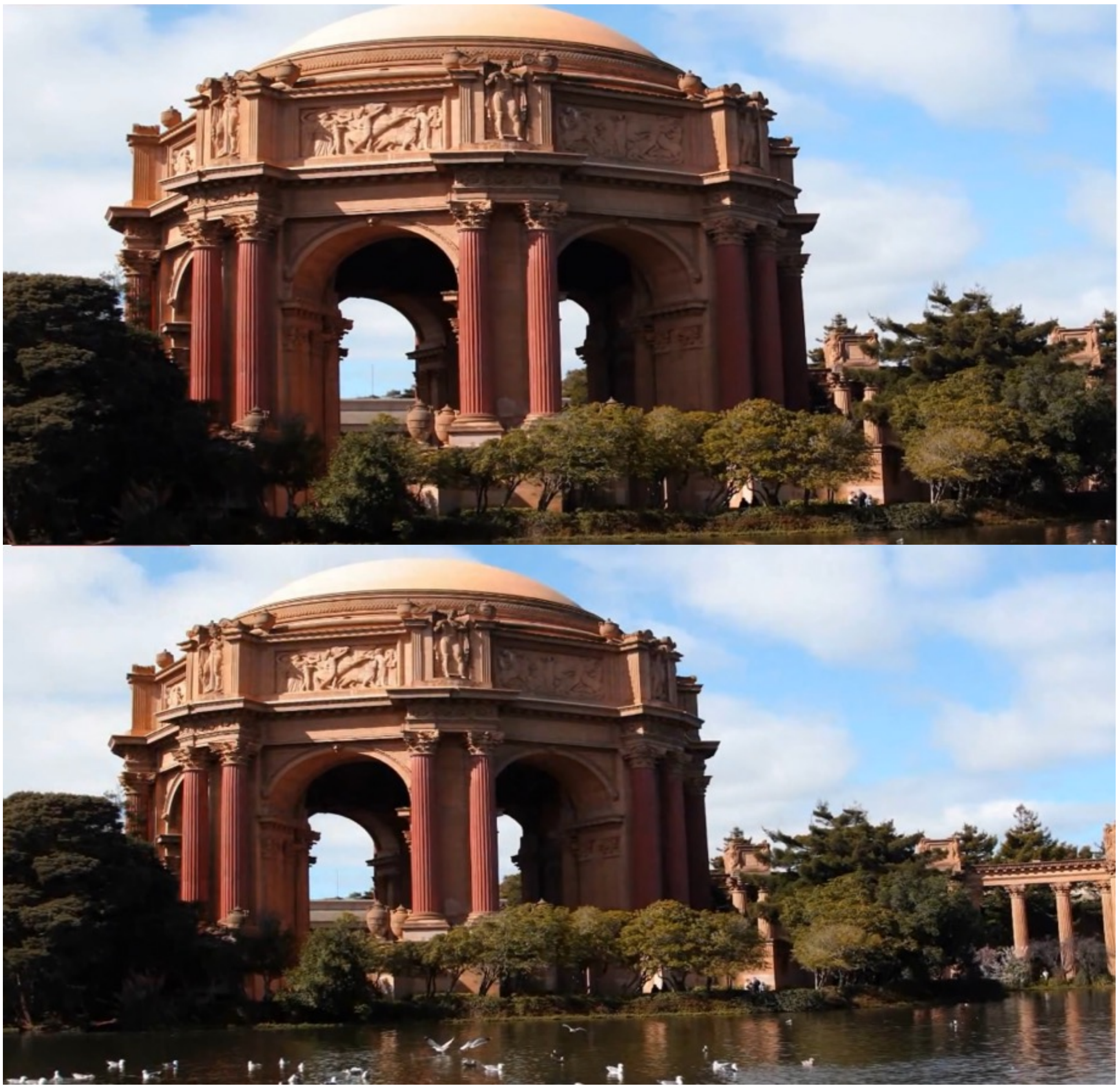


Figure 31: Edited video were a zoom and move effect was applied to a photo. No depth information can be calculated.

Thermal distortion / vibrating rolling shutter camera

In some cases the videos suffered from high thermal effects (Figure 33) or, causing similar effects, a rolling shutter camera sensor with high drone vibrations. This causes a continuous *ripple* effect on the video. An extreme example of this effect is shown in Figure 32, where the harp strings appear to be bendy. This effect, even mildly, could have a big impact on depth reconstruction as matched feature points between frames easily reach errors >5 pixels.

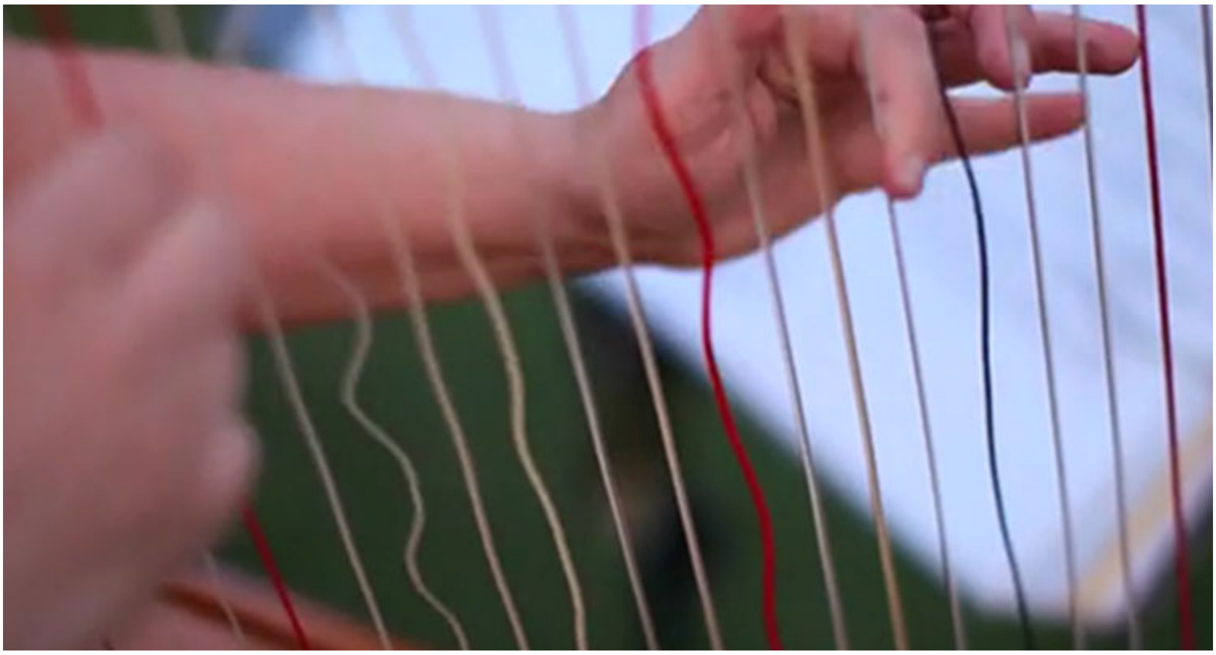


Figure 32: Rolling shutter effect. The harp strings are not straight.

Figure 33: One of the videos with high thermal / rolling shutter effect.

Suitable data for small-scale objects (bench, chair, etc.) reconstruction was also present in the first iteration (Figure 34). The recovered data was deemed very useful and suitable for 3D reconstruction. However automatic querying for this data is more difficult.

