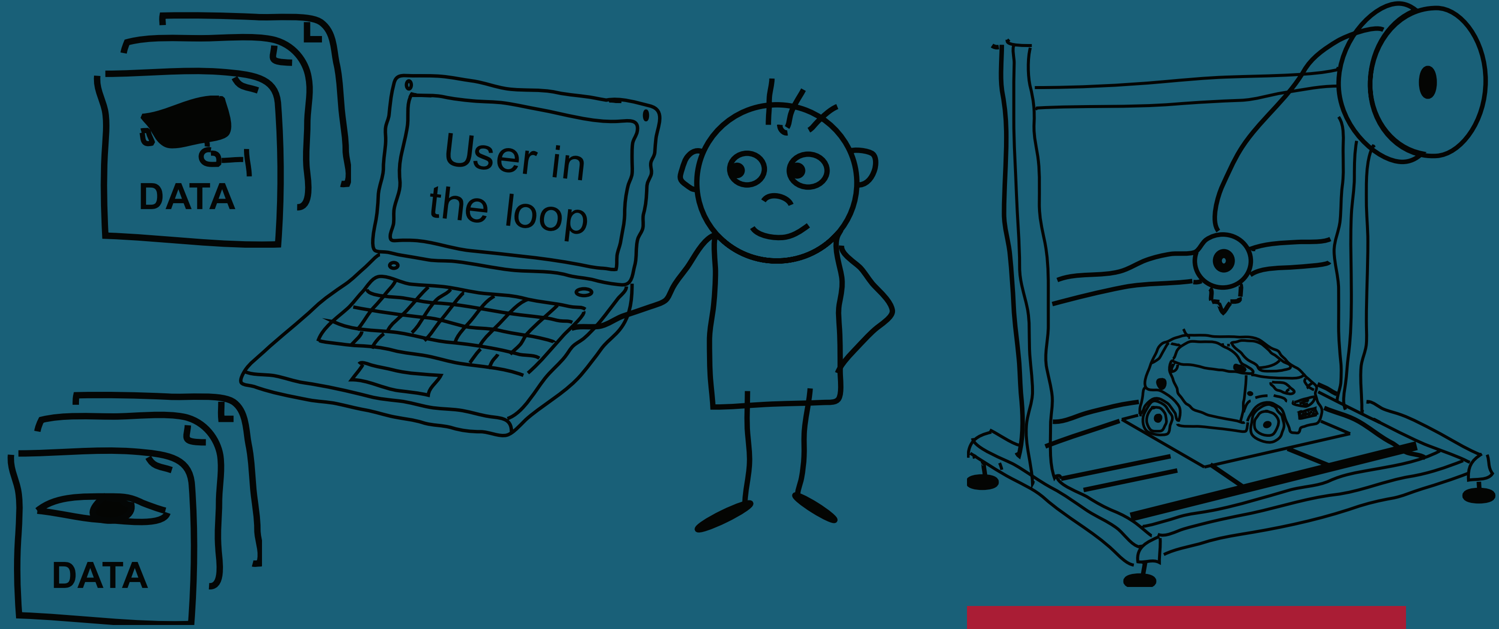
JULIUS SCHÖNING

INTERACTIVE 3D RECONSTRUCTION

2017

INTERACTIVE 3D RECONSTRUCTION

Julius Schöning



Interactive 3D Reconstruction

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
im Fachbereich Humanwissenschaften der
Universität Osnabrück

vorgelegt
von

Julius Schöning

Osnabrück, 2017

This is a cumulative doctoral thesis that includes articles published in peer-reviewed journals and conference proceedings. The articles are not included in this electronic version of the thesis due to copyright reasons. Instead, only the abstracts are embedded in Part P, but the full bibliographical reference and the corresponding DOI of each article are provided.

Supervisor and first examiner: Prof. Dr. Gunther Heidemann
Co-examiner: Prof. Dr. Daniel Weiskopf
Co-examiner: Prof. Dr. Kai-Uwe Kühnberger

Day of disputation: 26. April 2018

Abstract

Interactive 3D Reconstruction

Applicable image-based reconstruction of three-dimensional (3D) objects offers many interesting industrial as well as private use cases, such as augmented reality, reverse engineering, 3D printing and simulation tasks. Unfortunately, image-based 3D reconstruction is not yet applicable to these quite complex tasks, since the resulting 3D models are single, monolithic objects without any division into logical or functional subparts.

This thesis aims at making image-based 3D reconstruction feasible such that captures of standard cameras can be used for creating *functional* 3D models. The research presented in the following does not focus on the fine-tuning of algorithms to achieve minor improvements, but evaluates the entire processing pipeline of image-based 3D reconstruction and tries to contribute at four critical points, where significant improvement can be achieved by advanced human-computer interaction:

(i) As the starting point of any 3D reconstruction process, the object of interest (OOI) that should be reconstructed needs to be annotated. For this task, novel pixel-accurate OOI annotation as an interactive process is presented, and an appropriate software solution is released. *(ii)* To improve the interactive annotation process, traditional interface devices, like mouse and keyboard, are supplemented with human sensory data to achieve closer user interaction. *(iii)* In practice, a major obstacle is the so far missing standard for file formats for annotation, which leads to numerous proprietary solutions. Therefore, a uniform standard file format is implemented and used for prototyping the first gaze-improved computer vision algorithms. As a sideline of this research, analogies between the close interaction of humans and computer vision systems and 3D perception are identified and evaluated. *(iv)* Finally, to reduce the processing time of the underlying algorithms used for 3D reconstruction, the ability of artificial neural networks to reconstruct 3D models of unknown OOIs is investigated.

Summarizing, the gained improvements show that applicable image-based 3D reconstruction is within reach—but nowadays only feasible by supporting human-computer interaction. Two software solutions, one for visual video analytics and one for spare part reconstruction are implemented.

In the future, automated 3D reconstruction that produces functional 3D models can be reached only when algorithms become capable of acquiring semantic knowledge. Until then, the world knowledge provided to the 3D reconstruction pipeline by human computer interaction is indispensable.

Contents

Listing of figures

Abbreviations

2D	two-dimensional
3D	three-dimensional
ANN	artificial neural network
AOI	area of interest
AR	augmented reality
CAD	computer-aided design
CNN	convolutional neural network
conv	convolutional
CV	computer vision
EEG	electroencephalography
fc	fully connected
GPU	graphics processing unit
GUI	graphical user interface
HCI	human-computer interaction
HRC	human-robot collaboration
iSeg	interactive annotation and segmentation
LiDAR	light detection and ranging
maxp	max pooling
mp	megapixel
MVG	multiple view geometry
NUI	natural user interfaces
OOI	object of interest

PC	point cloud
RGB	red, green, and blue
RGB-D	red, green, blue and depth
RNN	recurrent neural network
SAGTA	semi-automatic ground truth annotation
SfM	structure from motion
SLAM	simultaneous localization and mapping
UI	user interface
VA	visual analytics
ViPER	video performance evaluation resource
VR	virtual reality
VVA	visual video analytics

0 Introduction

Estimating the three-dimensional (3D) structure of an object of interest (OOI) within a scene is usually a trivial task for humans. Image-based 3D reconstruction is such a trivial task for a human that even missing, i. e., occluded, elements of the OOI can be completed correctly [47, 82]. In contrast to human vision, and despite more than three decades of research in this field, computer vision (CV) still does not solve the problem of getting the 3D structure, including its geometrical constraints, from two-dimensional (2D) projections like images or video sequences, in a satisfying way. However, gaining information from a 3D scene with its 3D OOIs is the prerequisite for many modern applications, such as engineering in augmented reality (AR), as well as virtual reality (VR), automated video surveillance, autonomous car driving, and human-robot collaboration (HRC).

By taking images and video sequences, the real world with its 3D content is reduced to a 2D projection. Therefore, image-based reconstruction must estimate the lost dimension to get back the 3D nature of the scene. In theory, infinitely many different 3D geometries can create the same 2D projection. Thus, image-based 3D reconstruction is a mathematically ill-posed problem [92, 36], i. e., this inverse problem does not have a unique solution. This ill-posed nature of 3D reconstruction could be the reason why humans are so much better at solving this problem, even based on single images. In general, for solving ill-posed problems, additional *a priori* information and constraints are needed, also referred to as semantic knowledge. Nowadays, 3D reconstruction pipelines make use of low-level geometrical assumptions, e. g., the Manhattan-world-like characteristics of human-made objects [56, 20] for solving ambiguities. These ambiguities are mainly caused by the 2D projection and the pixel discretization within the image acquisition hardware. Nevertheless, to accomplish industrially-relevant and applicable 3D models, this thesis investigates how interactive processes in which the humans' semantic knowledge about the 3D geometry is combined with the computational power of today's computers, can improve the ill-posed problem of image-based 3D reconstruction.

0.1 Image-Based 3D Reconstruction – in a Nutshell

Software applications for image-based 3D reconstruction, such as *3DF Zephyr* [1], *Agisoft PhotoScan* [3], *ARC 3D* [95], *ContextCapture* [9], *ProFORMA* [63], and *VisualSFM* [99], use a feed-forward processing pipeline composed of various functional steps for estimating the 3D geometry of the scene. As illustrated in the uppermost row of Figure 0.2, five steps, in general, are necessary for getting from the very first step, i. e., the *image acquisition* hardware, to the final step, i. e., a 3D model which can be *printed*, *rendered*, and *used*, e. g., for computer-aided design (CAD) applications. Depending on the intended level of detail, the 3D reconstruction pipeline can be represented with more or less processing steps. As an introductory subject matter of 3D reconstruction, however, the five steps in Figure 0.2 give an appropriate level of detail.

Starting with *data acquisition* hardware, different sensors capture the whole scene, an area of interest (AOI) from the scene, or the OOIs within the AOI. Commonly, three different types of sensors can be distinguished: standard monocular cameras with red, green, and blue (RGB) color channels, depth sensors with a depth channel only, and active depth cameras with four channels, respectively a red, green, blue and depth (RGB-D) channel. In this first processing step of the pipeline, only the acquisition technology and its output values, i. e., resolution, measurement accuracy, noise characteristics, and data formats, are considered for this thesis.

Depending on the acquisition sensors, the next processing step *takes photos*, *takes video sequences*, *takes* RGB-D *images*, etc., will result in different collections of data. A collection of planned photos is, for example, taken by setting the OOI on a turntable and capturing pictures only from specific viewpoints, while the turntable is rotating. For illustration, consider the collection of planned images in the Middlebury multi-view stereo data set [79]. Other options for taking input images and recording video sequences are, among many others, image collections from the Internet (cf. Rome16k [57]), camera walks or journeys through streets [31], object capture under quite controlled conditions with moving cameras [53], computer-generated data sets rendered on virtual 3D models [15], and scene captures with only a few images and almost planar camera movement [83].

Based on the kind of input data, different methods have been developed for the *fea-*

ture detection and description step. Here, unique keypoints are extracted from the input and, by using features, are described in machine-understandable terms—most commonly by numerical feature vectors. For aligning images and frames of video sequences with other images or frames in which the same keypoints with the same features occur, *feature matching* is performed. To cope with the amount of input data and its associated complexity, a variety of approaches, such as incremental, hierarchical and global feature matching [77], have been developed. Once the decision has been made which images or frames show the same parts of a scene, the spatial location of the acquisition hardware at capturing time, with respect to the other viewpoints, is estimated by methods of *multiple view geometry* (MVG). Afterwards, the 3D position of each keypoint is triangulated, which leads to a 3D point cloud (PC) of keypoints representing the 3D scene, AOI, or OOI. Note that the last three processing steps, *feature detection and description*, *feature matching*, and *multiple view geometry*, of the 3D reconstruction pipeline might vary, depending on the data acquisition sensor.

Since the points of the PC are not connected in any way, *3D point cloud clearing and meshing* must be performed as the last step, before getting a 3D model. This *3D point cloud clearing and meshing* means that 3D points that were identified as outliers or that did not belong to the OOI, e. g., will be removed from the PC. This cleared PC will then be meshed, i. e., points next to each other will be meaningfully connected with lines which then form faces, so that a 3D model is created.

After considering the general pipeline of 3D reconstruction, the question arises, however, whether the created 3D models have currently any practical use. The answer to this question is usually *No*, because, due to the monolithic nature of the reconstructed 3D models, they are inappropriate for simulation tasks, reverse engineering, object replication, and other demanding tasks. In monolithic 3D models, the OOI is not represented on its own. Furthermore, scene and objects cannot be distinctively divided, because no semantic information regarding the scene, its objects, and the subparts, also known as subassemblies, of these objects are included in the 3D model. For the industrially-relevant use, 3D models are required which, in addition to the required accuracy, also contain a semantic division into logical subparts, including rigid or flexible connections between the subparts. Only such kinds of 3D models allow the 3D printouts, in addition to looking like the original, also to work like the original object.

In order to obtain these desirable 3D models of high industrially-relevance from images and video sequences captured by standard RGB-cameras, ways of incorporating human semantic knowledge about the 3D characteristics of the scene and its objects are researched in this thesis. The research objective is to improve the ill-posed problem of image-based 3D reconstruction, by the incorporation of human knowledge using as little human-computer interaction (HCI) during the 3D reconstruction process [48, 37] as possible. As starting point, therefore, the entire processing pipeline of image-based reconstruction is evaluated to identify shortcomings that can be significantly improved by

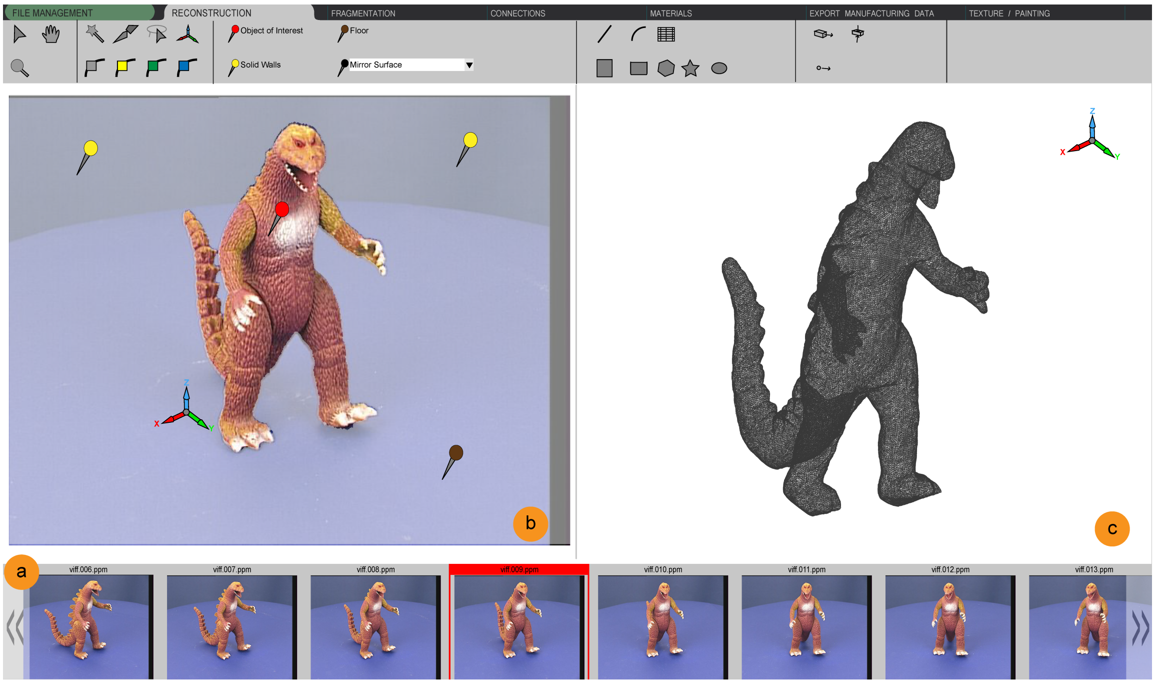


Figure 0.1: Snapshot of the designed graphical user interface for interactive 3D reconstruction [Paper IV, Figure 2 (a)] showing the first step of a reconstruction. The user selects the object of interest by marker pins on one image or frame (cf. ⓑ) which instantly leads to a first intermediate 3D model (cf. ⓒ).

means of HCI. For developing sustainable solutions of high practical use, the investigation of shortcomings in the 3D reconstruction pipeline is explicitly performed top-down. Top-down means here, starting from the top of every software—the graphical user interface (GUI) with its user interface (UI) metaphors—which then leads down to the different processing steps and their algorithms. Note that the term “image-based” includes collections of images and video sequences. Both serve as 2D projections and will be considered, unless explicitly stated otherwise, as the same, since frames of a video sequence can be seen as a collection of images and vice versa.

More precisely, the GUI, including its UIs, must equip the user with options for recognizing, as well as filling in, missing information into the 3D reconstruction pipeline to get industrially-relevant, i. e., CAD-ready, 3D models. Determining potential starting points that substantially improve current reconstruction software, different snapshots of a GUI for interactive image-based 3D reconstruction are designed.

As illustrated in Figure 0.1, to reconstruct one or multiple OOIs, the user identifies them using a marker pin on a single image or frame from the video sequence. Triggered by this user interface metaphor, the initialization of the 3D reconstruction process is started. Since the marker pin defined the semantic information for recognizing the objects which are needed for the reconstruction, the marker pin UI metaphor adds knowledge to the process. Considering this additional knowledge, pixel-accurate segmentation based on the 2D projections can be performed. Thereby, all pixels representing the OOIs will be distinguished as foreground, and all the remaining pixels will be considered as background. Due to pixel discretization, as well as the fact that the objects are projected to the 2D image plane, the fore- and background segmentation by itself is also ill-posed. In consequence, interactive pixel-accurate OOI annotation and the leveraging segmentation processes are researched and the insights gained are implemented as software.

The prototyped interactive annotation and segmentation software allows a rapid fore- and background segmentation, but due to the use of classical UI devices, keyboard and mouse, it is still tedious and time-consuming. By considering, e. g., motion, gesture, and gaze sensing input devices, more natural user interfaces (NUI) for this task could be provided. In order to realize NUIs for image segmentation, the fusion of CV algorithms with human sensory data is of high importance. For supporting research on the integration of sensory data in CV applications, a uniform standard file format is introduced. Based on this format, instantaneous visualization and sonification are possible, as well as the merge of multiple data sets into one. These merged data sets facilitate, e. g., the evaluation of sensory enhanced algorithms or the training of sensory-inspired artificial neural networks (ANNs).

By combining the high-level knowledge gained through HCI, it can be seen that pixel-accurate recognition of the OOI within the 3D reconstruction pipeline, and in a second step, the recognition of its subparts on the 2D projections, significantly improves the resulting 3D model. Thus, the objective of applicable image-based 3D reconstruction is moved closer to its completion. Since the effect of 3D recognition on images and video sequences seems particularly beneficial, its analogies with the human visual system are investigated for proving or disproving its importance.

Processing time results in “waiting time” and is therefore lethal for every HCI. Within the 3D reconstruction pipeline, the estimation of the 3D PC by MVG is the bottleneck in terms of process time. In order to significantly shorten this time to meet the acceptable “waiting times” for HCI [80], a feasibility analysis on the replacement of MVG by ANNs for the 3D reconstruction of unknown objects is conducted.

To demonstrate the primary finding of this thesis, i. e., that pixel-accurate segmentation of the OOI on the input data significantly improves the usability of the resulting 3D objects, practical software applications of industrially-relevance, which use the interactive 3D reconstruction pipeline, are developed. For emerging technology of 3D printing, the interactive replication of spare parts using only a smartphone is prototyped, demonstrating an engineering show case. As second show case within the domain of video surveillance, visual video analytics (VVA) improved by reconstructed 3D objects, as well as improved by users’ sensory data, here gaze data, is implemented and evaluated.

In summary, it can be stated that a true copy of an OOI, create by image-based 3D reconstruction, is not only feasible in Hollywood films and TV series like “Enemy of the State” and “CSI: Miami”. By the use of HCI, i. e., the collaboration of users with the computer to generate semantic information, this can already be realized today. With the increasing amount of semantic information about the OOI, like how many subparts it consists of, or what kind of connections exist between the subparts, image-based 3D reconstruction and related topics such as OOI annotation in video sequences, become industrial applicable. In the future, with advancing algorithms, the interactive 3D reconstruction, which obtained applicable 3D models, can be systematically automated. Thus, 3D reconstruction gets more industrial value.

0.2 Structure of this Thesis – in a Nutshell

The research question introduced above, i. e., how to integrate human high-level semantic knowledge into the 3D reconstruction pipeline for getting industrially-relevant 3D models, spawns the body of research in this cumulative doctoral thesis. Covering the broad spectrum of research on this topic, most of the 19 peer-reviewed articles in both journals and conference proceedings fall into the field of CV. A quite heterogeneous corpus of articles at first glance, however, under the single heading “interactive 3D reconstruction”, they reflect my research on improving image-based 3D reconstruction in the past three years in the biologically-inspired CV group at the Osnabrück Institute of Cognitive Science.

Highlighting those aspects, my research contributes in particular, all 19 articles are linked to the corresponding processing steps of the 3D reconstruction pipeline. These links lead to the overview matrix in Figure 0.2 that illustrates the general structure of my research. While the horizontal axis represents the introduced processing steps of the 3D reconstruction pipeline, the vertical axis is defined by the six main focus areas to which my articles contribute. Each of these foci corresponds to an individual chapter in the remainder of this thesis. More precisely, each chapter will introduce the reader to the particular research question, which is studied by the referred articles. These articles, which contribute my research, are cited, in particular, as Paper I to Paper XIX and for further reading they are reproduced in Part P of this thesis. Due to the interdisciplinarity of bio-inspired CV, the articles are sometimes written in styles of other disciplines and not necessarily in the presentation style of computer science. Since two articles had been accepted for publication, but are not published at the time of this writing, they have been marked accordingly and have also been included in Part P.

The remainder of this thesis consists of six chapters, each characterized by a single focus area in the vertical dimension of Figure 0.2, and one addition concluding chapter. As the first focus area in the vertical dimension, Chapter 1 discusses the existing and

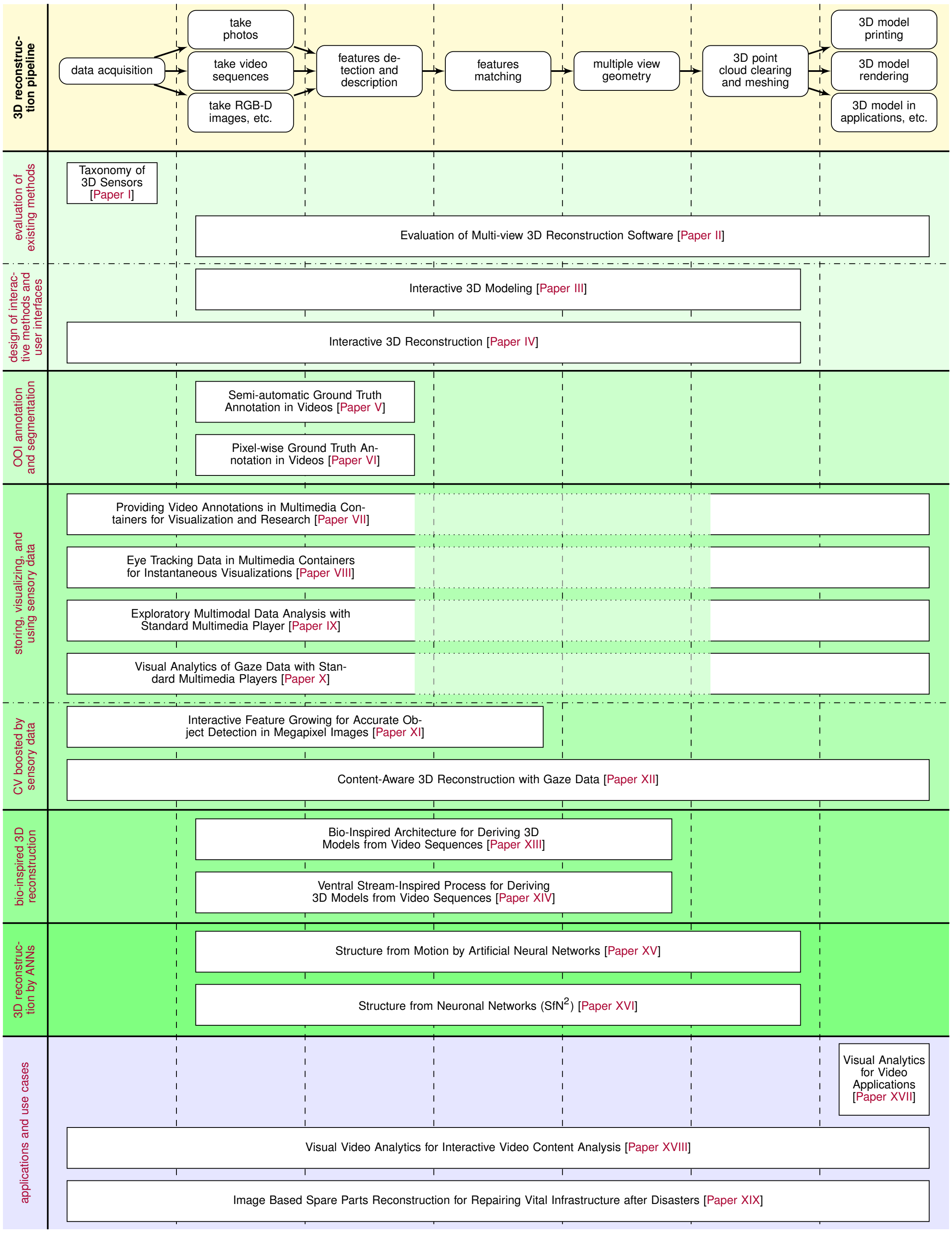

3D reconstruction pipeline
data acquisition
take photos
take video sequences
take RGB-D images, etc.
features detection and description
features matching
multiple view geometry
3D point cloud clearing and meshing
3D model printing
3D model rendering
3D model in applications, etc.
evaluation of existing methods
Taxonomy of 3D Sensors [Paper I]
Evaluation of Multi-view 3D Reconstruction Software [Paper II]
design of interactive methods and user interfaces
Interactive 3D Modeling [Paper III]
Interactive 3D Reconstruction [Paper IV]
OOI annotation and segmentation
Semi-automatic Ground Truth Annotation in Videos [Paper V]
Pixel-wise Ground Truth Annotation in Videos [Paper VI]
storing, visualizing, and using sensory data
Providing Video Annotations in Multimedia Containers for Visualization and Research [Paper VII]
Eye Tracking Data in Multimedia Containers for Instantaneous Visualizations [Paper VIII]
Exploratory Multimodal Data Analysis with Standard Multimedia Player [Paper IX]
Visual Analytics of Gaze Data with Standard Multimedia Players [Paper X]
CV boosted by sensory data
Interactive Feature Growing for Accurate Object Detection in Megapixel Images [Paper XI]
Content-Aware 3D Reconstruction with Gaze Data [Paper XII]
bio-inspired 3D reconstruction
Bio-Inspired Architecture for Deriving 3D Models from Video Sequences [Paper XIII]
Ventral Stream-Inspired Process for Deriving 3D Models from Video Sequences [Paper XIV]
3D reconstruction by ANNs
Structure from Motion by Artificial Neural Networks [Paper XV]
Structure from Neuronal Networks (SfN^2) [Paper XVI]
applications and use cases
Visual Analytics for Video Applications [Paper XVII]
Visual Video Analytics for Interactive Video Content Analysis [Paper XVIII]
Image Based Spare Parts Reconstruction for Repairing Vital Infrastructure after Disasters [Paper XIX]

Figure 0.2: Overview matrix illustrating the aspects in which the articles of this thesis improve the current knowledge of the 3D reconstruction pipeline (uppermost row), in the horizontal dimension. In the vertical dimension, the articles are clustered into six focus areas, each described in more detail in a single chapter of this thesis.

interactive methods of 3D reconstruction. Furthermore, this chapter describes the user-centric top-down approach based on GUIs designed for usable 3D reconstruction by HCI which identifies starting points for the research. With regards to the question of how OOI annotation and segmentation can be performed, Chapter 2 introduces the interactive, polygon-based and semi-automatic method iSeg. This method combines the strengths of users and computers, i. e., the computer actively asks for “help” if it recognizes that it might do incorrect annotations. Having polygon-based shape annotations of, e. g., video sequences, leads directly to the question of how to store these annotations in a meaningful and generally accessible format. Chapter 3 discusses this issue of storing, visualizing, and using sensory data, with particular focus on the holistic consideration of how to store annotations and also other metadata that might be generated during close HCI, like gaze trajectories or other sensory data. To achieve the full potential of HCI, the idea suggests itself that CV boosted by sensory data drastically simplifies the interaction in object detection and 3D reconstruction tasks. These two subtopics are covered by Chapter 3 on CV and gaze data. Aiming at bio-inspired 3D reconstruction, Chapter 4 discusses the analogies between the psychophysical evidence that “humans encode 3D objects as multiple viewpoint-specific representations that are largely 2D” [11] and image-based 3D methods. Based on the found analogies, already made and needed improvements, for the 3D reconstruction pipelines are implemented and evaluated. With the focus on reducing the computational complexity to ensure a smooth HCI, Chapter 5 investigates whether the computationally intensive parts of MVG methods could be replaced with 3D reconstruction by ANNs. This analysis aims to develop ANNs which achieve the same results as MVG-based 3D reconstruction methods and which are, like MVG methods, also able to perform 3D reconstruction without *a priori* object knowledge. Representing the final focus area, Chapter 6 highlights the practical applications and use cases that have been improved by or developed in the course of the research reported in this thesis. The spectrum of improved or developed applications ranges from VVA via pixel-accurate video annotations to image-based spare parts reconstruction. Finally, the conclusion summarizes the main findings gained by this research and discusses open issues and limitations to define the directions of future work for accomplishing industrially-relevant and practical 3D models by interactive 3D reconstruction.

You don't have to reinvent the wheel
just attach it to a new wagon.

Mark McCormack

1

Existing and Interactive Methods

In order not to reinvent existing and valuable methods and approaches, a careful exploration of the current body of research is necessary. That is why this chapter summarizes the literature survey performed during the research, and, furthermore, includes additional recent findings. This analysis of the current state-of-the-art then serves as the line of argument for the design of interactive architectures which can integrate humans' semantic knowledge into the 3D reconstruction process. For the design of such adequate interactive architectures as the framework for practical software solutions, two requirements are relevant: *(i)* the process must return applicable 3D models of industrial relevance and *(ii)* the UI must be reduced to a minimum, but in case UI is necessary, it must be intuitive.

1.1 Evaluation of Existing Methods

Data acquisition, or, more specifically image acquisition, as the first processing step is the starting point for any 3D reconstruction pipeline (cf. uppermost row in Figure 0.2). The spectrum of sensors for image acquisition ranges from monocular cameras over stereo cameras, also known as dual cameras, to active depth cameras. These active depth sensors usually return an RGB-D image or video sequences, where, in addition to the color channels, a depth channel provides values for the distance between the camera center and the captured scene. In particular, after the first consumer RGB-D camera, the *Kinect* camera, was released in 2010 by *Microsoft*, the number of scientific publications

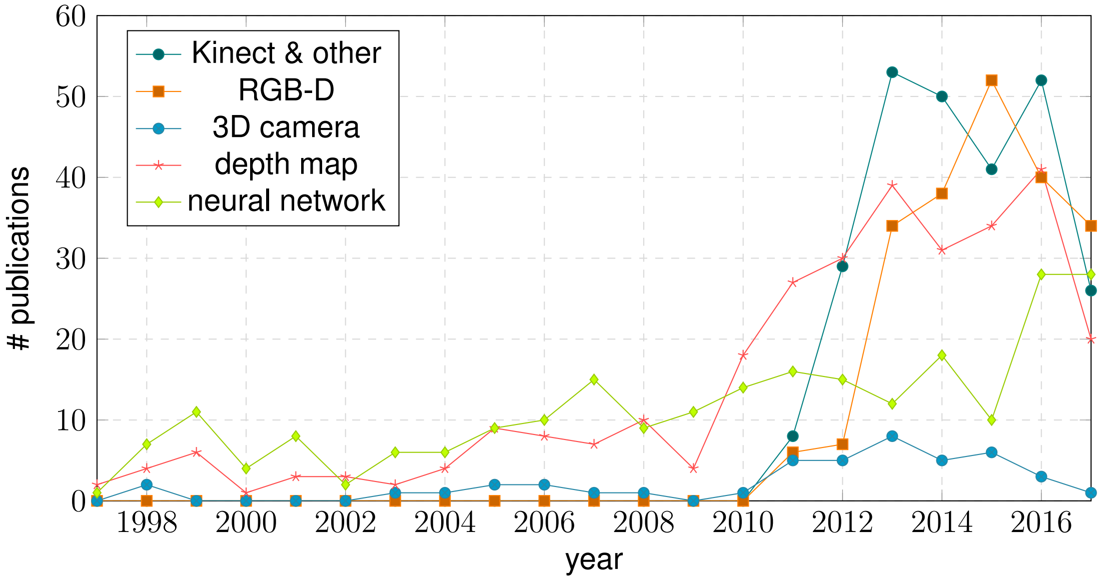


Figure 1.1: Number of publications with the keywords 3D reconstruction or 3D modeling and Kinect & other, RGB-D, 3D camera, depth map or neural network; calculated from Elsevier's Scopus webservice [26].

using these devices has increased rapidly, as shown in Figure 1.1.