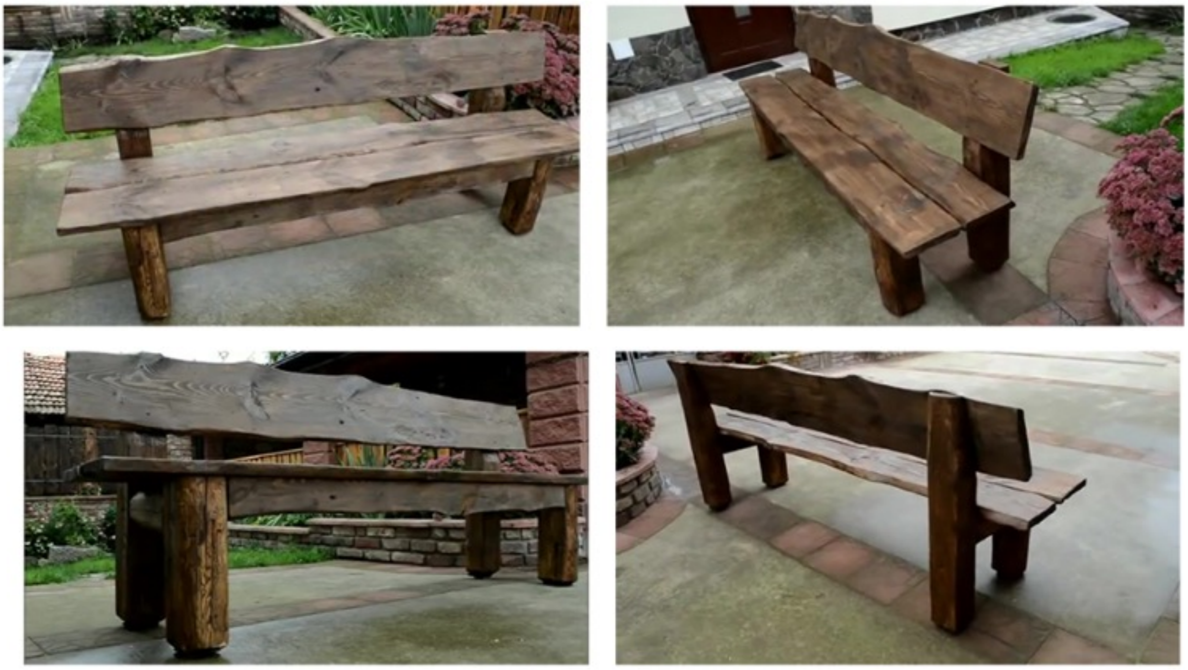


Figure 34: non-drone video frames extracted from youtube

Online Image repositories

A second possibility to find data for certain structures is searching for images (rather than videos) online. Similarly to video repositories, online image repositories exist containing material that is free for repurposing. The number of resulting images when querying specific objects (such as buildings or statues) is high and outperforms the video search by several numbers of magnitude. Figure 35 for instance shows the results for a query on "Eiffel Tower". Some important assessments can be made:

- The number of retrieved images is high, especially for well-known and much photographed buildings, such as the *Eiffel Tower* or the *Istana Pagaruyung*.
- The queries typically result in a high percentage of unusable imagery that is tagged with the search string but contains other information. The search for the Eiffel Tower for instance yielded imagery of similar towers in the US, a view of the Paris skyline and pictures of miniature replicas (Figure 35 bottom row). Our analysis showed that only 28% of the retrieved imagery actually depicted the actual Eiffel Tower and is useful for 3D reconstruction (Figure 36).
- In many cases the retrieved images contain metadata information about the camera and the lens. This is important because such information aids the reconstruction process. In the case of the Eiffel Tower, approximately 85% of the retrieved imagery contained information about the focal length (Figure 37).
- The resolution and image quality from digital photos is higher than video material. This can be seen in Figure 38.
- The use of an automated crawling unit for images limits the possible objects that can be queried to uniquely identifiable objects.

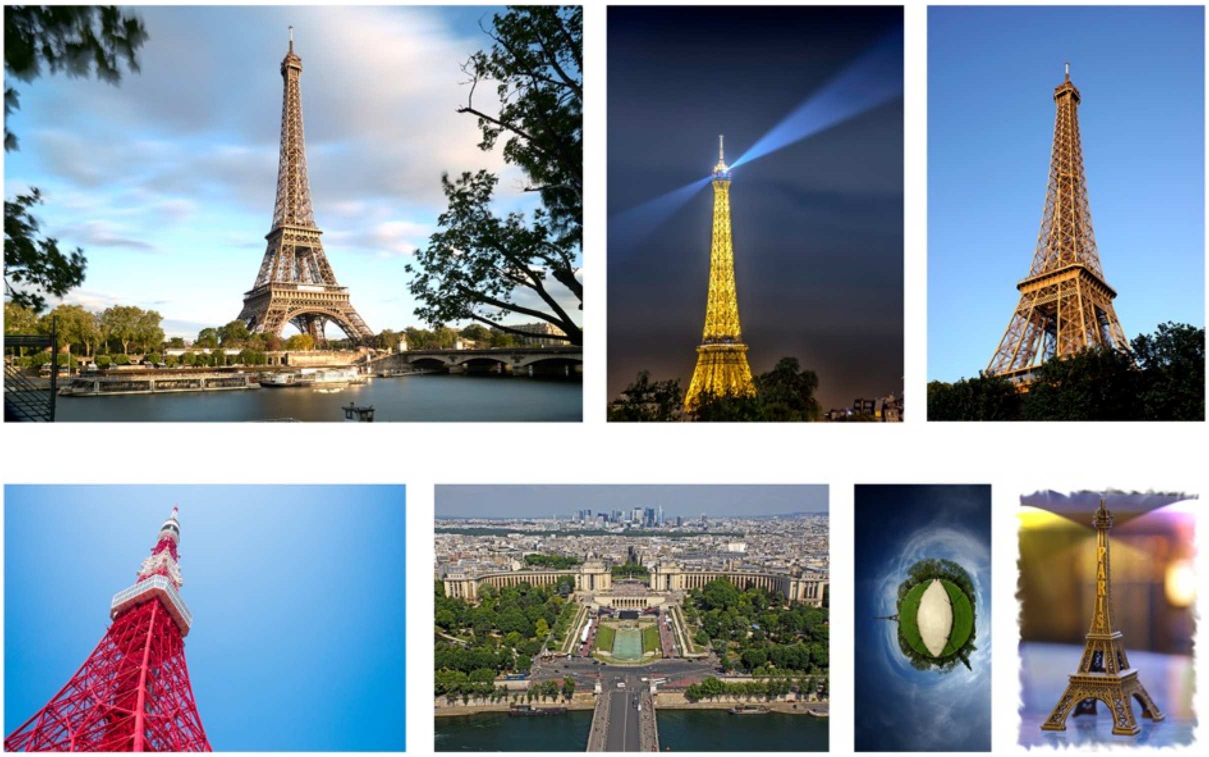


Figure 35: Online crawling results from Eiffel tower. The bottom row shows many undesirable results are present.

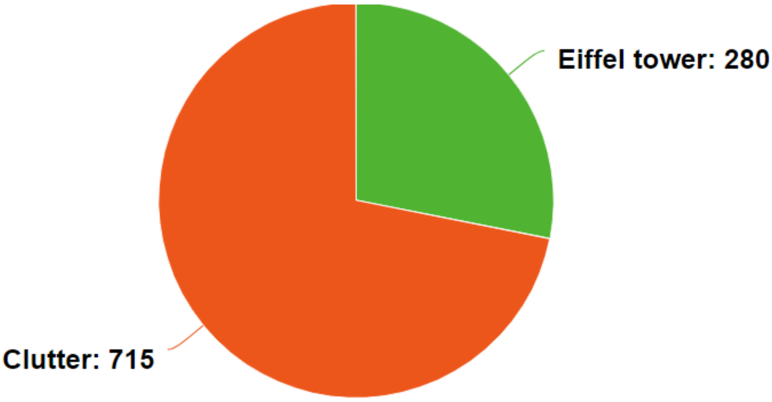


Figure 36: Analysis of the retrieved imagery for a search on “Eiffel Tower”. Only 28% of the retrieved images showed the actual Eiffel Tower and are useful for 3D reconstruction.

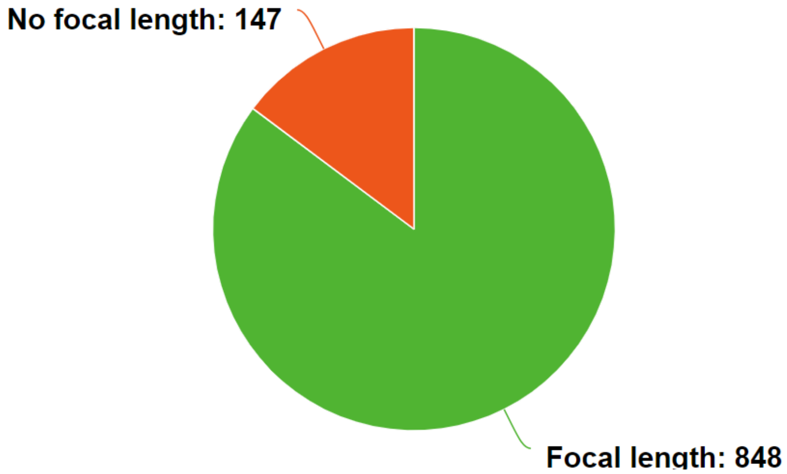


Figure 37: 85% of the images of the Eiffel Tower contain metadata information about the camera and lens.