RGB-D cameras, are used in different areas, for instance in scene reconstruction [35, 17], building mapping [38, 42], forensics [24, 61], robotics [25, 106], and various other applications [7, 30]. One might ask, however, whether RGB-D cameras of depth sensors should indeed be the preferred choice when it comes to the realistic reconstruction of scenes or OOIs with the goal of returning non-monolithic 3D models, as they are necessary in industrial uses? Therefore six, RGB-D cameras and also three depth sensors below 5.000€ were benchmarked. Contrary to initial expectations, do-it-yourself academic RGB-D cameras like [111] end up producing very high overall cost, clearly exceeding the previously introduced limit of 5.000€. However, they seem to be a good choice for highly specialized, but not for general use cases.

With a particular focus on the usability and measurement accuracy for 3D reconstruction purposes, a first benchmark on depth data acquisition sensors [Paper I] confirms Henry et al.'s [38] claim that RGB-D cameras only provide a depth sensing range of up to a limited scanning distance of approximately five meters. Secondly, if the working principle of depth measurement relies on a structured light pattern, very smooth or transparent objects will lead to blind spots in the depth values. These blind spots are due to the fact that the light pattern is not reflected as expected by the measurement hardware. Thirdly, the outdoor use of all selected depth cameras, as well as sensors, remains very limited, because of their vulnerability to weather conditions, such as sunlight, humidity, rain, and snow, which further limits the maximal depth perception distance—in extreme cases, to zero meters.

In general, both depth and RGB information can be considered for improving the 3D reconstruction pipelines. However, the motivation for research on 3D reconstruction pipelines using "standard", monocular, RGB cameras for the data acquisition is strengthened by the following three shortcomings of RGB-D sensors: *(i)* the limited range at which these sensors can observe depth, *(ii)* even within depth sensing range certain material characteristics lead to blind spots of the depth map, and *(iii)* the fact that these sensors, even nowadays, are not available in every-day products like smartphones and laptops. Note that in line with this survey [Paper I] and for focusing the line of research, all other special hardware devices like dual cameras [51, 55] and light detection and ranging (LiDAR) sensors [39, 21] have deliberately been omitted. Thus, only monocular RGB photo- as well as video-cameras which return colored 2D projections of 3D scenes and OOIs will be considered as data acquisition hardware for the remainder of this research. Going beyond these technical considerations, it should also be pointed out that humans are capable of perceiving 3D object shapes even when using only one eye (cf. Chapter 4). Thus, a monocular imaging sensor for data acquisition in combination with certain semantic knowledge, like a set of underlying geometrical 3D objects [10, 43, 109], must be sufficient for almost any 3D reconstruction task. An additional reason for the commitment to monocular cameras is the vast amount of available data sets as well as recordings, e. g., from Internet image collections, and the almost exclusive use of monocular cameras in current electronic devices.

Boosted by the emerging technologies of 3D printing, VR and AR in the last decade, several reconstruction software solutions for image-based 3D reconstruction have been released. As initial data, all of these applications use 2D projections in the form of collections of images, video sequences or frames of video sequences showing either the whole scene or showing the AOI or the OOI from slightly different viewpoints. Due to the fact that for image-based 3D reconstruction the different viewpoints are typically created by camera as well as object motion, this method is also known as structure from motion (SfM) or, in older references, as stereophotogrammetry. Assessing the quality of existing methods, the four most widely used image-based 3D reconstruction software applications are evaluated [Paper II]. For this evaluation, *Agisoft PhotoScan Standard Edition* [3] and *Autodesk 123D Catch* [5], as non-academic software solutions, as well as *VisualSFM* [99, 100, 98] and *ARC 3D* [95], as academic software, have been chosen. Concerning practical applicability in real-world and set-up capturing scenarios, the benchmarks are completed on a real-world [83] and a planned [96] multi-view image data set. Note that all benchmarked software tools perform the image-based 3D reconstruction based on the SfM principle, which is comparable to the processing steps as shown by the uppermost row in Figure 0.2.

In order to provide qualitative and also quantitative results, a new method for comparing the reconstructed model with the ground truth is defined and applied in Paper II. This method returns the same number of measurement points for each model, which is needed to get comparable histograms, statistics, and meaningful figures. So far, conventional practice [86, 90] has been to make the comparison between the ground truth as the reference and each reconstructed model. The newly introduced method does the comparison the other way round. This way, the amount of measurement points depends only on the ground truth PC. Recently, this new method has been used for computing, e. g., empirical cumulative distribution functions and meaningful heart map visualizations for benchmarking new approaches [62, 89]. The performed benchmark [Paper II] shows that it is indeed possible to rank available software solutions for multi-view 3D reconstruction with respect to reconstruction quality and processing time also known as runtime. However, due to the large number of application cases, such as the 3D reconstruction of entire cities or small heritage artifacts, it is not feasible with our benchmark to provide a general ranking and announce the best software, because each software has a slightly different use case. The more general result of this benchmark is that almost all evaluated software obtains monolithic 3D models sufficient in accuracy if a lot of input images without occlusion are available and if the scenes or the OOIs do not exhibit too many shape irregularities.

1.2 Design of Interactive Methods and User Interfaces

The state-of-the-art surveys and benchmarks [78, 62, Paper II] show that the complete 3D reconstruction pipeline works in general, but only for specific tasks. However, would an architect or an engineer call the resulting 3D model a CAD-like one? Is an inexperienced user able to apply and use these tools? Can users create applicable 3D models of real-world OOIs, that are manufacturable with 3D printers in order to get an authentic and functioning replica? The answer to all these questions is a *No*—but why? By conducting a broad-based comparison of automatic as well as interactive image-based 3D reconstruction techniques [Paper III], the reason for this answer should be identified. In doing so, all methods are analyzed with a focus on the practical usability, as well as on the characteristics and the amount of input data.

In short, all of the eleven examined techniques [Paper III] require specific, planned input images or video sequences which show the OOI ideally from all perspectives. Furthermore, the handling of delicate structures, textureless surfaces, hidden boundaries, illumination, specularity, or dynamic or moving objects, as they occur in normal captures, are not taken into account by any of these reconstruction techniques. In addition, if only few images are available, such methods are not able to construct 3D model at all [49]. Another major drawback is that, like the already benchmarked software solutions [Paper II], all techniques considered here also return monolithic models as PC or

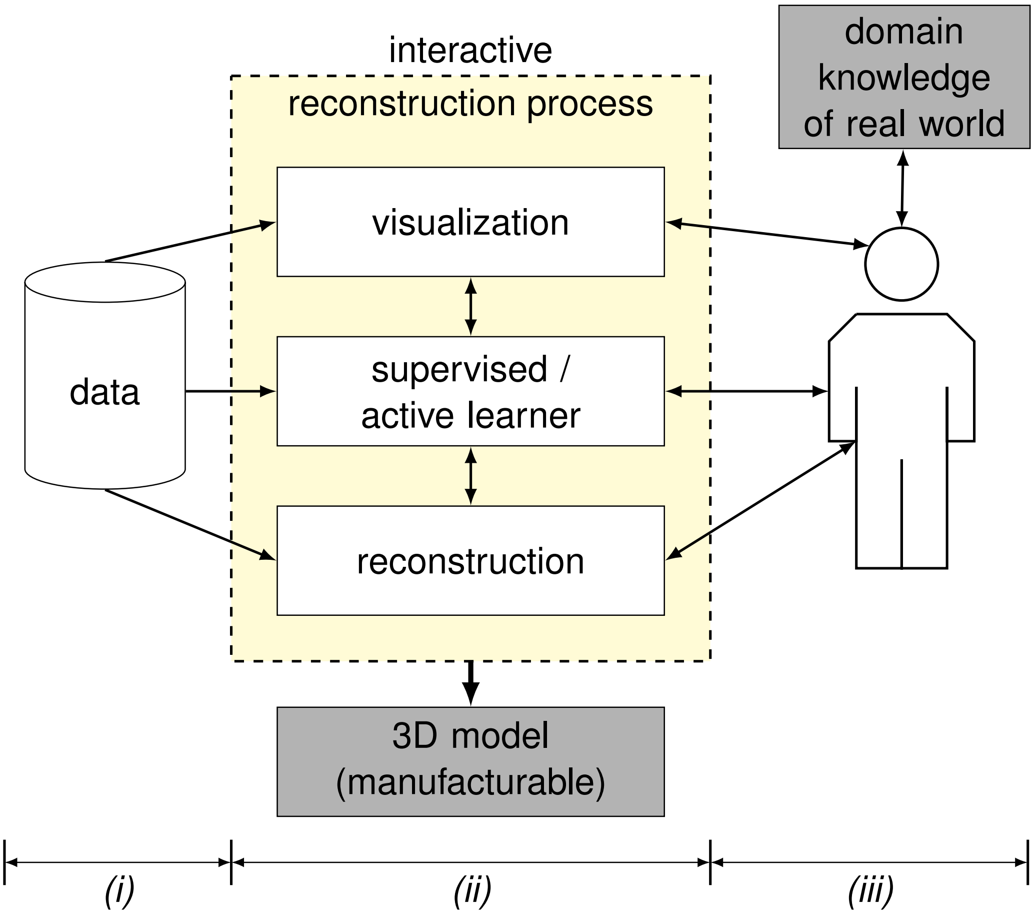


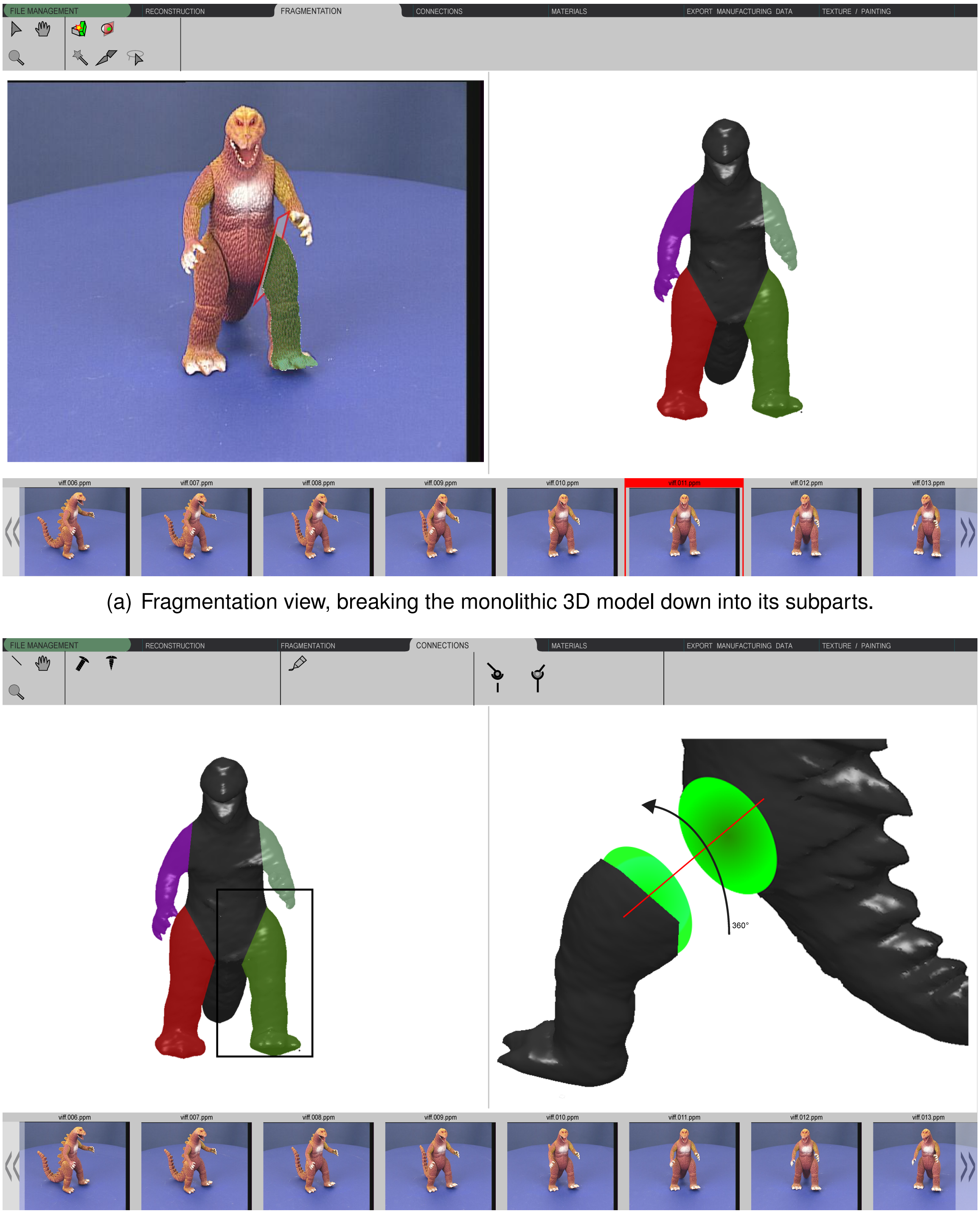
Figure 1.2: Interactive reconstruction architecture [Paper III, Paper IV] comprised of three parts: (i) the input data in the form of image acquisition as well as additional data, (ii) the interactive reconstruction process itself and (iii) the user as a domain expert of the real world.

as meshes without any declaration of subassemblies. In consequence, these kinds of monolithic models can be used for visualization, but cannot be used in CAD applications like simulation tasks and AR manuals.

In order to bridge this obvious gap of, for instance, missing subpart definitions or missing geometrical shape information caused by occlusion, additional information must be given to the 3D reconstruction process. This lack of semantic knowledge and also the lack of shape information by the ill-posed nature of image-based reconstruction can currently only be solved by interactive 3D reconstruction architectures, processes and frameworks. In this context, interactive also includes semi-automatic processes. As a consequence, HCI is the only solution for creating manufacturable 3D models of high practical use. HCI incorporates high-level knowledge provided by users into the reconstruction pipeline [48, 37]. In the near future, this semantic knowledge cannot be integrated into existing algorithms explicitly, because of the sheer complexity this would create. Nonetheless, in order to reliably reconstruct 3D models from any collection of images and video sequences in the foreseeable future, Figure 1.2 outlines the first draft of an interactive 3D reconstruction architecture. As illustrated, three main parts characterize this architecture: *(i)* the input data captured by monocular cameras, as well as additional data, such as physical interrelationships, *(ii)* the interactive reconstruction process itself, and *(iii)* the user as a domain expert of the real world with all her / his high-level knowledge. In consideration of the data analysis methodologies for the pattern and structure discovery in VVA tasks [Paper XVIII, Paper XVII, 40], where the input data will be enhanced from a pixel-based to a high-level symbolic representation by HCI, this interactive 3D reconstruction architecture uses the same conceptual clustering into three parts. Thus, this interactive 3D reconstruction, like the VVA architecture, joins the computational power of today's computers with the semantic knowledge of users to solve issues that are computationally infeasible at the moment and to a certain extent, ill-posed. In this way, the computer remains the "workhorse" of the reconstruction process, while machine learning algorithms, including ANN, shift the workload from the user to the computer. As a consequence, 3D reconstruction from arbitrary and unplanned collections of 2D projects becomes possible, which will broaden the range of applications and use cases (cf. Chapter 6). For example, everyone will be enabled, e. g., to reverse-engineer needed mechanical parts including smaller subparts [Paper XIX], to summarize videos sequences with focusing on specific AOIs or OOIs [Paper XVIII], to annotate AOIs and OOIs with their eyes [Paper XI, Paper VI], or to reconstruct non-rigid, i. e. deformable, objects such as cuddly toys.

To further strengthen the interactive architecture sketched in Figure 1.2 and for developing sustainable solutions of high practical value, possible GUIs with the corresponding interface metaphors are prototyped [Paper IV]. These so-called GUI mockups give an indication which UI metaphors are essential to obtain the desired 3D models of high industrial relevance from images. More precisely, this top-down consideration of interactive 3D reconstruction identifies with which software functionalities the user must be equipped for incorporating their *a priori* high-level knowledge about the 3D shape of the OOI, the number of subparts, the rigid or flexible connections between the subparts, etc. for getting a CAD-like model.

The three most important snapshots of the prototyped GUI are illustrated in Figures 0.1 and 1.3. They show the three major HCIs components needed for interactive image-based reconstruction. In Figure 0.1, which is discussed in detail in the introduction, HCI is used for identifying and segmenting the OOI on the 2D projections which are needed to generate an intermediate, monolithic 3D model. For breaking this monolithic model down into its subparts, the HCI of Figure 1.3(a) enables the user to define edges on the images or frames, based on which the 3D model is then divided into parts. The HCI of Figure 1.3(b) can be used for defining rigid, flexible, detachable or permanent



(a) Fragmentation view, breaking the monolithic 3D model down into its subparts.

(b) Connection view, adding detachable or permanent connections and their degree of freedom.

Figure 1.3: The designed GUI of interactive 3D reconstruction [Paper IV, Figure 2]. Here showing different snapshot of different operating stages and their interface metaphors.

connections between the subparts. All these snapshots are chronologically sequenced, and interactive OOI annotation and segmentation must thus be investigated first.

A lot of the future of search is going to be about pictures instead of keywords.
Computer vision technology is going to be a big deal.
Ben Silbermann

2 Object of Interest Annotation and Segmentation

In visual attention tasks, humans usually outperform computers and the same is also true for object recognition and tracking tasks (cf. Chapter 3). On these kinds of tasks, users can easily follow an AOI or OOI over several frames even if the area or the object is occluded partially or completely. As already mentioned in the finial section of previous chapter, the identification and segmentation of one or multiple OOIs, which includes detection, masking, and cropping, is a major step within the interactive 3D reconstruction pipeline. The importance of OOI identification, also referred to as annotation, is highlighted by the fact that about 300 hours of video sequences are uploaded to *YouTube* [107] every single minute, not even considering other video platforms such as *Vimeo* and *Flickr*. For real time annotation, 18,000 operators would be necessary just for *YouTube* videos. Unfortunately, annotating video sequences and collections of images in such way that they can be used as ground truth in, e. g., scientific tasks, cannot be done in real time or faster, because pixel-accurate annotation is an ill-posed task whose complexity scales up with the resolution of the input data.

Even though a large diversity of approaches for AOI and OOI tagging, annotating as well as tracking over frames has been released in the last decades, useful software tools are still rare. A possible reason might be the high complexity of the task, in both the spatial and the temporal domain. The complexity of this task is so high that even the pixel-accurate annotations created by humans differ on the same data set [110, 29].

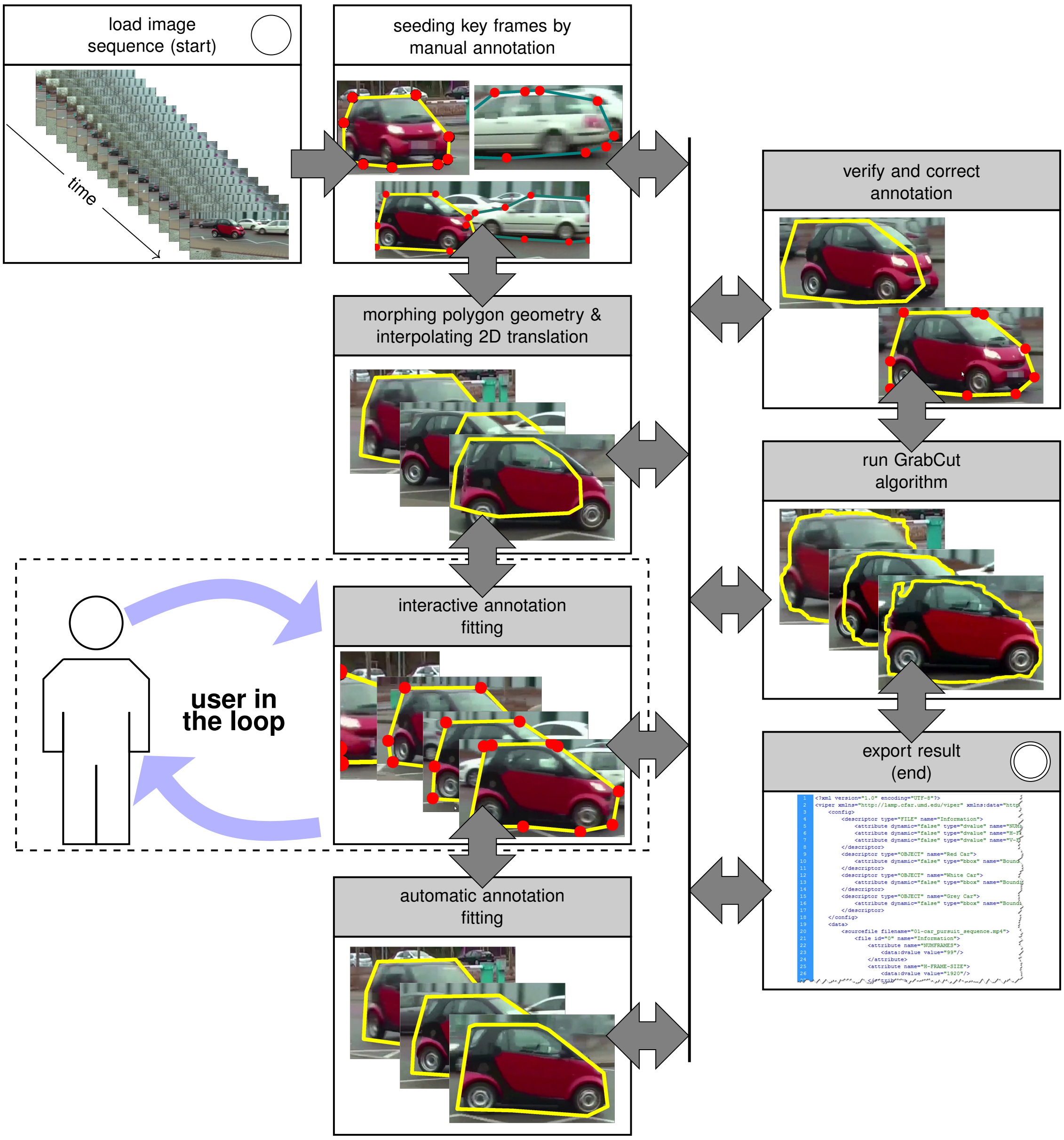
A further difficulty is that an enormous amount of pixel-accurate annotated video sequences is necessary for training, testing and validation, in particular, when ANNs are used.

The developed interactive annotation and segmentation method (iSeg) [Paper V, Paper VI] provides these pixel-accurate annotated video sequences with only little user interaction. The iSeg tool with its novel interactive approaches for polygon-based, i. e., pixel-accurate, OOI annotation as well as segmentation is introduced in the following. The chapter proceeds to highlight how the iSeg software has been used successfully in current independent research [108, 6], elaborates on the link to bio-inspired 3D reconstruction, and refers to further applications based on iSegs' framework (cf. Chapter 4 and 6 for more detailed discussion).

The benchmark of several image and video annotation tools based on different criteria by Dasiopoulou et al. [19], such as annotation granularity, location, and expressivity, triggered the research of the iSeg process. This benchmark finds that only one software provides the option of pixel-accurate AOI and OOI annotation for video sequences. All other evaluated tools provide only simple geometric primitives for annotations, such as rectangles and ellipses. As has become apparent by the evaluation of video annotation tools [Paper V] like the video performance evaluation resource (ViPER) [23] and semi-automatic ground truth annotation (SAGTA) tool [101] another drawback of these software solutions is that they do not provide useable interface metaphors for an easy concurrent annotation of several frames. Furthermore, important semantic information like inter-object relations, e. g., "the red car is occluded by the white one", cannot be added to the resulting annotations. Based on this evaluation, however and despite these drawbacks, positive stands out the annotation export format of ViPER. This XML based export format is appropriately defined and specified by an XSD schema. This standardized export format might will be the reason why ViPER is still broadly used despite its many shortcomings.

In contrast to these solutions, the iSeg software provides intuitive, interactive and pixel-accurate annotation features as modular processing blocks, as schematized in Figure 2.1. By design, the iSeg software combines both the computational power of computers and the high-level semantic abilities of human knowledge. The architectural framework of the interactive reconstruction [Paper III, Paper IV], which is based on the architecture of VVA [Paper XVIII, Paper XVII, 40], is very well-tailored to this ill-posed task. As a consequence, the iSeg application significantly increases the quality of the AOI and OOI annotations from the fact that the polygon-shaped annotations and inter-object relations can be encoded, which contain much more information than the current annotations with geometric primitives.

The key features of the released iSeg software are first, its workflow that is composed of eight modular process blocks, as shown in Figure 2.1, and, second, the underlying architecture. Only the two initial processing blocks, marked with a white headline, are



```
<?xml version="1.0" encoding="UTF-8"?>
<viper xmlns="http://lamp.cfar.umd.edu/viper" xmlns:data="http
    <config>
        <descriptor type="FILE" name="Information">
            <attribute dynamic="false" type="dvalue" name="NUM
            <attribute dynamic="false" type="dvalue" name="H-F
            <attribute dynamic="false" type="dvalue" name="V-F
        </descriptor>
        <descriptor type="OBJECT" name="Red Car">
            <attribute dynamic="false" type="bbox" name="Bound
        </descriptor>
        <descriptor type="OBJECT" name="White Car">
            <attribute dynamic="false" type="bbox" name="Boundi
        </descriptor>
        <descriptor type="OBJECT" name="Grey Car">
            <attribute dynamic="false" type="bbox" name="Bound
        </descriptor>
    </config>
    <data>
        <sourcefile filename="01-car_pursuit_sequence.mp4">
            <file id="0" name="Information">
                <attribute name="NUMFRAMES">
                    <data:dvalue value="99"/>
                </attribute>
                <attribute name="H-FRAME-SIZE">
                    <data:dvalue value="1920"/>
```

Figure 2.1: Overview of the interactive annotation and segmentation (iSeg) process [Paper VI]. Process blocks with white headlines are obligatory when running the iSeg tool. Blocks with gray headlines are optional and can be used in any order at any time. The block interactive annotation fitting actively asks for user interaction in case the automatically generated annotations appear to be incorrect. Thus, the computer remains the "workhorse" of the process, while the close cooperation with the user allows the iSeg software to improve results in both annotation accuracy and time.

obligatory and must be processed first and in the specified order. The remaining blocks can be arranged, executed and concatenated in an arbitrary order interactively by the user until the desired annotations are reached. Due to the immediate visualization of annotations on top of the video sequences, iSeg's process enables the user to verify and correct the annotations directly. For that purpose, UI metaphors for the annotation process, by the use of mouse and keyboard as input devices, are defined. Naming exemplary metaphors: By clicking and dragging the inside of an annotation shape, its entire polygon contour can be relocated. By double-clicking, a new vertex is added, which adjusts the shape of the annotation. By single-clicking and dragging a vertex, its position can be manipulated, and by pressing the SHIFT-key and clicking on a vertex, it can be deleted. Based on the finding that interactive CV applications are greatly improved by recording and incorporating users' gaze trajectories during the performance of tasks [Paper XI, Paper XII], new gaze-based interface metaphors are currently being prototyped. These gaze-based metaphors as NUI allow hands-free annotation during playback of video sequences, just by gaze at the OOI (cf. Chapter 7).

Facilitating the annotation of areas and objects, in an interactive manner, where the user is put into the loop, is should by avoided that users are annoyed by long “waiting times“ [80]. These ”waiting times“ are normally caused by the processing time of computationally expensive algorithms, like the extraction and description of keypoints on all frames of a given video sequence. Unfortunately, the sheer number of keypoints and the complexity of extracting these keypoints significantly increases with the resolution of the input frames. Given the high resolution of current devices, reaching 1080p (1920×1080) or 4k (4096×2160), keypoint extraction requires high-performance computer hardware to ensure a smooth HCI without ”waiting times“. Allowing smooth HCI even on laptops, iSegs' interactive annotation fitting block, in contrast, only extracts keypoints in or close to the annotation for adjusting it to the ”real“ AOI or OOI location. This way, it can handle even high-resolution video sequences. In case the computer cannot find a sufficient amount of keypoints for fitting the annotation, it asks the user for help. The detailed implementation of this functionality is based on the activity diagram [Paper V, Figure 2] which defines the algorithm's behavior when activity asking for user interaction. It should be mentioned that this implementation can cope with moving OOIs, moving cameras as well as moving OOIs *and* moving cameras, i. e., if both motions happen simultaneously. By increasing the smoothness of the interactive annotation fitting, multithreading is implemented since iSeg version 0.0.5.

The concept of the semantic timeline [Paper VI] further improves the annotation quality. This means that, in addition to a pixel-accurate annotation of AOIs and OOIs, the relations between them can also be specified. For encoding these inter-object relations, every annotation has its separate semantic timeline, where manually annotated frames are distinguished from automatically or interactively created annotations. The order of the semantic timelines can be changed by dragging and dropping either single or multiple frames to represent the z-order of the annotation. Here, z-order relates to the inter-object relation and means how close, for instance, a particular object is to the camera center. By encoding the z-order with the semantic timeline in this way, additional information is added to the annotation, e. g., when an object is completely or partially occluded by another object.

The export of the annotations created with the iSeg tool can be done in XML format, which is compatible with and valid for ViPER's XSD scheme [23]. Furthermore, the latest version of iSeg (1.0.0) provides the option to directly export annotations and inter-object relations as subtitle formats, which then can be integrated into multimedia containers for instantaneous visualization (cf. Section 3.1). As it is standard for software, the annotations and their inter-object relations can be stored in iSeg's binary project files, allowing an easy save-and-load of an entire project.

Summing up, the lack of video annotation tools allowing pixel-accurate annotations which can handle video sequences in standard resolutions inspired the development of the iSeg software. Based on the fact that such tools are needed, e. g., for ground truth generation, the iSeg prototype software is compiled for the most popular operating systems—Windows, Ubuntu, and MacOS. Beyond that, the source code is GPLv3 licensed so that anyone can assemble or improve iSegs' implementation on any platform. For instance, recently, Balloch and Chernova [6] use the iSeg software for the dense annotation of *Kinect v2* sensor data for scene segmentation by convolutional neural networks (CNNs). In the same domain, Zeng et al. [108] annotate the agent as well as the risky region in video sequences with the iSeg software for agent-centric risk assessment by recurrent neural networks (RNNs).

A picture is worth a thousand words.
An interface is worth a thousand pictures.
Ben Shneiderman

3

Computer Vision and Gaze Data

Computer vision and human vision—both are often compared to each other, but only in rare cases are they combined for creating interactive CV. The cooperation of the user with the 3D reconstruction pipeline is the general concept of interactive reconstruction. In this vein why not use the users' visual system and NUI to boost the HCI as well as the underlying CV algorithms? The obvious and most straightforward information which can be measured by sensors from the human visual system are the points the eyes are looking at, the so-called gaze points. Gaze points are usually measured on a 2D target plain, typically a computer screen, and, over time, they result in gaze trajectories. With every task execution, the user provides these gaze trajectories as unconscious high-level information indicating, e. g., which areas are especially interesting. For realizing NUIs for OOI annotation and segmentation, the fusion of CV algorithms with human sensory data is of major importance. In order to enable efficient research on how CV boosted by sensory data, in particular by gaze data, a uniform standard file format for storing, visualizing, and using sensory data is mandatory and introduced in the following. As a result, sensory data of users watching video sequences or collections of images can be used, *inter alia*, for designing NUIs, improving CV methods, evaluating sensory-enhanced algorithms and training sensory-inspired ANNs.

3.1 Storing, Visualizing, and Using Sensory Data

Starting with the created pixel-accurate OOI and AOI annotations (cf. Chapter 2), it should be good practice to encapsulate these annotations and the corresponding video sequences in a standardized format [Paper VII], which is then easy to use by off-the-shelf software. Such a practice for storing text plus metadata, such as bookmarks, hyperlinks and comments has become standard since the introduction of the PDF container. Even today, metadata, i. e., research data such as OOI annotations and tags, are stored as an addition to the video in a single or multiple files. For storing these metadata such diverse formats as plain text, XML, *MATLAB* format, and binary are currently used. The plain text and the binary formats in particular are often unique data formats. Special tools are thus necessary to access and visualize these metadata on top of the video sequences. As a consequence, the accessibility of these data is quite hard for experts and for non-experts, infeasible.