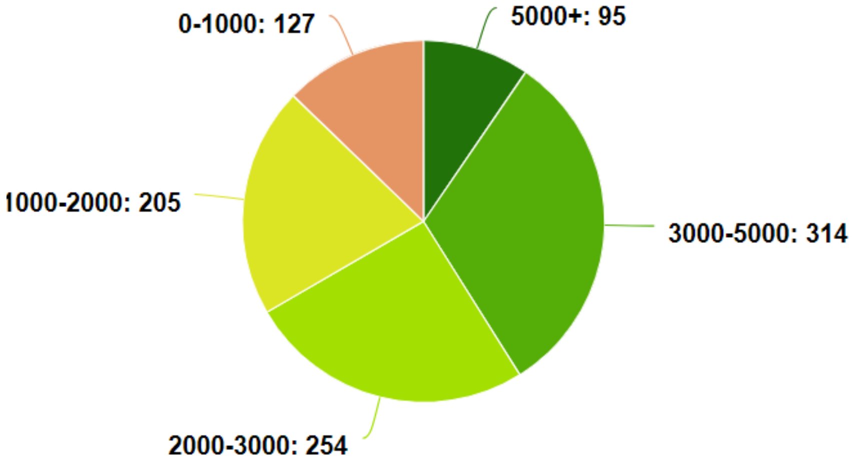


Figure 38: Distribution of photo resolution, expressed as the number of pixels of the image width. For comparison: full HD video = 1920 pixels, 4k video = 3840 pixels.

3.3.2 3D reconstruction with prior knowledge

As discussed in section 3.1.1, in order to retrieve correct geometrical 3D information from 2D image sources, it is mandatory to have recordings of scenes/objects from different viewpoints. In cases where only a single image is present no correct depth information can be determined by traditional Multiview reconstruction techniques. However, due to the substantial amount of such data that was found during the current iteration, a review of the current state of the art on single view reconstruction techniques was performed. All these single view reconstruction techniques rely on additional prior knowledge or assumptions of depth information. The term 'prior knowledge' refers to additional information of what exactly is present in the scene and additionally how these detected objects could be represented in a 3D environment.

The following single view datatypes were found in the current image database that contain the necessary prior information for 3D reconstruction.

Symmetrical items

Due to the specific nature of symmetrical items, reconstruction may be possible from a single view. More specifically surface of revolution **(SoR)** objects such as vases and pillars are potentially suitable. To reconstruct a surface of revolution it is important to determine its central axis and generatrix. Depending on the input image quality and camera type, reconstruction may be possible by detecting the contours and symmetry line of the object. By request of KUL, EF queried and collected an initial surface of revolution dataset (Figure 39).

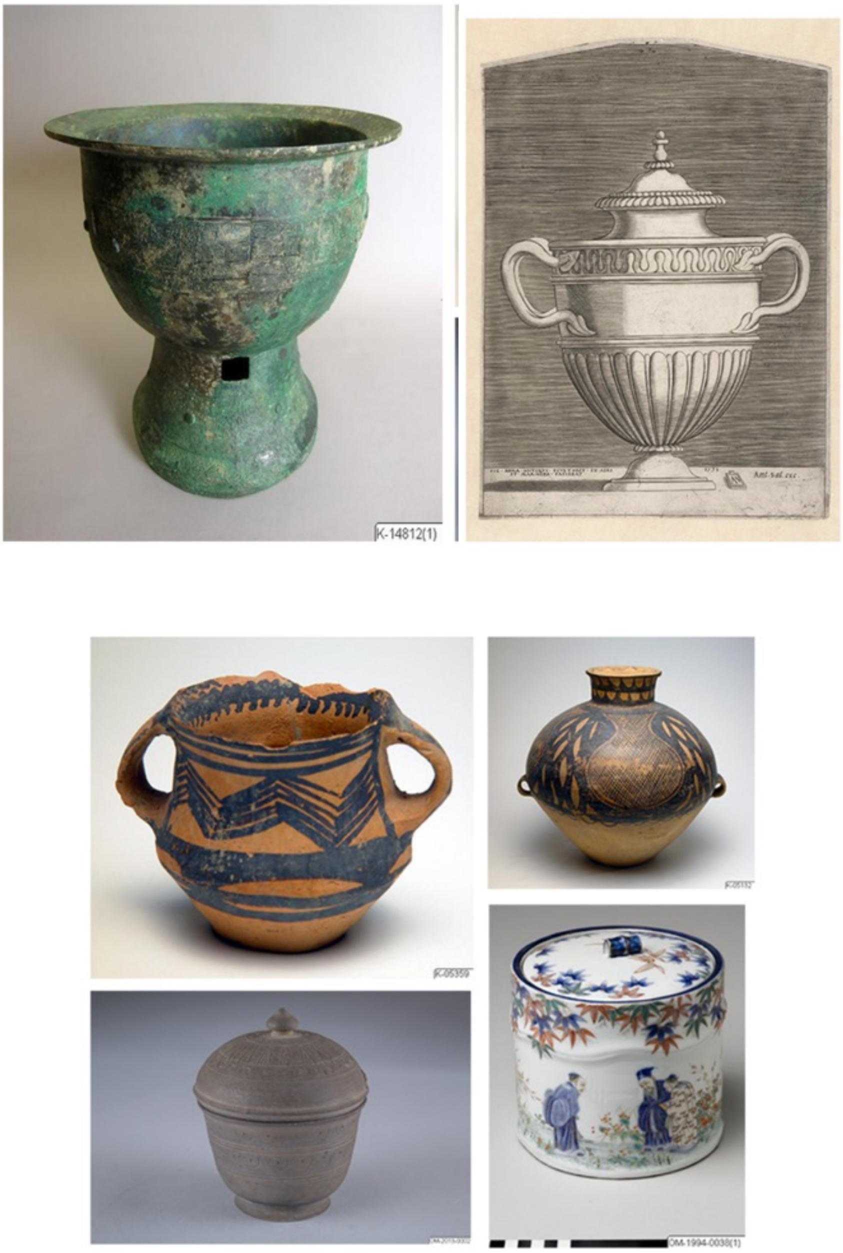


Figure 39: Several Symmetrical items provided by EF.

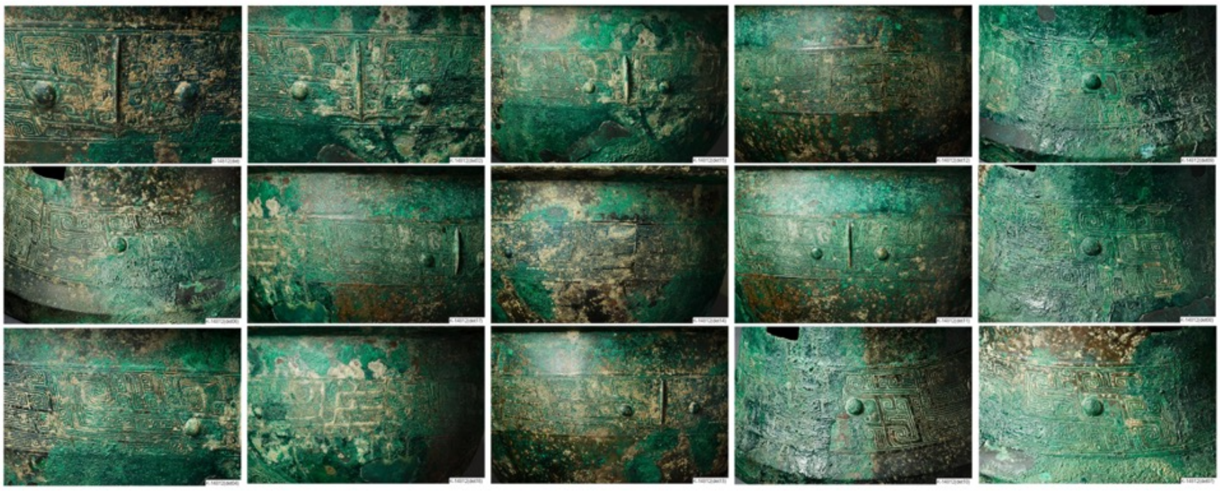


Figure 40: For some objects additional detailed images are provided. Detailed shots from the leftmost jar is Figure 39 is displayed.

A large portion of these images were provided by museums to the EF database. Many of them provide a homogenous background colour and high resolution, a requirement towards automatic reconstruction. Reconstruction of imperfections or parts such as handles that do not confine to the surface of revolution model may not be possible with this technique (Figure 42). Additional surface of revolution objects may also be found in data from other providers (Figure 41). However, in this case no prior knowledge on this data is given so an automatic detection algorithm should be implemented/researched for detection of SoR objects in scenes.

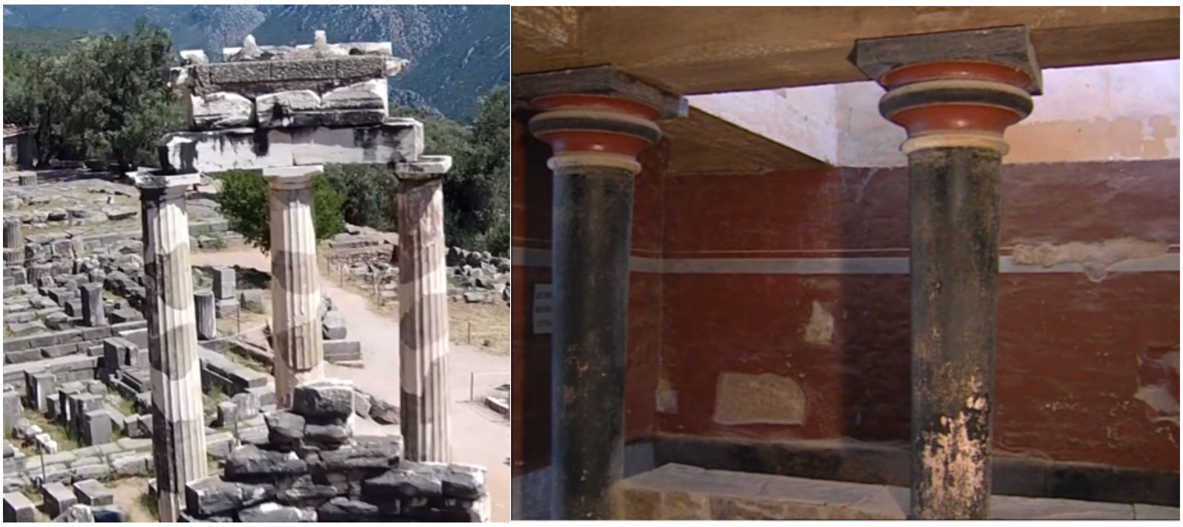


Figure 41: Pillars are a common SoR object. Left: Delphi, Right: Crete (AF)

Other well-captured objects

In addition to surface of revolution objects many other single-view objects are present in the database.

One particular branch of current state-of-the art research focuses on single view reconstruction. Work by (Huang, 2018) focusing mainly on chairs requires any background content to be filtered out of the image. Based on prior knowledge of chair structure and the joints in them an algorithm attempts to localize those joints in the input image. Resulting 3D information (heavily based on the prior knowledge data) will not contain texture or complex forms, such as the shape of a Victorian era chair leg (Figure 42, left) but rather a set of joints and sticks between them.

Other research attempts single view 3D reconstruction based on a prior knowledge in the form of a set of 3D models. 2D renderings are processed from many different angles for a limited selection of these models. Attempts to match the 2D rendered images to the original input image are made. Any resulting 3D reconstruction will be a rescaled (aspect ratio) original model (without reuse of texture or the detailed shape).

Attempts for automatic implementation of such methods required well-defined metadata (e.g. description of object in image) to be present. Currently, only certain datasets from EF contain such metadata.

The limited output from these state-of-the-art methods generally containing an incorrect geometrical representation of the object and missing texture do not meet the user requirements of repurpose. In conclusion, this input data is not suitable for the 3D reconstruction task.

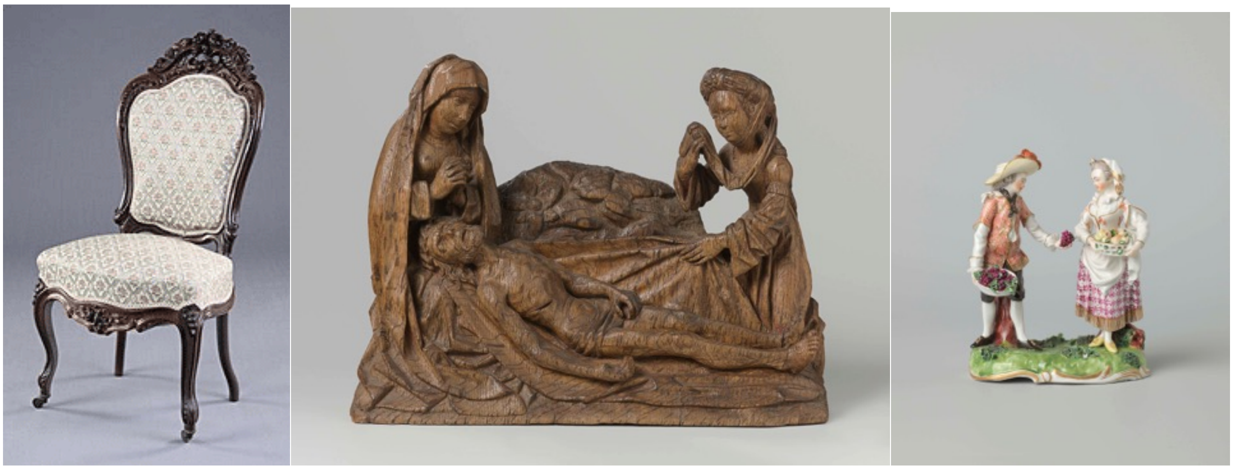


Figure 42: Non-symmetrical data provided. No 3d data can be extracted from a single image

4 EMPIRICAL STUDY OF THE CONTENT FOR SPATIO-TEMPORAL BUILDING AND OBJECT LOCALIZATION (STBOL)

4.1 Spatio-temporal building and object localization

The spatio-temporal building and object localization module is challenging because it requires the correct detection of objects or buildings in images or videos and the localization of each one of them using segmentation, where the goal is to classify each pixel into a fixed set of categories (objects). This module focuses on the content of images or videos, which are cut into batches of frames, in order to show to the end user the part of the image or video that contains a detected object or building. It is important for this module that the images, used as training or testing input depict buildings or objects. For training purposes, is also important for each dataset to be enriched with annotations for each image and video stream.

4.2 Internal Content

The Internal Content (IC) that was provided by V4Design partners and is related to STBOL is: IC001, IC002, IC005, IC006, IC007, IC008, IC009, IC010, IC011, IC012, IC013, IC014, IC015, IC016, IC017, IC018, IC019, IC020, IC021, IC022, IC023, IC024, IC025, IC026. We separate the datasets in two categories with respect to their modality: The first one contains datasets which are compiled from images while the second consists of videos.

4.2.1 Image content

It is important for the STBOL module that the input images contain buildings or objects. The IC001 dataset contains links to 63 object images with metadata focusing on Hans Scharoun. The dataset includes images that depict objects or buildings as presented in Figure 43 and are useful for the STBOL module, but also include many photos of Hans Scharoun himself, as the example in Figure 44, which could not be deployed, because the module focuses on buildings and objects and not on people.