Like the standard PDF format, state-of-the-art multimedia containers encapsulate video sequences, audio sequences, subtitles, chapter tags, cross-references, etc. for home entertainment systems in a single file. These containers have become common due to the everyday use by a broad audience. The best-known and most-used multimedia-encapsulation formats are the MP4 [44], OGG [105] and MKV [59] containers. Besides that, there is, in all research fields, an ever-increasing amount of video data sets which comprise additional metadata, such as annotations, tagged events, and inter-object relations (cf. Chapter 2). Consequently, an evaluation of the capabilities of multimedia containers has to be performed, focusing on the encapsulation of various kinds of research-relevant annotations, such as point markers, rectangular bounding boxes, and polygon-based bounding shapes.

As the first presented solution, AOI and OOI annotations, marked with ⧫ in Figure 3.1, were encapsulated in multimedia containers [Paper VII]. This way, different existing multimedia containers formats can be evaluated, and three possible implementations were identified. Two of these implementations use the *Matroška* container format (MKV) [59] with different types of subtitle formats, one being the *Advanced Sub Station Alpha* (ASS)[84] and another one being the *universal subtitle format* (USF) [65]. The third possible implementation is based on MPEG-7 [45] embedded into MPEG-4 [44] multimedia container data. Since, based on literature research, hardly any of the available standard multimedia players support MPEG-7, only the MKV-based approaches were prototyped and tested.

In order to support the decision to only implement the MKV-based prototypes, all three approaches are presented in a comparison matrix [Paper VII, Table 1]. Based on this matrix, the USF encapsulated in the MKV container as well as the MPEG-7 container seem the best suitable formats for providing instantaneous visualizations of pixel-accurate annotations in multimedia players. However, the prototype, where ASS is encapsulated

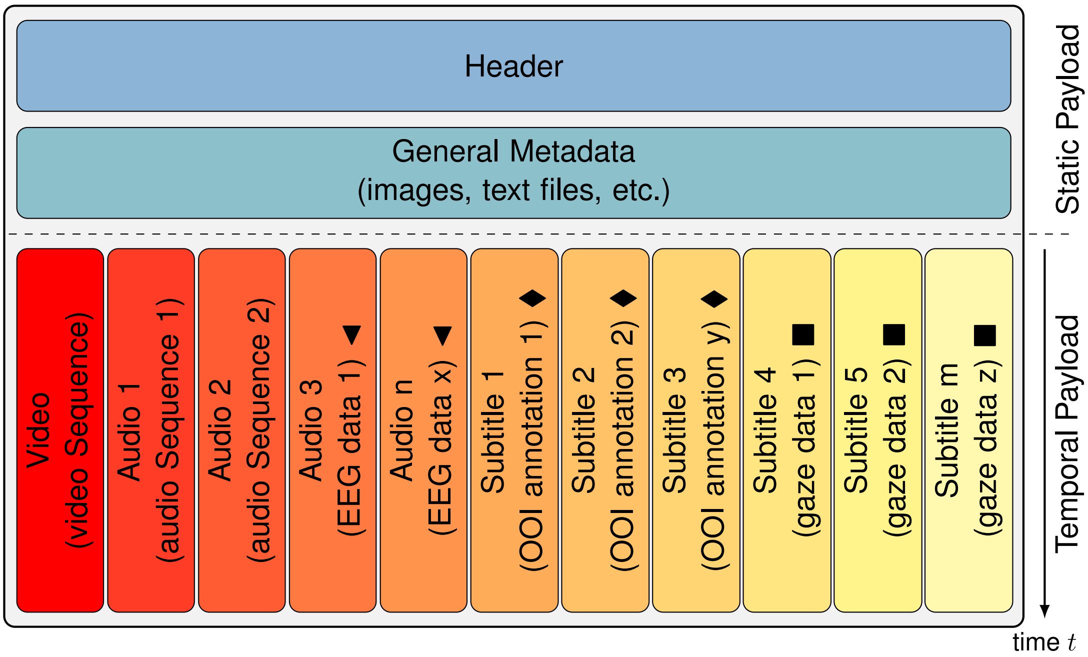


Figure 3.1: The general data structure of a multimedia container is split into static and dynamic payload, enabling streaming without transmitting the whole container beforehand. The encapsulation of several OOI annotations [Paper VII] is marked with ♦, gaze data [Paper VIII] from several users is marked with ■, and EEG data [Paper IX] in different sonifications is marked with ◀. Methods using research data in the proposed multimedia container for data visual analytics research [Paper X] simply visualize several subtitle tracks simultaneously or next to each other.

in the MKV container, has one significant advantage in comparison to other. It can be visualized by any state-of-the-art multimedia player without any modification of the players. Nevertheless, the best-suited format that can describe all relevant research data is MPEG-7, but, as mentioned before, there is a lack of MPEG-7 support in software libraries and multimedia players. In order to adequately demonstrate the potential of annotations stored in multimedia containers, seven video sequences of the Berkeley Video Segmentation Dataset [29], where OOIs are pixel-accurately annotated using the iSeg software (cf. Chapter 2), are muxed into a multimedia container. The resulting data set is public and can be downloaded by everyone. Thus, by the use of a standard multimedia player, the broad audience can watch the created OOI annotations and researchers can easily extract the annotations for their projects.

Advancing these first prototype implementations, eye tracking data is also integrated into the multimedia containers [Paper VIII]. The raw eye tracking data, as well as the video stimuli, are encapsulated into the MKV multimedia container, and for visualizing the gaze trajectories, they are additionally converted into a subtitle track, as illustrated with ■ in Figure 3.1. For promoting this approach, three well-known gaze data sets [70, 52, 16] in which the gaze trajectories of several participants are encoded and OOI and AOI annotations exist, are muxed into multimedia containers and released. Thus, the benefit of this approach is highlighted, and it becomes possible to instantaneously visualize all encoded metadata with standard video players. These converted gaze data sets show additionally that eye-tracking data in multimedia containers will significantly boost the accessibility and shareability for experts, researchers from other fields, and the broad audience.

Going even further, by enabling multimodal research in multimedia containers in both a visual and an auditive manner, metadata is sonificated using the audio tracks (cf. ◀ in Figure 3.1). As a consequence, it is shown that any research data can be visualized and sonificated by the use of multimedia containers. The main characteristics of such data can be summarized as follows: *(i) stimuli* like video sequences, images, as well as audio sequences, *(ii) metadata of the stimuli* like capturing details, object tags, subtitles, as well as labels, *(iii) additional object or scene data* like 3D descriptions, inter-object relations, online links, as well as scene maps and *(iv) sensory data of one or several participants* like gaze trajectories, heart rate, and electroencephalography (EEG) curves. Based on these containers, standard multimedia players can be used as tools for exploratory multimodal analysis of this data. Focusing on the exploratory multimodal data analysis on video and image stimuli, additional data sets are converted within the scope of first usability studies. For instance, the data set of Açık et al. [2] which consist of gaze trajectories of multiple participants watching a short video sequence followed by a freeze frame of it, is converted into the proposed containers to demonstrate instantaneous visualization of video sequences and still images. Showing other possibilities of the approach, further EEG data from neurophysiological recordings are sonificated and are encapsulated into a multimedia container together with the corresponding visualized gaze recordings as well as the stimuli video sequence. In a pilot study one expert and one non-expert perform explorative multimodal analysis with a multimedia player on theses two data sets. As a general result, both users highlighted the usefulness of research data in multimedia containers and were able to gain first impressions of the data sets only by the use of standard software.

Furthermore two add-ons for the VLC player are developed [Paper X] for visual analytics (VA) using standard and research data encoded in multimedia containers. Preliminary user interviews [Paper IX] identified the need for a parallel visualization of different items, like OOI annotation and gaze trajectories of several participants for detecting, e. g., inter-object relations. Since no known media player is designed to visualize several subtitle tracks in parallel, two VA add-ons, *SimSub* and *MergeSub*, are developed. *SimSub* visualizes every item, i. e., annotations and gaze trajectories in an additional video window [Paper X, Figure 5(a)]. All these video output windows are synchronized with the video controls of the main window. The user can then navigate through the video stimulus and directly see the changes by comparing the output windows. The second add-on, *MergeSub*, merges all items into a single subtitle track, which is visualized in the main window [Paper X, Figure 5(b)]. In this way, tiny differences between items can be found quickly and inter-object relations can be identified easily.

A more detailed user study with nine participants has conducted [Paper X] to get more feedback on both the approach of encapsulating research data in multimedia containers and the designed VA add-ons for the VLC player. After introducing the participants to the concept of storing research data in multimedia containers that allow direct visualization and sonification, they were asked to its significance for different user groups [Paper X, Figure 6]. By performing a given VA task using the designed VLC add-ons, all participants decided to use the *MergeSub* add-on only. One participant argued that *SimSub* cannot be used for a comparison of gaze points, because it is impossible to concentrate on two or more video windows at the same time. Four participants recommended the assignment colors to each metadata item as an improvement of *MergeSub*.

For encouraging more application as well as add-ons using these multimedia containers, in total six data sets comprising 319 video sequences with their corresponding research data such as annotations, eye tracking data, audio tracks, EEG measurements, and still frames were converted. The resulting 319 multimedia containers of data sets [Paper VII, Paper VIII, Paper IX] were published to facilitate multimodal research with standard multimedia players and to further enable novel VA in combination with the released VLC add-ons [Paper X]. One feature of this approach, however, has not been highlighted yet: By muxing metadata into multimedia containers, the data become streamable via the Internet, i. e. time-dependent data like video stimuli and their annotations are downloaded during playback and not beforehand. Consequently, the accessibility and shareability of research data will be significantly increased, so that the use of in multimedia container encapsulated research data facilitates the development of new applications and potentially of novel areas of research as well. One such area could be cognitive learning, where ANNs mimic human-like sensory input. Such ANNs might be used for testing existing theories of human sensory processing and for gaining new insights into how the human brain works. Another new area of research could be assistive technology, where gaze data might be used to highlight objects in movies for the visually impaired. An additional potential area would be the fusion of CV algorithms with human sensory data for creating reliable algorithms and new intuitive NUIs, as demonstrated in the next section.

3.2 Computer Vision Boosted by Sensory Data

A particular focus of research on the fusion of CV and sensory data is the question of how knowledge of the human visual system can be meaningfully captured and integrated into algorithms. The easiest method to get sensory data from the visual system nowadays is the capture of users' gaze points. Various hardware [87, 69, 76] and software [64, 18] solutions for gaze point capturing already exist. A logical first step in the development of an NUI for the iSeg software (cf. Chapter 2) is to investigate whether gaze trajectories improve object detection in megapixel (mp) images, both with respect to accuracy and processing time. Most current approaches to object detection are either based on keypoints with their feature vectors [66, 60, 12] or based on ANNs [27, 85, 50]. Usually, these approaches are benchmarked on datasets such as *ImageNet* [74], *SUN Database* [104] or *FlickrLogos-32/47* [71], whose image resolutions are significant below that of modern cameras. Current cameras have resolutions of, e. g., 2 mp for 1080p, 8 mp for 4k cameras and up to 50 mp for professional digital photography equipment.

In order to evaluate the performance of gaze-improved object detection in comparison to keypoint-based methods, a data set was created with a scene resolution of 5152×3864 (19.9 mp) with nine different objects with various resolutions. The developed interactive feature growing process [Paper XI, Figure 1], where users' gazes boost the detection of keypoints in order to improve detection accuracy and processing time, is benchmarked against the four most common keypoint feature detectors and their descriptors, namely *SIFT* [58], *SURF* [8], *ORB* [73], and *KAZE* [4].

Humans have a fast response time [88, 28] in visual search tasks. It is thus a reasonable hypothesis that the interactive feature growing process, which is affected neither by the image resolution nor the pixel density, should in general outperform the keypoint-based methods. For this interactive feature growing process, the users' gaze fixations on certain areas of the scene during a given visual search task are mapped. Based on these gaze maps, the AOIs the user's gazes fixate on are processed by the CV algorithms. As a consequence, users' unconscious knowledge and expectations encoded in their gaze trajectories reduce the use of traditional UI devices to a minimum and improve object detection within the scenes.

The result of the performed benchmark with ten participants [Paper XI, Figure 3] suggests, that interactive feature growing has a very high object detection rate of 95%. This detection rate is 6.3% better than the detection rate of the best keypoint-based method, which uses *KAZE* feature detectors and descriptors. Regarding processing time, the interactive feature growing is two times faster than the best keypoint-based method and has the same processing time as *SIFT*-based feature detectors and descriptors on scenes with 19.9 mp. As a general result, it is shown that users' gazes provide additional knowledge to the CV algorithms so that they outperform the other benchmarked approaches in accuracy, while the processing time is similar to the compared methods.

Going even further and using sensory boost CV not only for object detection, which can improve OOI annotation and segmentation, the benefit of users' eye movements in the 3D reconstruction pipeline is validated in a prototype implementation [Paper XII]. Since gaze features, such as saccades, smooth pursuit, and total time spend on a particular object, are indicators for task-relevant content, these features are used to prototype a content-aware 3D-reconstruction pipeline. Content-aware reconstruction means that only the OOI itself or the most important parts of the OOI should be reconstructed, which is contrary to all existing software solutions (cf. Section 1.1).

Building on the concept of the bio-inspired 3D reconstruction pipeline (cf. Chapter 4), which rests on pixel-accurate OOI recognition as well as the interactive feature growing, gaze data is used for the content-aware recognition of the AOIs. On the recognized AOIs, the OOI is segmented using the GrabCut [72] segmentation algorithm. By integrating gaze features into the GrabCut algorithm, pixels of the recognized AOIs are set as possible foreground pixels that the users' gazes fixate on are set as certain foreground, and all other pixels are set to possible background. The prototype of the whole content-aware 3D reconstruction pipeline [Paper XII, Figure 1] is tested on an eye tracking data set [52], where participants were asked to follow an OOI, in this case a red car, in a video stimulus.

The resulting 3D model only shows the red car almost without any artifacts of the scene. The qualitative reconstruction accuracy is similar to the accuracy gained by the bio-inspired 3D reconstruction architecture, but does not require OOI annotations with, e. g., the iSeg software. By combining this architecture with the emerging hardware of wearable eye trackers, new application areas become accessible. To name but a few applications, content-aware 3D reconstruction can facilitates HRC where the operators' gaze will help a robot to, e. g. understand where to grasp a particular component so that it can be mounted by the operator. A similar field of application is assistive technology, where content-aware 3D reconstruction could provide additional 3D shape information so that patients can easily guide their prosthetic arms or hands with the unconscious support of their gazes.

Vision is the process of discovering from images what is present in the world, and where it is.

David Courtnay Marr

4

Bio-Inspired 3D Reconstruction

In contrast to current CV methods for 3D reconstruction (cf. Section 1.1), perceiving the 3D shape of OOIs is a trivial task for humans [47, 82], even in challenging lighting conditions or with object occlusion. In case the differences between the CV algorithms and the human visual system are identified, the existing CV algorithms can be improved accordingly. Since these improvements are gained by the consideration of a biological system, they can be called bio-inspired approaches.

Summarizing the findings on humans' 3D perception, Ullman [91] developed the view-combination scheme. In this scheme, "cells along the hierarchy from *V1* to *V4* also show an increasing degree of tolerance for the position and size of their preferred stimuli" [91, p. 152] which leads to multiple pictorial representations of different views in which "an object appears to be represented in *IT*" [91, p. 152] cortex. Under the consideration of the view-combination scheme, a bio-inspired 3D reconstruction pipeline [Paper XIII, Paper XIV], which improves the existing 3D reconstruction methods, is designed. This bio-inspired pipeline facilitates the 3D model reconstruction of an OOI from several video sequences and image collections. Like humans, this pipeline should be capable of combining semantically different input sources, in order to perceive the 3D shape of the object.

As illustrated in Figure 4.1, the bio-inspired 3D reconstruction pipeline is comprised of six processing blocks. The first five processing blocks have significant analogies to humans' 3D perception. Thus they are bio-inspired. For obtaining applicable CAD-like 3D models of industrial use, one last technically-inspired processing block is also needed for creating a meshed model. Due to the current psychophysical evidence that "humans encode 3D objects as multiple viewpoint-specific representations that are largely 2D" [11], this last processing block has no analogies to the human visual system, and thus it is only technically-inspired. Like the abstraction level of data in the sensemaking process of VVA task [Paper XVIII, Figure 3], the value of the data representation increases with every processing block, while the amount of data decrease. In this case, the abstraction level of the raw image and video data increases from unclassified pixels, via OOI annotations to the symbolic 3D model of the OOI as result of the reconstruction process.