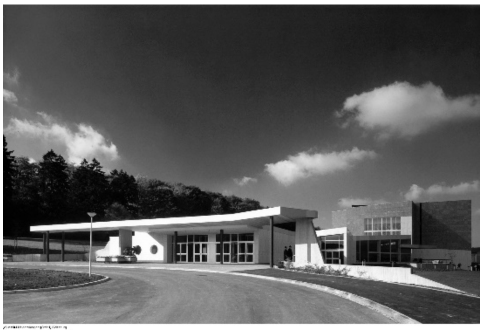


Figure 43: A sample of a building image in Hans Scharoun dataset.

Figure 44: A sample of image which depicts Hans Scharoun.

The IC002 dataset contains links to 838 objects and their metadata focusing on Hong Kong, Kowloon and Victoria Harbour but most of them are photos of text and maps (see for instance Figure 45) which are not useful for the STBOL module, because they do not contain any architectural objects.

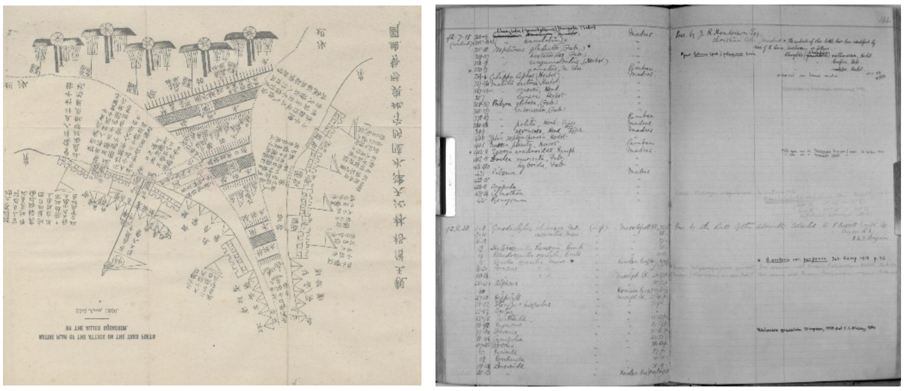


Figure 45: A sample image of text and map of IC002 dataset.

The IC005 dataset contains 1200 photos focusing on Japanese art. These images depict artefacts and buildings, which could be used for testing the STBOL module and could be useful for the implementation of PUC2, which focuses on designing interior architectural objects inspired from Asian art. Some examples are shown in Figure 46.

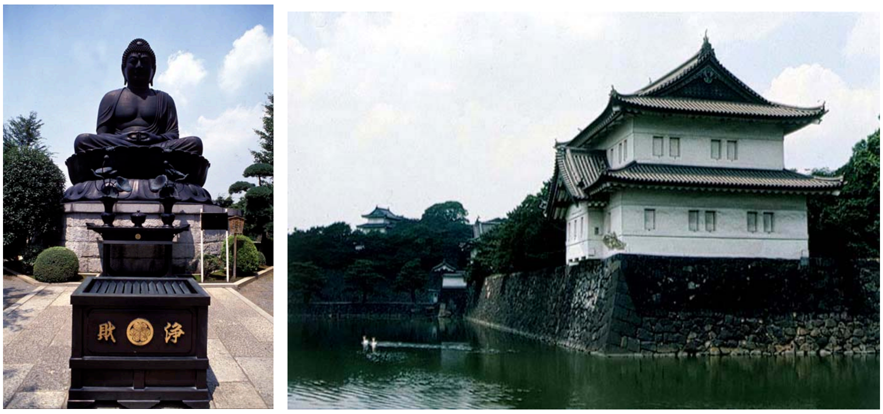


Figure 46: A sample of an artefact and a building from the IC005 dataset.

The IC007 dataset includes links to object and building images that could be fed to the STBOL module. In addition, the IC011 dataset includes 402 images in total. Some of these are historical photos from the garden and interior, sketches of decoration and statues which can be used to train our module and find interior and exterior architectural objects, but also includes photos of text about the palace which could not be used. Moreover, the IC012 includes 11 images and a video of buildings and objects focusing on Bruno Mathsson. Figure 47 depicts some of these images.

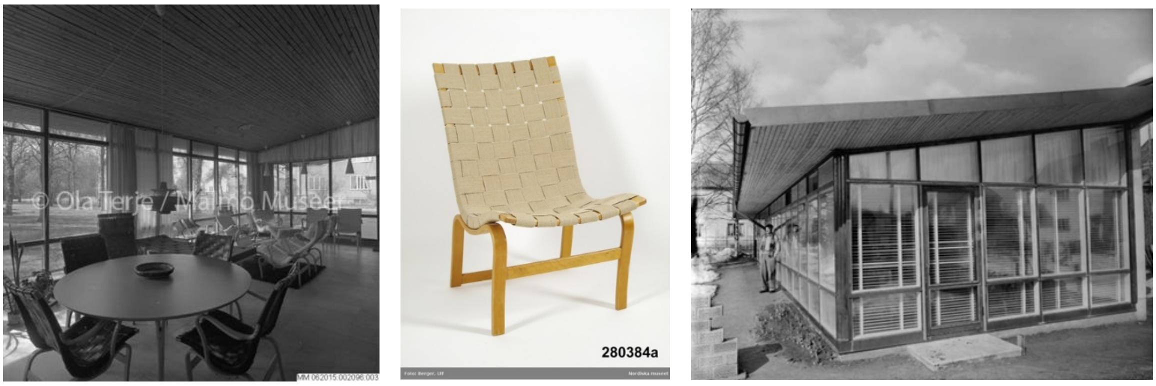


Figure 47: Images of objects and buildings from IC012 dataset.

Furthermore, the IC014 dataset includes 46 objects, among which some images depicting the Notre Dame church, which is very useful for PUC1. It also includes text, which could not be deployed by the STBOL module. The IC015 dataset includes 97 images and 2 videos of Hans Scharoun which are very similar to the IC001 dataset. IC015 also contains other useful images, as well as images which can not be deployed because they do not depict architectural objects. Another dataset is the IC019, which includes 326 images. 200 of these depict text which could not be used in the STOBL module but the rest depict objects and buildings of ancient Greek architecture focusing on Delphi, as shown in Figure 48 and are therefore useful for the STBOL module because the content is relevant to V4Design.

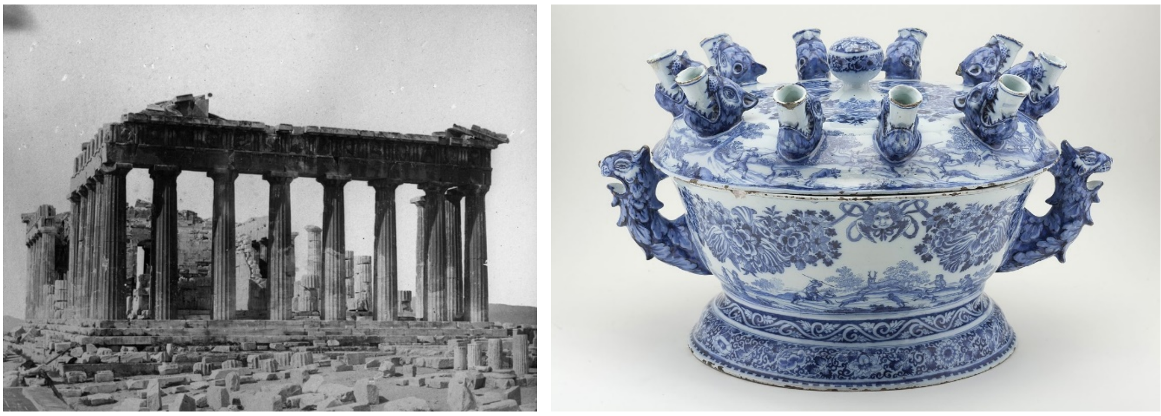


Figure 48: A sample of images of ancient Delphi.

The following five datasets also include images and videos with useful content for V4Design STBOL module. More specifically, the IC020 dataset, includes 33 images of German pre-1950 architecture focusing on Berlin, which is very useful for the implementation of PUC4 that focuses on reviving historical buildings from archival footage. The IC021 dataset, includes 23 videos focusing on different architectural structures which is useful for V4Design STBOL module. The IC022 dataset includes 2693 images focusing on Japanese and Chinese material design which are useful for object segmentation and for PUC2. The IC023 dataset, which consists of 600 object images from Japan, China and other Asian art is also helpful in order to test the STBOL module and finally the IC026 dataset contains images focusing on the

Gendarmenmarkt and is a very useful dataset for the module and PUC4 (i.e. VR and historical buildings).

4.2.2 Video content

In this section we discuss the video content, such as movies, episodes and personal collections that can be used for the spatiotemporal analysis towards the detection and localisation of buildings and objects. First, the IC006 dataset includes videos with Japanese art that can be used by the STBOL module in order to deploy PUC2. The IC008 dataset includes videos of architectural buildings. The IC009 dataset includes videos of buildings, such as the ones depicted in Figure 49, and can be used so as to extract objects and buildings from archival material, while the IC016 dataset reiterates the IC009 dataset introducing annotations. Figure 49 shows frames from the "Great Expectations" video from the IC009 dataset.

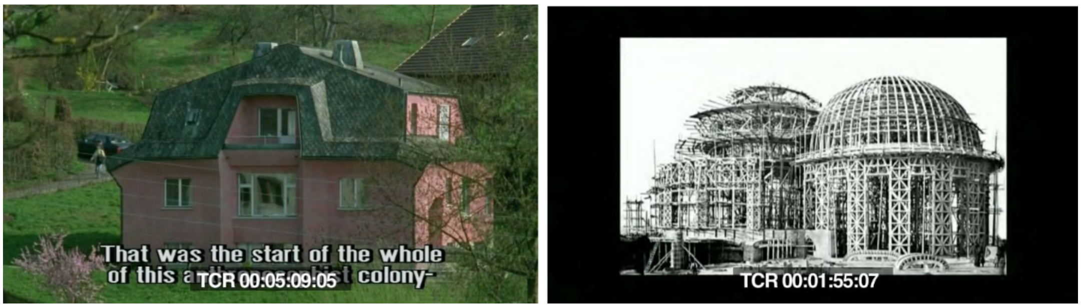


Figure 49: A sample of frames of the Great Expectations video from IC009 dataset.

The IC010 dataset includes 230 episodes of the DW series "Nico's Weg", including the trailer, which are very useful for the implementation of PUC3. We have already used a sample of this dataset and in Figure 50 we present the results of our module. The IC021 dataset includes 23 videos focusing on different architectural structures. The videos depict many different buildings and objects, and are useful for the STBOL module because their content is also relevant to V4Design use cases. The IC025 dataset includes 230 video objects of Nico's Weg web-based exercises.

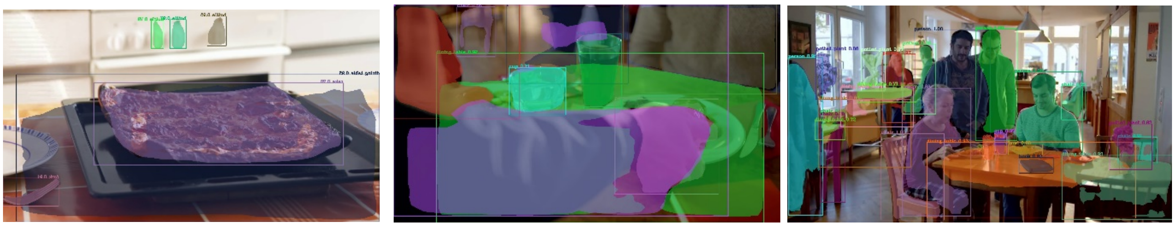


Figure 50: A sample of results of our model on images from the Nico's Weg series.

The IC013 dataset includes video clips that show buildings and objects, while the IC017 dataset includes over 600 short clips, 2-5 minutes focusing on screeners for contemporary art. The IC024 dataset includes 8 videos with architecture objects of Bali, Japan and China but the most frames depict people (as presented in Figure 51) and are not useful for V4Design.

Figure 51: Samples of video frames of IC024 dataset, depicting people and therefore not interesting for V4Design

4.3 External content

The External Content (EC) that has been collected from free web sources and other benchmarks datasets and is related to STBOL, is present in the datasets with the following IDs: EC002, EC006, EC007, EC008, EC009, EC010, and EC011. We will use this content to build our initial object detectors and scene recognition algorithms. Below we provide some more details about the different datasets.

The EC002 dataset contains 663 images crawled from the aforementioned Wikipedia webpages and is considered a small dataset, while the EC006 is the Microsoft Common Object in Context (COCO) dataset (Lin 2014) which has more than 200,000 images that depict 91 categories that occur in both indoor and outdoor objects. This dataset also contains both bounding boxes and object segmentation output. A sample is presented in Figure 52.

Figure 52: A sample of images of COCO dataset.

The EC007 dataset is the Open Images (storage.googleapis.com/openimages/web) dataset of around 9 M images that have been annotated with image-level labels and object bounding boxes. This training set contains 14.6M bounding boxes for 600 object classes on 1.74M images, making it the largest existing dataset with object location annotations today. The boxes have been largely manually drawn by professional annotators to ensure accuracy and consistency. The images are very diverse and often contain complex scenes with several objects (8.4 objects per image on average). Moreover, the dataset is annotated with image-level labels spanning thousands of classes.

The EC008 dataset is the ImageNet (www.image-net.org) dataset, a large visual database designed for use in visual object recognition software research. Over 14M images have been hand-annotated by ImageNet to indicate what objects are depicted; in at least one million of the images, bounding boxes are also provided. ImageNet contains over 20 thousand categories. Currently, bounding boxes for over 3000 popular synsets are available. For each synset, there are on average 150 images with bounding boxes.

The EC009 dataset is the Places2 (places2.csail.mit.edu) dataset, which contains more than 10M images comprising over 400 unique scene categories. A sample of the Places2 dataset is depicted in Figure 53.

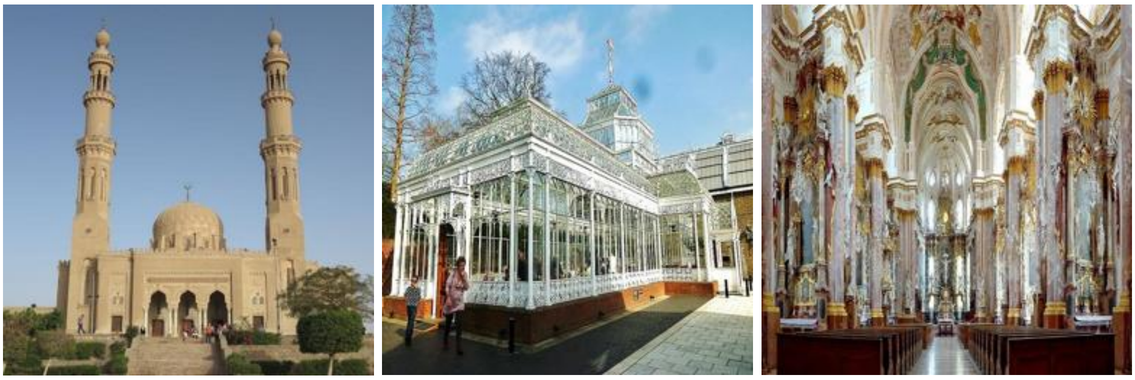


Figure 53: A sample of buildings of Places2 dataset.

The EC010 dataset is the SUN397 (groups.csail.mit.edu/vision/SUN) database containing 397 categories of environmental scenes, places and the objects within. The number of images varies across categories, but there are at least 100 images per category, and 108,754 images in total. Finally, the EC011 is the Oxford Buildings dataset which consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. Figure 54 shows several samples.