Starting with unclassified pixels such as input data, the first processing block converted these data into a data format capable of being processed. Here, as illustrated in the row of exemplary frames in Figure 4.1, any number of video sequences and also any collections of images can serve as data input. Pixel-accurate OOI recognition is the second processing step. In this essential processing block, it must be identified as to whether the OOI is part of the scene and in case it is, whether it is occluded. It must also be determined which pixels belong to the OOI, i. e., the OOI must be segmented as foreground. Especially for input data, showing the OOI in its real world, the extraction of the semantic information of "what" the OOI looks like, i. e., to recognize the OOI, is very challenging for current CV methods [Paper XI]. Building on this information, all background pixels not belonging to the OOI are nullified and frames as well as images, which do not contain any pixel belonging to the OOI, are also removed. By fitting the boundaries of the frames and images to only non-nullified pixels, multiple pictorial representations of the OOI are created. These multiple pictorial representations as 2D projections of the real-world OOI ideally show the OOI from various perspectives and scales. As discussed in the introduction, these different perspectives and scales are caused by taking images and video sequences. Based on these pictorial representations MVG methods solve the ill-posed problem and estimate the lost dimension, which lead to a 3D PC of the OOI. Finally, in the technically-inspired processing block mesh the PC to obtain an applicable 3D model.

During the implementation of these bio-inspired and with the focus on representation level, a further significant analogy became apparent. The transition of the multiple pictorial representations to the 3D PC corresponds to Ullman's statement, "recognition by multiple pictorial representations and their combinations constitutes a major component of 3D object recognition" [91, p. 154] of humans. This significant increase in the level of data abstraction, which happens in this transition, is possible only if the OOI is recognized on the 2D projections. For this recognition, semantic knowledge about "what" OOI is looking like is more important, than "where" the OOI is located in the real world.

These "what" and "where" tasks are associated in general with the ventral and dorsal visual stream of the visual cortex [94, 93]. The introduced process, first identified on different inputs if the specific OOI is shown and by generating multiple pictorial

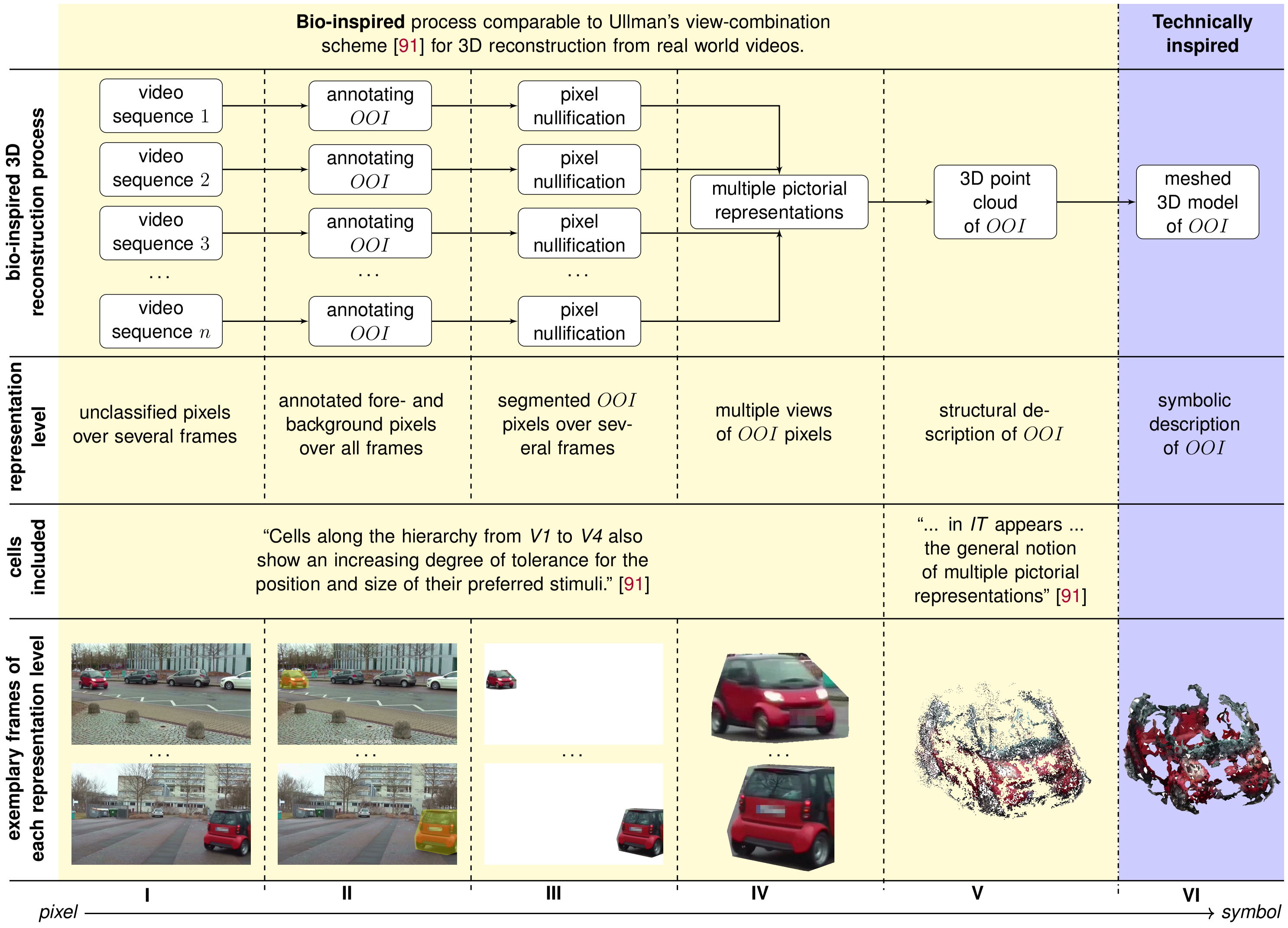

Bio-inspired process comparable to Ullman's view-combination scheme [91] for 3D reconstruction from real world videos.
Technically inspired
bio-inspired 3D reconstruction process
video sequence 1
video sequence 2
video sequence 3
...
video sequence n
annotating OOI
annotating OOI
annotating OOI
...
annotating OOI
pixel nullification
pixel nullification
pixel nullification
...
pixel nullification
multiple pictorial representations
3D point cloud of OOI
meshed 3D model of OOI
representation level
unclassified pixels over several frames
annotated fore- and background pixels over all frames
segmented OOI pixels over several frames
multiple views of OOI pixels
structural description of OOI
symbolic description of OOI
cells included
"Cells along the hierarchy from *V1* to *V4* also show an increasing degree of tolerance for the position and size of their preferred stimuli." [91]
"... in *IT* appears ... the general notion of multiple pictorial representations" [91]
exemplary frames of each representation level
...
...
...
...
I
II
III
IV
V
VI
pixel
symbol

Figure 4.1: Comparison of the bio-inspired 3D reconstruction process to the view-combination scheme by Ullman [91].

representations the spatial information about “where” the OOI is located, is removed. Thus, the proposed bio-inspired pipeline [Paper XIII, Paper XIV] has more similarities to the ventral stream because only the information “what” OOI looks like is used as additional input data. In contrast, 3D reconstruction approaches based on simultaneous localization and mapping (SLAM) estimate the locations of the moving camera and reconstruct the scene thereby [97, 13]. Therefore, SLAM methods mainly rely on “where” objects are located vis-à-vis the camera, i. e. these methods might have analogies to the dorsal visual stream. Since SLAM-based 3D reconstruction approaches are not capable of taking into account different video sequence as well as collections of images and in addition only work on a single video sequence without any cuts, these methods are neglected in this research.

For evaluating the bio-inspired 3D reconstruction process, three different OOI, a car, a table and a desk are reconstructed, each from two different video sequences [Paper XIV]. To analyze the quantitative result, the table and the desk provide ground truth geometry, since they are OOIs from computer-generated video sequences [15, 35]. As real-world video capture, the first two video sequences, showing a moving red car, of Kurzhals et al. [52] gaze tracking data set is chosen. Because no ground truth 3D shape of the car is available, the reconstruction quality of this OOI can be assessed only qualitatively but serves as a basis of comparison, e. g. gaze improved reconstruction pipelines [Paper XI]. Exemplary intermediate frames resulting from each processing block during the bio-inspired 3D reconstruction based on the red car are illustrated in Figure 4.1.

As a result of this evaluation, all three OOIs could be reconstructed as 3D PC [Paper XIV, Figure 2], such that only the OOI should be represented by the resulting PCs and not the whole scene. However, a closer consideration of the PCs obtained shows an enormous amount of noise, so that without colorization, even humans, could not perceive the 3D surface of the OOIs. By comparing the resulting PCs created by the bio-inspired reconstruction process to the resulting PCs of standard reconstruction software, it can be noticed that, particularly for the red car, no 3D model could be reconstructed [Paper XIV, Figure 3(a)]. For the computer-generated OOIs, the standard software solutions reconstruct the whole scene with the OOIs, instead of only the OOIs. By considering the quantitative result, which of course could be performed only on the virtual OOIs, no significant difference between the bio-inspired and the standard approaches was determined [Paper XIV, Figure 4].

By ventral stream-like object recognition, the bio-inspired 3D reconstruction process, shown in Figure 4.1, obtains 3D models of the OOI or its subparts only. Therefore, pixel-accurate recognition of the OOI is needed, which cannot automatically be done currently with CV algorithms. Nevertheless, for the implementation of this processing pipeline, interactive OOI annotation and segmentation methods could be used. For the evaluation performed, the iSeg software is used. Considering the qualitative results only, the last technically-inspired processing block requires significant improvements to mesh the PC of the OOI into an applicable CAD-like model. Due to the amount of noise within the resulting PCs, no standard meshing algorithms can be applied in a meaningful way for getting from the PC to a meshed model. Solving this issue, top-down approaches, i. e., from a 3D model to the PC might be an option. As technically-inspired top-down approaches, therefore, wireframes [109] or CAD models [103] can be fitted into the obtained PCs to create the optimal 3D geometry. For solving this problem by a bio-inspired approach, Biederman's idea of geons [10, 43], geometrical primitives which can be combined / subtracted to any 3D geometry, is a promising solution. A significant benefit of this solution is that these geons can approximate any 3D shape, even unknown shapes, without having a huge database of CAD models for example. Since nowadays, handcrafted process pipelines like the proposed one, are regularly outperformed by ANN-based approaches and their ability to reconstruct unknown objects are researched in the next chapter. Geon-based PC meshing, however, will be further considered in the conclusion of this thesis.

Standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy, in a very specific and satisfying sense.

Kurt Hornik, Stinchcombe Maxwell, and Halbert White

5

3D Reconstruction by Artificial Neural Networks

Over the last five years, the computing power of graphics processing units (GPUs) has improved remarkably due to the use of several thousand parallel computing cores. Thus, GPUs allow the massive parallelization of algorithms, so that the training as well as the execution of ANNs, even deep configurations with millions of training weights, become feasible. Up to now, it has been shown that ANNs outperform handcrafted algorithms in many fields of applications such as object recognition tasks [50, 32, 81]. In case ANNs can learn the underlying mathematical principle of MVG, then the trained ANNs can reconstruct any unknown 3D OOI without seeing it during training. Such an image-based reconstruction ANN implemented in interactive 3D reconstruction pipeline might, in contrast to the MVG algorithms, guarantee a smooth HCI, even in cases where the number and resolution of the input data make MVG infeasible.

In theory, as stated by Hornik et al. [41] “standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy, in a very specific and satisfying sense”. Hence, because the MVG methods usually consist of a concatenation of several algorithms, it must be possible to approximate its underlying mathematical principle by an ANN. For the research on this topic and in contrast to existing data sets like *ShapeNet* [14] and *ModelNet40* [102], a data set is needed that does not encode any object information, i. e., does not have shape priors of object categories. In addition, the data set must be scalable in its complexity and provide a large body of sample with ground truth data, ensuring the learning of even deep network configurations.

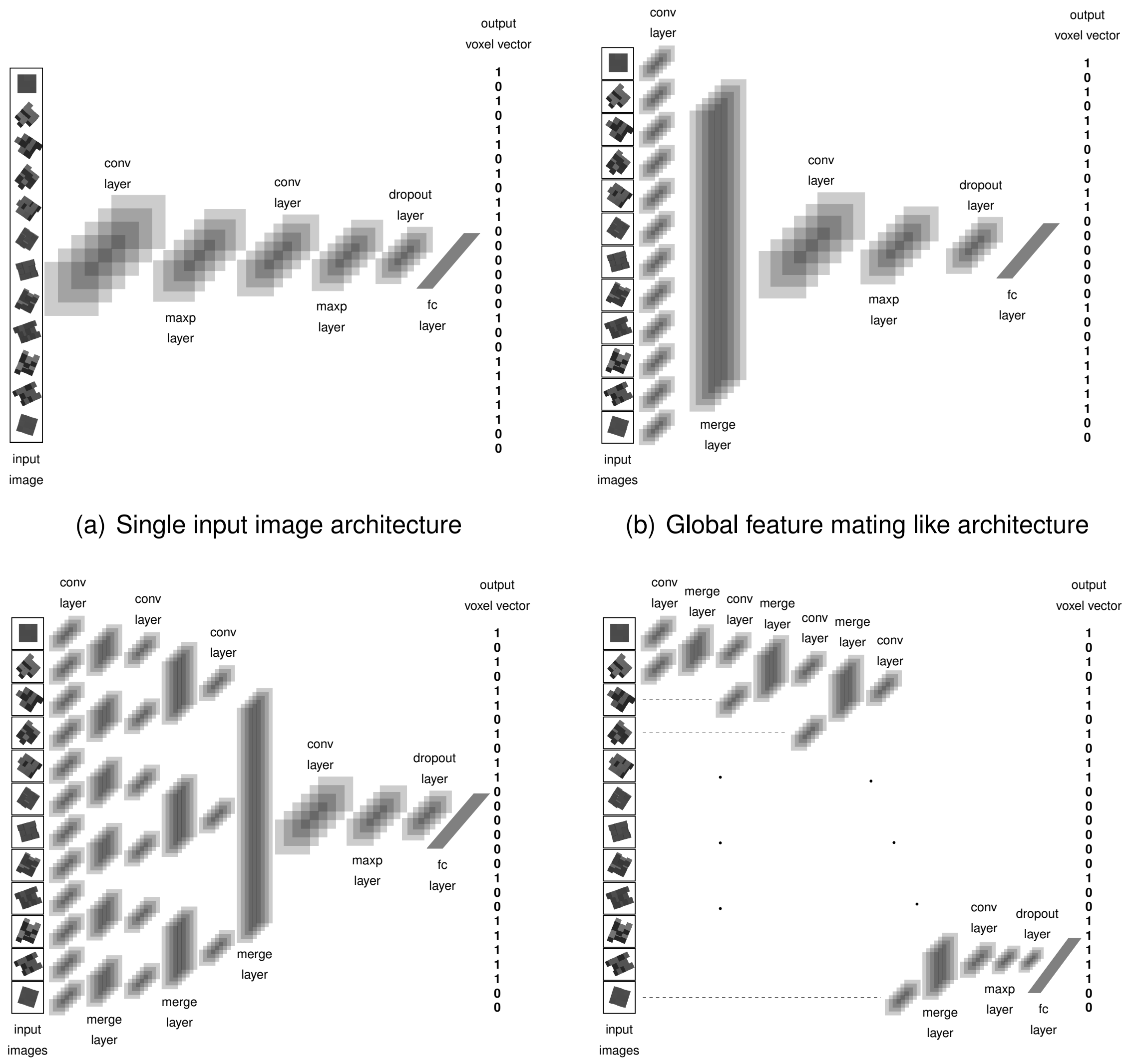
Because such a data set does not exist, a scalable cube data set is designed [Paper XV]. Preventing object categories, this data set provides random 3D objects and the associated 2D projections from various viewpoints. Therefore, a 3D object generator is developed, which returns the random objects in a common CAD format. This generator randomly creates, set and unset, i. e. visible and invisible, voxels in a $\langle r \times r \times r\rangle$ space, where for scalability the parameter r could be varied. The resulting output OOIs—random $\langle r \times r \times r\rangle$ cubes—are then stored as 3D objects and in addition, the information whether a voxel is set or not is saved. Obtaining the necessary 2D projections, showing the 3D OOI from different viewpoints, an image renderer is also implemented. For providing w different images of the objects in various perspectives, a minimal bounding sphere around the object is defined. On this sphere, the Fibonacci lattice [33] from the south- to the north pole is projected for evenly distributing the w camera centers. With the camera focus on the sphere center, then w grayscale images with the resolution $x \times x$ of the object are rendered. Due the bounding sphere design, this image renderer can generate w images from almost all 3D objects, not only from the generated cubes.