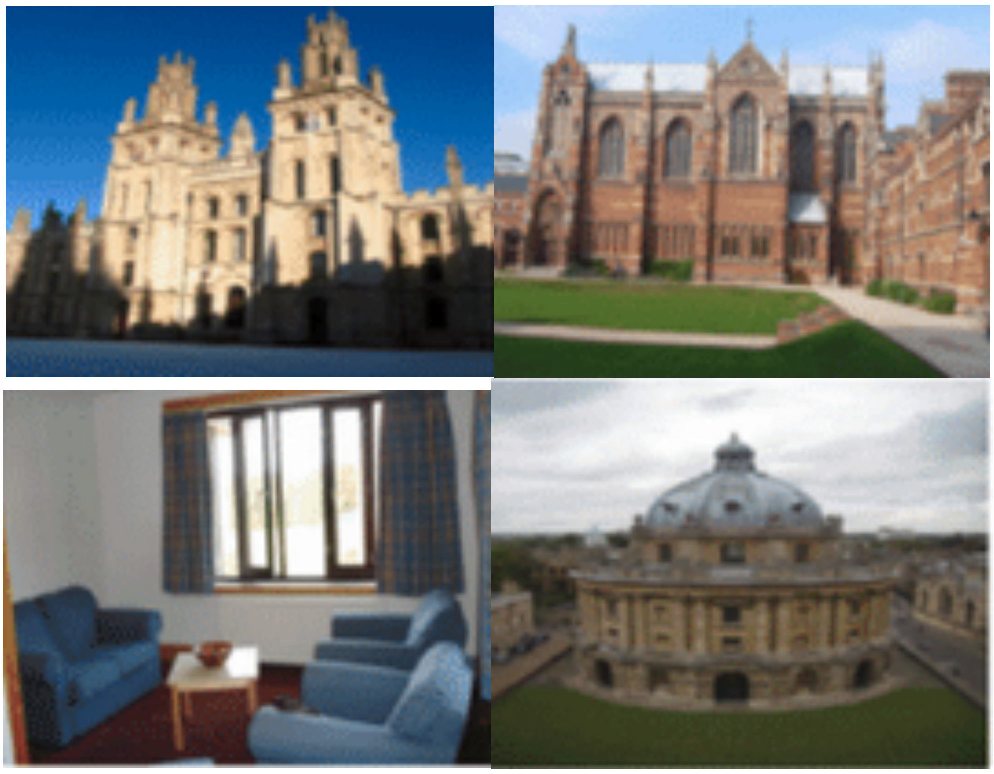


Figure 54: Sample of Oxford buildings dataset.

This collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This results in a set of 55 queries over which an object retrieval system can be evaluated.

5 EMPIRICAL STUDY OF THE CONTENT FOR AESTHETICS EXTRACTION AND TEXTURE PROPOSALS (AE&TP)

5.1 Aesthetics extraction and texture proposals

Aesthetics extraction is a module, which is based on computer vision classification methods, aiming to classify each image with respect to style or creator. The model requires a significant amount of data in order to be trained. Therefore, many painting images with annotation metadata included per class and per creator are expected to support the development of the AE module. Regarding the TP part of the module, it requires as input one content image and one style image in order to generate one new image with the content of the first one and the style of the second one.

5.1.1 Internal content

The Internal Content (IC) that have been given from V4Design partners and is related to AE&TP: IC029 and IC032. Both datasets have no information about style, genre and creator at the moment, but they can be used for testing V4Design modules and then they will be evaluated by experts. The IC029 dataset includes 13,988 images focusing on collection of pre-1950 painting artworks in EF, for AE&TP tasks. The model for style classification requires a number of images per class for training, validation and testing. The IC029 dataset is suitable for testing purposes. Two samples are shown in Figure 55.

Figure 55: A sample of two painting images that are contained in IC029 dataset.

The IC032 dataset contains 41,083 painting images from EF-Wiki extracted from their SPARQL endpoint containing paintings with at least one associated genre or style tag. Two samples are shown in Figure 56.

Figure 56: A sample of two painting images that are extracted from Wiki.

5.1.2 External content

The External Content (EC) that has been collected from free web sources and other benchmarks datasets and is related to AE&TP consists of EC012, EC013 and EC014**.** We used this content to build our initial aesthetics extraction and texture proposals algorithms.

The Pandora (Florea 2016) (EC012) paintings dataset has a collection of 18,720 paintings from many different sources. The image collection has been distributed among 18 style classes, having approximately 1,000 images. Engineers have ensured that only the relevant part of images is shown and art experts also ensured that the artistic annotation is valid. Figure 57 illustrates a sample of the Pandora dataset from the style "cubism".

Figure 57: A sample of painting images from Pandora dataset which belongs to Cubism style.

The Paintings-91 (Khan 2014) (EC013) dataset contains 4,266 painting images from 91 different painters. The artists in this dataset belong to different eras. There are a variable number of images per artist, ranging from 31 to 56. The large number of images and artist categories make the problem of computational painting categorization extremely challenging.

The Wikiart (www.wikiart.org) paintings (EC014) dataset is an image collection of 81,472 paintings images, from more than 1,000 artists. This dataset contains 27 different styles and 45 different genres. To our knowledge, this is currently the largest digital art dataset publicly available for research purposes. 81,446 paintings are used for style classification, while only 10 genres with more than 1,500 paintings are chosen, with a total number of 64,995 samples. Similarly, only a subset of 23 artists with more than 500 paintings is chosen, with a total number of 19,051 images for artist classification. A sample from the Wikiart dataset is presented in Figure 58 and Figure 59**.**

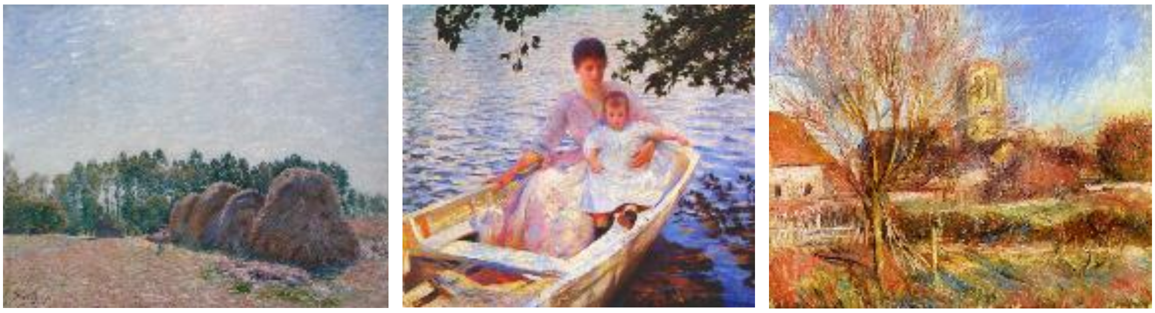


Figure 58: A sample of painting images from Wikiart dataset belongs to Impressionism style.

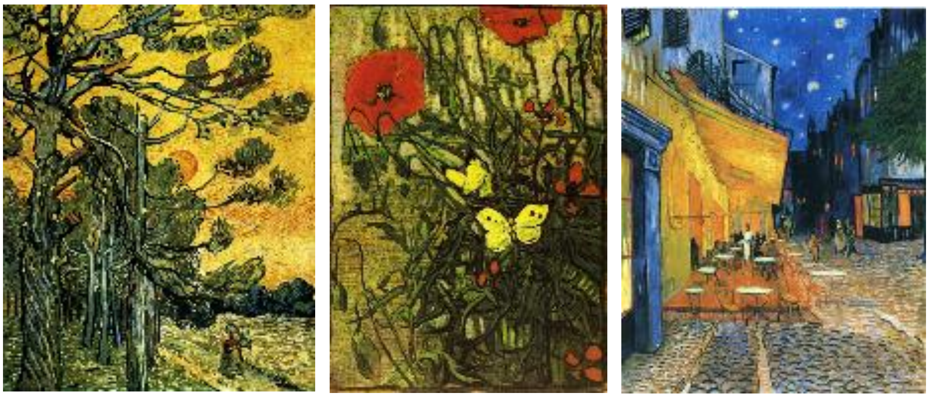


Figure 59: A sample of painting images from Wikiart dataset created by Vincent Van Gogh.

Furthermore, we use painting images to transfer their style on images with other content. Some results are presented in Figure 61. in which we have extracted the style of the *"The café terrace"* painting of Vincent van Gogh (third image of Figure 59 on the right) and transfer it on images from Gendarmenmarkt square (Figure 60).