For comparable benchmarks, static imprints of the cubes in the 3×3×3, 4×4×4, and 8×8×8 setup with 100,000, 300,000, and 430,000 different objects are released, next to the generators tools [Paper XV]. Theron, a systematic investigation as to if and to what extent ANNs can learn the image-based 3D reconstruction on unknown geometries can be performed. The result of this investigation might lead to a better understanding of which network architecture designs are preferable for such a multi-classification task. In this case the designed ANN need to predict r^3 values of binary nature, i. e. whether or not each voxel is set.

On image-based classification tasks, simple feedforward CNNs have shown high performance. The MNIST handwritten digits data set [54], e. g., can be sufficiently solved by a five-layer ANN. As a starting point, an ANN working on the MNIST data set with five layers is adapted to the slightly different task of image-based voxel prediction. This adapted network [Paper XV], illustrated in Figure 5.1(a), concatenates all w images into one single input image and outputs as all benchmarked ANN architecture a voxel vector. The transition, i. e., the connection of the input image to the output voxel vector, should be learnt by a six-layer network comprising a convolutional (conv), a max pooling (maxp), a second conv, a second maxp, a dropout and a fully connected (fc) layer. After four iterations of training on the first 30,000 objects of the 3×3×3, this architecture predicts 88.68% of all voxels correctly. In the 4×4×4, the same architecture obtains a voxel accuracy of 70.80%. Unfortunately, even the voxel accuracy of 88.68% only leads to a cube accuracy, i. e., all voxels in a cube are predicted correctly, of 4.84% [Paper XVI].

In order to achieve a higher cube accuracy resulting from a higher voxel accuracy, deeper architectures, containing more trainable weights, are designed. Within 3D reconstruction pipelines, a global, a hierarchical, or an incremental method for feature matching is used depending on the amount and complexity of the data [77]. Assigning these three high-level concepts to the architecture of ANNs, the networks schematized in Figure 5.1(b) – Figure 5.1(d), are designed [Paper XVI]. As illustrated in these Figures, these networks do not concatenate the w images, instead each image is handled separately. For ensuring comparable benchmark results of the global, hierarchical, and incremental network, all of them have by design the same number of trainable weights.

The best performance in these benchmarks is achieved by the global setup with a cube accuracy of 58.60% resulting from a voxel accuracy of 98.04%. It is significant in the design of the global setup that after the first conv layer, all neurons are merged. In the direct comparison [Paper XVI, Table 3] the other architectures perform only slightly worse. An additional result is that the filter size of each conv layer does not influence the accuracy strongly, but alter the learning performance over the iteration of learning. Whether the voxel space is scaled up is quite critical to the performance of all designed networks: the global, hierarchical, incremental as well as the simple feedforward network. Even in the 8×8×8 voxel space, the voxel performance is only 66.74%. Thus all designed networks do not scale. This might be due to the fact that the resolution r of the space leads to 2^{r^3} possibilities, which the network needs to predict. By further scaling up, this leads almost to an infinite-dimensional problem. Accordingly, for voxel space of technical use with, e. g., r=100, ANN-based approaches for 3D reconstruction of an unknown object are not promising yet. The question whether ANNs can learn the underlying principle of MVG methods could not be answered in general, since the voxel-based out need for ANNs is not directly comparable to the vertice-based output of MVG approaches. Therefore, a shift from a voxel-based to either a vertice-, edge-, or face-based output is necessary, which also might solve the scalability issue of a voxel-based output.



(a) Single input image architecture

(b) Global feature mating like architecture

(c) Hierarchical feature mating like architecture

(d) Incremental feature mating like architecture

Figure 5.1: Simplified schemes of the different benchmarked ANN architectures for image-based 3D reconstruction. In architecture (a) all w images from various viewpoints are concatenated to a single input images. Based on this single input image a simple six layer architecture is used for the prediction of each single voxel [Paper XV]. The architectures (b)–(d) are inspired by the three different feature matching methods: global, hierarchical, as well as incremental [Paper XVI]. All w images from the object are used individually as input for the networks. As output all networks have the a binary vector of r^3 voxels—depending on the resolution of the voxel space.

Grau, teurer Freund, ist alle Theorie
Und grün des Lebens goldner Baum.
(All theory, dear friend, is grey
but the golden tree of life springs ever green)
Johann Wolfgang Goethe

6

Applications and Use Cases

The findings of interactive 3D reconstruction gained in this research are made even more valuable by developing applications that improve use cases. Bringing research from theory-heavy CV labs to applications of real practical use, our everyday life might indeed be positively improved. With the aim ensuring safety, the amount of video surveillance constantly increases—one will be filmed,for instance, by around 300 different cameras [67] during an average day in London. As a consequence, massive collections of video sequences are recorded and have to be analyzed. For an effective as well as efficient analysis, VVA is needed, but due to the high complexity of video data in the spatial and temporal domain, it cannot be automated yet. In addition, fictitious possibilities such as the seemingly infinite magnification of digital images, reliable face recognition including, e. g., the removal of sunglasses, and 3D reconstruction of occluded objects, are carried out by artificial intelligence in books and movies only. In reality, these VVA tasks are still performed by specially trained human specialists.

For enhancing the VVA in video surveillance and other use cases, such as eye tracking analysis [Paper XVII], an interactive architecture is designed, which supports the specialist during the cognitive process of sensemaking [75, 68]. In this architecture, the analyst is still the core as she / he incorporates semantic knowledge as well as the task definition into the sensemaking process. In general, the architecture [Paper XVIII, Figure 2] can be split into two recurrent alternating parts: the *content extraction and representation* and the *content-based reasoning*. In close interaction of humans and computers, the content extraction and representation translates pixels of a given video sequence into meaningful content such as AOIs and OOIs [46, 40]. By the support of supervised machine learning, frames showing such content are preprocessed to the analyst's needs in order to minimize of HCI that is necessary for operating efficiently. On a real task, content extraction and representation can mean that an analyst identifies a specific person in a crowd, marks the person as an OOI and the machine learning algorithm returns all frames that contain this OOI. Performing such a task automatically is not yet feasible, but is clearly desired by security agencies [34]. From a purely the architectural point of view, the creation of AOIs and OOIs results in the remaining data having no relevance any more. The amount of data can, thus be reduced, and, at the same time, the abstraction level of the data increases [Paper XVIII, Figure 3]. Based on the identified content, the analyst performs the reasoning process step-wise by monitoring events, understanding situations, creating hypotheses and, finally, proving or disproving these hypotheses. In case the specialist needs more content for the reasoning process, she / he can return to the content extraction and representation process of the architecture, and vice versa, until the hypotheses are either conformed or rejected.

Considering content extraction and representation, software for, e. g., accurate content extraction from video sequences is needed. The developed iSeg software [Paper V, Paper VI] is perfectly suitable for effectively annotating AOIs and OOIs. In particular for content representation, the use of the proposed multimedia containers (cf. Section 3.1) for encapsulating both the extracted content as metadata and the video sequences are extremely valuable. Since 3D models of OOIs have a higher data abstraction level than annotated 2D projections, the 3D models of OOIs can be integrated into practical VVA software applications [Paper XVIII]. As illustrated in Figure 6.1, the iSegs' GUI is extended with a PC viewer, so that the analyst can mark analysis-relevant parts within the interactively generated PC of the OOI. Based on the marked parts, this prototypical software provides the analyst with an adaptive video playback and that way it improves the overall reasoning process. With respect to VVA of eye tracking data, on the other hand, a second software is prototyped [Paper XVIII, Figure 4 (b)], which specifically builds on the VVA architecture. Using this software, the analyst can then formulate a proposition on the basis of the extracted content, additional metadata, such as eye tracking or EEG data, and logical grammar elements. For example, the analyst might create the proposition that “the gaze point of participant C intercepts the OOI“. Then all following frames, in which the proposition is satisfied, will be highlighted on the timeline.

Considering another possible field of application, interactive 3D reconstruction can enable survivors of natural disasters to reconstruct broken spare parts for the efficient temporary repair of devastated vital infrastructure. Such applications are, and will be, very relevant considering that in 2015 alone 98.6 million people were affected by 346 reported natural disasters [22]. With the interactive 3D reconstruction method, using only equipment that can be found in everyone's pockets or shelters, spare parts could be reconstructed to facilitate the repair of vital infrastructure until rescue teams arrive.

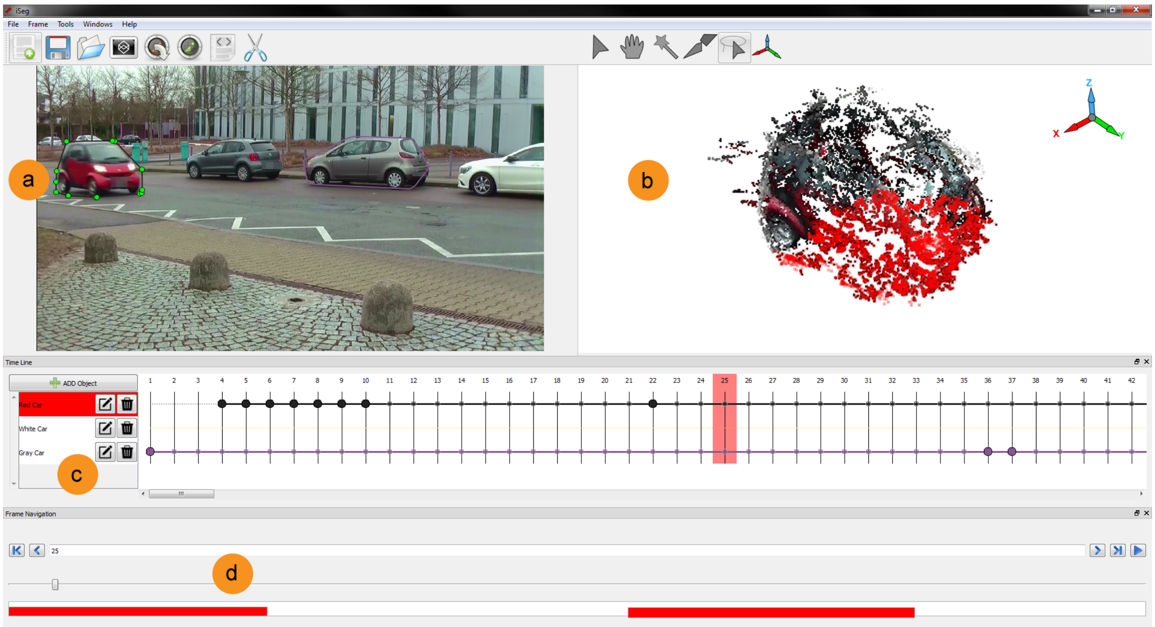


Figure 6.1: Exemplary application, for the adaptive playback of frames based on the marked parts of the OOI [Paper XVIII, Figure 4 (a)]. After the interactive creation of a 3D model, the analyst selects parts of interest in the PC viewer ⓑ and frames, where these parts are visible will be highlighted on the video timeline ⓓ. ⓐ and ⓒ are unchanged UIs of the iSeg application.

Therefore, an applicable system for image-based 3D reconstruction using only a smartphone or tablet computer and some additional materials such as newspaper and dust is developed and prototyped [Paper XIX]. After defining the scale dimension of the reconstructed 3D model using a new straightforward UI, the model must be transferred back from the virtual world to the real world. In the unlikely case that a 3D printer is available, the model can then easily be printed. However, since the existence of standard printers is more likely, the proposed system is capable of generating printable stencils that can be stacked for producing the spare part. Demonstrating this system, a gear wheel of a water pump is experimentally reconstructed as a 3D printout and as a stacked paper model [Paper XIX]. The functionality of the reconstructed part is tested as well, resulting a working water pump and for proving further support for this approach.

In all use cases where interactive image-based 3D reconstruction was prototyped, it returns applicable models with sufficient accuracy. However, especially in the spare part reconstruction scenario, the processing time of MVG prevents a smooth interaction between computers and humans. Solutions to this particular challenge and to others as well, are the topic of the following discussion.

7

Conclusion

In this cumulative dissertation, 19 articles on interactive 3D reconstruction have been presented which contribute to the ultimate aim of obtaining exact copies of OOIs, created with the help of image-based 3D reconstruction. The gained insights are already implemented in software applications (cf. Chapter 6). As demonstrated, the resulting 3D models can be printed for the temporary replacement of spare parts [Paper XIX] and the creation of video summaries [Paper XVIII], but unfortunately not yet for CAD-use cases.

CAD-use cases, however, are not fully covered by the presented research. In order to obtain CAD-like 3D models by image-based 3D reconstruction, two challenges still need to be solved. The first partially solved, challenge is the semantic division of OOIs into logical subparts. This division into subparts can already be done on 2D images using OOI annotation and segmentation (cf. Chapter 2) with, e. g., the iSeg software. In conjunction with the bio-inspired reconstruction architecture (cf. Chapter 4), this then results in noisy PC of the subparts. Performing the 3D PC clearing, in consideration of subpart identification, however, is still a completely manual task and needs to must be researched further. Once the PC is noise-free, regular meshing algorithms can compute a surface of the 3D model which is typically based on triangular faces. These meshed 3D models are suitable, for instance, for VR, AR, and 3D printing, but unfortunately not for CAD-tasks, since in CAD the 3D geometrical shapes are be expressed as curves in order to ensuring that, e. g., manufacturing machines are able to create smooth surfaces.

In order to create CAD applicable models and to address the second major challenge that sill needs to be solved, one main focus for further research should be the fitting of geometrical primitives into the 3D PC. This will ensure both the noise handling within the PC and the CAD-like representation of the 3D model which would no longer consists of a collection of triangular faces. Such an approach is similar to the ways engineers design 3D model in CAD software and reflects the idea of geons [10, 43]. Since computational power is now significantly larger than at the time of the first attempts of fitting geons into 3D shapes fail. The use of geons as geometrical primitives, approximating the 3D geometry of the OOI using GPU-implemented optimization algorithms, seems to be a highly promising solution.

In addition to the two challenges discussed above, it should be noted that even though image-based 3D reconstruction is feasible, the amount of user interaction is so high and the nature of this interaction so complex that completely manual 3D modeling is faster than the currently interactive implementation. The OOI annotation and segmentation with traditional UI (cf. Chapter 2) is particularly time-consuming. In order to improve on the current limitations, a gaze-based NUI for OOI annotation is currently being prototyped. The preliminary results, based on two video sequences, indicate that in 92% of the cases, majority pixels belonging to the OOI can be identified correctly. Carrying out more elaborate studies with the same principal design as this preliminary study may well benefit current research on content-aware 3D reconstruction [Paper XII].

Finally, the yet unanswered question whether ANNs can learn the underlying principles of MVG methods to reconstruct unknown OOIs (cf. Chapter 5) lies at the center of what will most likely be the most active research area in the context of image-based 3D reconstruction in the next decade. This trend is clearly visable in the number of articles that mention the keywords “ANN” and “3D reconstruction“ shown in Figure 1.1. The latest ANNs which are capable of reconstruct know 3D objects from 25 different object classes in a at maximum 32×32×32 voxel space. Bas on the evaluation of 3D reconstruction by ANNs, however, a shift from a voxel-based to either a vertice-, edge-, or face-based output is necessary to tackle the scalability issue of the voxel space. RNNs for 3D reconstruction seem to be particularly well-suited to this task, as with these kind of networks the number of input as well as output can be varied during the reconstruction process, which is similar to the number of keypoints in the classical MVG pipelines.

Summing up the research presented in this thesis, interactive 3D reconstruction can be used to create 3D models based on almost any kind of video sequences or image collections. Despite the considerable issues in the research area of image-based 3D reconstruction, this thesis has significantly improved the current reconstruction pipeline by advanced human-computer interaction. As the discussion in this shows there are still many challenging and existing question to be answered with respect to image-based 3D reconstruction—this research presented here is a step towards finding adequate and satisfying solutions.

P

Referenced Published Articles

Paper I: Taxonomy of 3D Sensors

A Survey of State-of-the-Art Consumer 3D-Reconstruction Sensors and Their Field of Applications

Abstract:

Sensors used for 3D-reconstruction determine both the quality of the results and the nature of reconstruction algorithms. The spectrum of such sensors ranges from expensive to low cost, from highly specialized to out-of-the-shelf, and from stereo to mono sensors. The list of available sensors has been growing steadily and is becoming difficult to manage, even in the consumer sector. We provide a survey of existing consumer 3D sensors and a taxonomy for their assessment. This taxonomy provides information about recent developments, application domains and functional criteria. The focus of this survey is on low cost 3D sensors at an accessible price. Prototypes developed in academia are also very interesting, but the price of such sensors can not easily be estimated. We try to provide an unbiased basis for decision-making for specific 3D sensors.
In addition to the assessment of existing technologies, we provide a list of preferable features for 3D reconstruction sensors. We close with a discussion of common problems in available sensor systems and discuss common fields of application, as well as areas which could benefit from the application of such sensors.