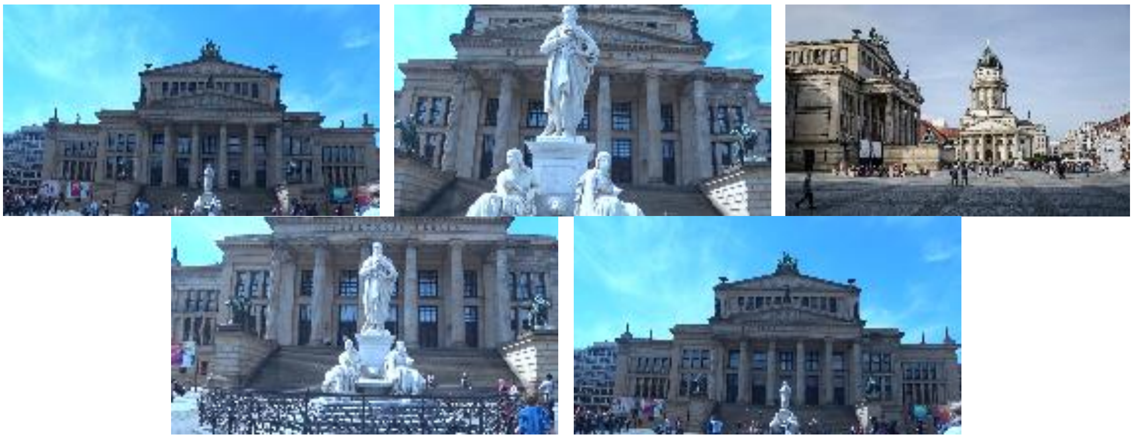


Figure 60: A sample of content images from Gendarmenmarkt square.

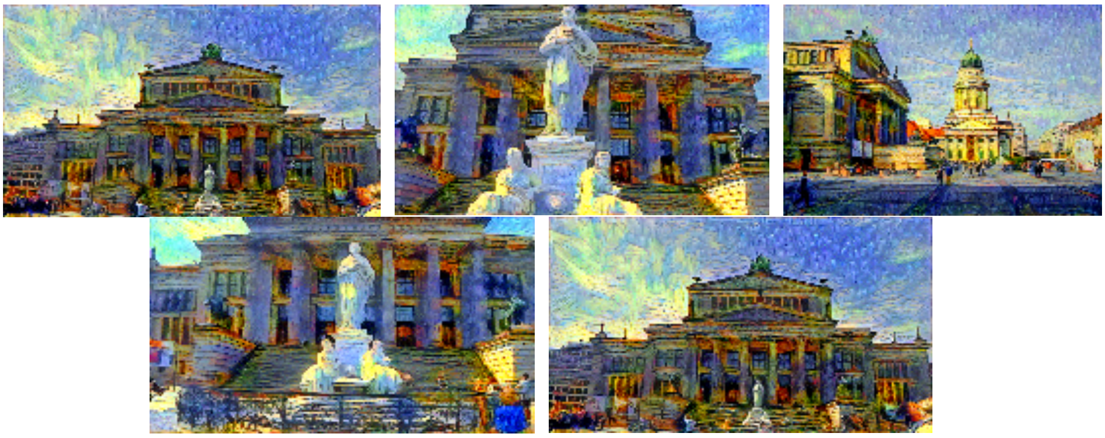


Figure 61: Results of TP from the Vincent Van Gogh painting ” *The café terrace*” to Gendarmenmarkt square images.

6 CONCLUSION

Not all 2D image or video data is equally suitable for 3D reconstruction, building localization or aesthetics extraction purposes. Some important constraints limit the applicability of the methods.

With respect to 3D reconstruction from 2D content, several requirements need to be met in part or in full in order to obtain an overall high-quality 3D model. The most important requirement is the presence of a **baseline** between camera positions since it is not possible to calculate depth information from images that are captured without changing camera position in between recordings. This requirement eliminates a large percentage of the available content. Another important limiting factor is the requirement of a **rigid scene** since moving objects like doors, people or cars interfere with the analysis of the 2D material in the 2D to 3D reconstruction.

The quality of the reconstruction depends on several factors, such as the presence of **distinct** features in the image content that can be uniquely matched between frames or images. It is found that content that lacks such details (e.g. homogeneous walls) causes 3D reconstruction to fail or produce undesirable results. Other important impacting factors are **image** and **spatial resolution** which directly influence the resolution of the 3D model. Films provided in resolution PAL or NTSC or lower, and without the requested camera movements for example have turned out to be suboptimal to the 2D to 3D transformation purposes. Finally, the **type** of scene or object that is recorded is important as well. It is determined that complex objects and reflective structures cause faulty depth information in the 2D to 3D transformation.

Based on analysis by KUL a full HD resolution or 4K provides much higher quality 3D results. Since much of the data provided by the content partners was captured some time ago – thus being filmed on film and transferred to PAL or NTSC video, or filmed on PAL or NTSC video – the material is mostly of low resolution. Some of the material was shot more recently and is thus provided in full HD. Today many productions are filmed in 4K. The exception being the DW series #DailyDrone since it exclusively consists of drone footage shot in HD or 4K.

The content from the different content providers was analysed and this process showed that most data from EF does not meet the necessary requirements for use in a 3D reconstruction pipeline. Image baseline, if present, is mostly too high for 3d reconstruction for this case. The general low quality footage and limited Multiview content rules out reconstruction using only this data. However, thanks to the available metadata, provided by EF, some imagery could be combined with other sources. Furthermore, EF provided several content collections that could be used for alternative 3D reconstruction methods.

The analysis of AF's data showed that most of the video footage did not contain any parallax (moving cameras). In those cases where AF's image collections contained multiple viewpoints from a single scene, these generally suffer from a low baseline. The data from SLRS is similar in that only a handful of scenes can lead to 3D reconstructions. The resolution of the videos containing potentially usable scenes also have low resolution.

The content from *Nico's Weg*, provided by DW is typically shot by cameras in close proximity to the actors, causing areas behind them to be blurry. Most scenes are shot in an over-the-shoulder fashion, which results in no overlap between camera shots.

Data from different external sources was also analysed, including many YouTube videos from landmarks and buildings, as well as online image repositories, such as Flickr. The empirical study shows that these content sources are very promising and can be used for 3D reconstruction purposes.

Regarding the STBOL module, we presented examples of input images with the corresponding content, aiming to be as close as possible to buildings or objects relevant to the V4Design use cases scenarios. The STBOL module is expected to detect and localize buildings and objects, so it is required to be fed with images that depict architectural objects. In this empirical study we presented images and videos that could be used for the development, training and fine-tuning of the computer vision algorithms. The total number of images and videos from the interior datasets is certainly sufficient but the content is not so satisfying in some cases. There are many images which depict maps or text referring to a building, but not the building itself. Moreover, there are many videos which depict buildings and objects, such as a video from Mediterranean region, but there are also some of them in which people and their culture are dominant and there are no buildings or objects. On the other hand, exterior datasets for objects include many useful images, but not many of them refer to traditional objects of different countries. Furthermore, exterior datasets for buildings are also limited and should be expanded in order to include more different kinds of buildings.

Finally, with respect to the AE&TP module, we have already fruitful external datasets including a wealth of famous painting images annotated by their school of art or their creator. On the other hand, regarding the interior datasets, they include many painting images that have no annotation. The interior datasets are useful in the case of an evaluation procedure where the system estimates style or creator in an automatic way and then an expert evaluate the results.

REFERENCES

Huang, Q., Wang, H., & Koltun, V. (2015*). Single-View Reconstruction via Joint Analysis of Image and Shape Collections.* ACM Transactions on Graphics (TOG), 34(4), 1–10. http://doi.org/10.1145/2766890

Lin T-Y., Maire M., Belongie S, Bourdev L., Girshick R., Hays J., Perona P., Ramanan D., Zitnic C. L. (2014). *Microsoft COCO: Common Objects in Context.* CoRR, 1405.0312, https://arxiv.org/abs/1405.0312Florea C., Condorovici R. G., Vertan C, Boia R., Florea

L., Vranceanu R. (2016*). Pandora : Description of a Painting Database for Art Movement Recognition with Baselines and Perspectives*. CoRR, 1602.08855, https://arxiv.org/abs/1602.08855

Khan F.S., Beigpour S., van de Weijer J., Felsberg M. (2014). “Painting-91: a large scale database for computational painting categorization”, *Machine Vision and Applications,* 25(6), 1385-1397, https://doi.org/10.1007/ss00138-014-0621-6