Originally published as:

J. Schöning and G. Heidemann. Taxonomy of 3D Sensors. A Survey of State-of-the-Art Consumer 3D-Reconstruction Sensors and Their Field of Applications. In: *Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*. SCITEPRESS - Science, and Technology Publications, 2016, pp. 194–199

DOI:

10.5220/0005784801920197

Paper II: Evaluation of Multi-view 3D Reconstruction Software

Abstract:

A number of software solutions for reconstructing 3D models from multi-view image sets have been released in recent years. Based on an unordered collection of photographs, most of these solutions extract 3D models using structure-from-motion (SFM) algorithms. In this work, we compare the resulting 3D models qualitatively and quantitatively. To achieve these objectives, we have developed different methods of comparison for all software solutions. We discuss the perfomance and existing drawbacks. Particular attention is paid to the ability to create printable 3D models or 3D models usable for other applications.

Originally published as:

J. Schöning and G. Heidemann. Evaluation of Multi-view 3D Reconstruction Software. In: *Computer Analysis of Images and Patterns (CAIP)*. Springer International Publishing, 2015, pp. 450–461

DOI:

10.1007/978-3-319-23117-4_39

Paper III: Interactive 3D Modeling

Abstract:

3D reconstruction and modeling techniques based on computer vision show a significant improvement in recent decades. Despite the great variety, a majority of these techniques depend on specific photographic collections or video footage. For example, most are designed for large data collections, overlapping photos, captures from turntables or photos with lots of detectable features such as edges. If the input, however, does not fit the particular specification, most techniques can no longer create reasonable 3D reconstructions. We review the work in the research area of 3D reconstruction and 3D modeling with a focus on the specific capabilities of these methods and possible drawbacks. Within this literature review, the practical usability with the focus on the input data—the collections of photographs or videos—and on the resulting models are discussed. Upon this basis, we introduce our position of interactive 3D reconstruction and modeling as a possible opportunity of lifting current restrictions from these techniques, which leads to the possibility of creating CAD-ready models in the future.

Originally published as:

J. Schöning and G. Heidemann. Interactive 3D Modeling. In: *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. SCITEPRESS - Science, and Technology Publications, 2015, pp. 289–294

DOI:

10.5220/0005259402890294

Paper IV: Interactive 3D Reconstruction

New Opportunities for Getting CAD-ready Models

Abstract:

A multitude of image-based 3D reconstruction and modeling techniques exist, which have achieved significant success in recent years. However, these techniques still lack certain abilities. For example, current 3D reconstruction techniques cannot decompose an object into its individual subparts. Thus, a printed model will consist of one single monolithic piece, which does not allow composing or decomposing parts, does not allow movable or flexible parts, and does not allow manufacturing the model from multiple different materials like wood, metal, or plastic. I reviewed the work in the research area of 3D reconstruction and provide an analysis of neglected research objectives and current drawbacks. Furthermore, I propose a mock-up of an interactive tool as a guideline for future research which describes a possible architecture, user interfaces, and processing pipeline, to overcome existing drawbacks of 3D reconstruction techniques.

Originally published as:

J. Schöning. Interactive 3D Reconstruction. New Opportunities for Getting CAD-ready Models. In: *Imperial College Computing Student Workshop (ICCSW)*. vol. 49. OpenAccess Series in Informatics (OASIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2015, pp. 54–61

DOI:

10.4230/OASIcs.ICCSW.2015.54

Paper V: Semi-automatic Ground Truth Annotation in Videos

An Interactive Tool for Polygon-based Object Annotation and Segmentation

Abstract:

Knowledge extraction from video data is challenging due to its high complexity in both the spatial and temporal domain. Ground truth is crucial for the evaluation and the adaptation of algorithms to new domains. Unfortunately, ground truth annotation is inconvenient and time consuming. Common annotation tools mostly rely on simple geometric primitives such as rectangles or ellipses. Here we propose a novel, interactive and semi-automatic process, which actively asks for user input if the result of the automatic annotation appears to be incorrect. After a brief review of related tools for video annotation, we explain our proposed semi-automatic method *iSeg* using a prototype implementation. *iSeg* has been tested on two visual stimulus datasets for eye tracking experiments and on two surveillance datasets. The experimental results and the usability are compared to existing annotation tools. Finally, we discuss the properties and opportunities of polygon-based video annotation.

Originally published as:

J. Schöning, P. Faion, and G. Heidemann. Semi-automatic Ground Truth Annotation in Videos. An Interactive Tool for Polygon-based Object Annotation and Segmentation. In: *International Conference on Knowledge Capture (K-CAP)*. ACM Press, 2015, 17:1–17:4

DOI:

10.1145/2815833.2816947

Paper VI: Pixel-wise Ground Truth Annotation in Videos

An Semi-Automatic Approach for Pixel-wise and Semantic Object Annotation

Abstract:

In the last decades, a large diversity of automatic, semi-automatic and manual approaches for video segmentation and knowledge extraction from video-data has been proposed. Due to the high complexity in both the spatial and temporal domain, it continues to be a challenging research area. In order to develop, train, and evaluate new algorithms, ground truth of video-data is crucial. Pixel-wise annotation of ground truth is usually time-consuming, does not contain semantic relations between objects and uses only simple geometric primitives. We provide a brief review of related tools for video annotation, and introduce our novel interactive and semi-automatic segmentation tool *iSeg*. Extending an earlier implementation, we improved *iSeg* with a semantic time line, multi-threading and the use of ORB features. A performance evaluation of *iSeg* on four data sets is presented. Finally, we discuss possible opportunities and applications of semantic polygon-shaped video annotation, such as 3D reconstruction and video inpainting.

Originally published as:

J. Schöning, P. Faion, and G. Heidemann. Pixel-wise Ground Truth Annotation in Videos. An Semi-Automatic Approach for Pixel-wise and Semantic Object Annotation. In: *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. SCITEPRESS - Science, and Technology Publications, 2016, pp. 690–697

DOI:

10.5220/0005823306900697

Paper VII: Providing Video Annotations in Multimedia Containers for Visualization and Research

Abstract:

There is an ever increasing amount of video data sets which comprise additional metadata, such as object labels, tagged events, or gaze data. Unfortunately, metadata are usually stored in separate files in custom-made data formats, which reduces accessibility even for experts and makes the data inaccessible for non-experts. Consequently, we still lack interfaces for many common use cases, such as visualization, streaming, data analysis, machine learning, high-level understanding and semantic web integration. To bridge this gap, we want to promote the use of existing multimedia container formats to establish a standardized method of incorporating content and metadata. This will facilitate visualization in standard multimedia players, streaming via the Internet, and easy use without conversion, as shown in the attached demonstration video and files. In two prototype implementations, we embed object labels, gaze data from eye-tracking and the corresponding video into a single multimedia container and visualize this data using a media player. Based on this prototype, we discuss the benefit of our approach as a possible standard. Finally, we argue for the inclusion of MPEG-7 in multimedia containers as a further improvement.

Originally published as:

J. Schöning, P. Faion, G. Heidemann, and U. Krumnack. Providing Video Annotations in Multimedia Containers for Visualization and Research. In: *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017

DOI:

10.1109/wacv.2017.78

Paper VIII: Eye Tracking Data in Multimedia Containers for Instantaneous Visualizations

Abstract:

Nowadays, the amount of gaze data records of subjects associated with video sequences increases daily. These eye tracking data are unfortunately stored in separate files in custom-made data formats, which reduces accessibility even for experts and makes the data effectively inaccessible for non-experts. Consequently, we still lack interfaces for many common use cases, such as visualization, streaming, data analysis, high level understanding, and semantic web integration of eye tracking data. To overcome these shortcomings, we want to promote the use of existing multimedia container formats to establish a standardized method of incorporating content videos with eye tracking metadata. This will facilitate instantaneous visualization in standard multimedia players, streaming via the Internet, and easy usage without conversion. Using our prototype software, we embed gaze data from eye tracking studies and the corresponding video into a single multimedia container, which can be visualized by any media player. Based on this prototype implementation, we discuss the benefit of our approach as a possible standard for storing eye tracking metadata including the corresponding video.

Originally published as:

J. Schöning, P. Faion, G. Heidemann, and U. Krumnack. Eye Tracking Data in Multimedia Containers for Instantaneous Visualizations. In: *IEEE VIS Workshop on Eye Tracking and Visualization (ETVIS)*. IEEE, 2016, pp. 74–78

DOI:

10.1109/etvis.2016.7851171

Paper IX: Exploratory Multimodal Data Analysis with Standard Multimedia Player

Multimedia Containers: a Feasible Solution to make Multimodal Research Data Accessible to the Broad Audience

Abstract:

The analysis of multimodal data comprised of images, videos and additional recordings, such as gaze trajectories, EEG, emotional states, and heart rate is presently only feasible with custom applications. Even exploring such data requires compilation of specific applications that suit a specific dataset only. This need for specific applications arises since all corresponding data are stored in separate files in custom-made distinct data formats. Thus accessing such datasets is cumbersome and time-consuming for experts and virtually impossible for non-experts. To make multimodal research data easily shareable and accessible to a broad audience, like researchers from diverse disciplines and all other interested people, we show how multimedia containers can support the visualization and sonification of scientific data. The use of a container format allows explorative multimodal data analyses with any multimedia player as well as streaming the data via the Internet. We prototyped this approach on two datasets, both with visualization of gaze data and one with additional sonification of EEG data. In a user study, we asked expert and non-expert users about their experience during an explorative investigation of the data. Based on their statements, our prototype implementation, and the datasets, we discuss the benefit of storing multimodal data, including the corresponding videos or images, in a single multimedia container. In conclusion, we summarize what is necessary for having multimedia containers as a standard for storing multimodal data and give an outlook on how artificial networks can be trained on such standardized containers.

Originally published as:

J. Schöning, A. L. Gert, A. Açık, T. C. Kietzmann, G. Heidemann, and P. König. Exploratory Multimodal Data Analysis with Standard Multimedia Player. Multimedia Containers: a Feasible Solution to make Multimodal Research Data Accessible to the Broad Audience. In: *Computer Vision, Imagingand Computer Graphics Theory and Applications (VISAPP)*. SCITEPRESS - Science and Technology Publications, 2017, pp. 272–279

DOI:

10.5220/0006260202720279

Paper X: Visual Analytics of Gaze Data with Standard Multimedia Players

Abstract:

With the increasing amount of studies, where participants eye movements are tracked watching video stimuli, the volume of gaze data records is growing tremendously. Unfortunately, in most cases, such data are collected in separate files in custom-made or proprietary data formats. These data are difficult to access even for experts and effectively inaccessible for non-experts. Normally expensive or custom-made software is necessary for their analysis. We want to solve this problem by using existing multimedia container formats for distributing and archiving eye tracking and gaze data bundled with the stimuli data. We define an exchange format that can be interpreted by standard multimedia players and can be streamed via the Internet. We converted several gaze data sets into our format, demonstrating the feasibility of our approach and allowing to visualize these data with standard multimedia players. We also introduce two VLC player add-ons, allowing for further visual analytics. We discuss the benefit of gaze data in a multimedia container and explain possible visual analytics approaches based on our implementations, converted datasets, and first user interviews.

Originally published as:

J. Schöning, C. Gundler, G. Heidemann, P. König, and U. Krumnack. Visual Analytics of Gaze Data with Standard Multimedia Players. In: *Journal of Eye Movement Research* 10.5 (2017), pp. 1–14

DOI:

10.16910/jemr.10.5.4

Paper XI: Interactive Feature Growing for Accurate Object Detection in Megapixel Images

Abstract:

Automatic object detection in megapixel images is quite inaccurate and a time and memory expensive task, even with feature detectors and descriptors like *SIFT*, *SURF*, *ORB*, and *KAZE*. In this paper we propose an interactive feature growing process, which draws on the efficiency of the users' visual system. The performance of the visual system in search tasks is not affected by the pixel density, so the users' gazes are used to boost feature extraction for object detection.

Experimental tests of the interactive feature growing process show an increase of processing speed by 50% for object detection in 20 megapixel scenes at an object detection rate of 95%. Based on this method, we discuss the prospects of interactive features, possible use cases and further developments.

Originally published as:

J. Schöning, P. Faion, and G. Heidemann. Interactive Feature Growing for Accurate Object Detection in Megapixel Images. In: *Computer Vision – ECCV Workshops*. Springer International Publishing, 2016, pp. 546–556

DOI:

10.1007/978-3-319-46604-0_39

Paper XII: Content-Aware 3D Reconstruction with Gaze Data

Abstract:

3D reconstruction has been shown to be a successful method for creating accurate 3D models out of video data with moving objects. Typically, videos are captured by ordinary cameras; however, more egocentric video footage will be taken by wearable cameras. In this work, we present a 3D reconstruction pipeline that implements content awareness for combining a wearable camera (a scene camera of an eye tracker) with gaze information. The aim is to identify the object of interest (OOI) within the video sequence. The OOI is identified within each frame for boosting the results of classical Structure from Motion (SfM) approaches, using the bio-inspired approach from an earlier study. We implemented a prototype based on the concept of content-aware 3D reconstruction using gaze data. Lastly, we gave an extensive overview of possible use case scenarios in a broad range of fields, starting from spare part reconstruction in difficult-to-access areas to assistive technologies, including exoskeletons and prosthetic arms / hands.

Originally published as:

J. Schöning, X. Jiang, C. Menon, and G. Heidemann. Content-Aware 3D Reconstruction with Gaze Data. In: *International Conference on Cybernetics (CYBCONF)*. IEEE, 2017

DOI:

10.1109/cybconf.2017.7985769

Paper XIII: Bio-Inspired Architecture for Deriving 3D Models from Video Sequences

Abstract:

In an everyday context, automatic or interactive 3D reconstruction of objects from one or several videos is not yet possible. Humans, on the contrary, are capable of recognizing the 3D shape of objects even in complex video sequences. To enable machines for doing the same, we propose a bio-inspired processing architecture, which is motivated by the human visual system and converts video data into 3D representations. Similar to the hierarchy of the ventral stream, our process reduces the influence of the position information in the video sequences by object recognition and represents the object of interest as multiple pictorial representations. These multiple pictorial representations are showing 2D projections of the object of interest from different perspectives. Thus, a 3D point cloud can be obtained by multiple view geometry algorithms. In the course of a detailed presentation of this architecture, we additionally highlight existing analogies to the view-combination scheme. The potency of our architecture is shown by reconstructing a car out of two video sequences. In case the automatic processing cannot complete the task, the user is put in the loop to solve the problem interactively. This human-machine interaction facilitates a prototype implementation of the architecture, which can reconstruct 3D objects out of one or several videos. In conclusion, the strengths and limitations of our approach are discussed, followed by an outlook to future work to improve the architecture.

Originally published as:

J. Schöning and G. Heidemann. Bio-Inspired Architecture for Deriving 3D Models from Video Sequences. In: *Computer Vision – ACCV Workshops*. Springer International Publishing, 2016, pp. 62–76

DOI:

10.1007/978-3-319-54427-4_5

Paper XIV: Ventral Stream-Inspired Process for Deriving 3D Models from Video Sequences

Abstract:

The reconstruction of complex 3D objects from video sequences captured by surveillance, smartphone, and other cameras is a common technique in Hollywood blockbusters and TV series. Unfortunately, the automatic or interactive 3D object reconstruction from this kind of videos is not yet possible in the real world. Enabling computers to recognize the actual 3D shape of objects from complex video sequences, we developed a bio-inspired processing architecture, motivated by findings in the area of human object recognition. By utilizing viewpoint-specific object recognition, changes in position and size of the object of interest in video sequences can be eliminated to a great extent. The result is a representation, comprised of multiple pictures showing 2D projections of the object of interest (OOI) from different viewpoints. Based on this representation, a 3D point cloud (PC) from the object can be obtained. After a detailed description of our architecture and its similarities to the human view-combination scheme, we demonstrate its potency by reconstructing several OOI from complex video sequences. Because some processing modules of the architecture cannot yet be fully automatized, we introduced interactive modules instead. Thus the prototypical implementation of our approach could be realized. Based on the resulting PC, we evaluate our architecture and consider more analogies between human and computer vision, which improve image-based 3D reconstruction.

Originally published as:

J. Schöning and G. Heidemann. Ventral Stream-Inspired Process for Deriving 3D Models from Video Sequences. In: *New Trends in Image Analysis and Processing – ICIAP Workshops:* Springer International Publishing, 2017

DOI:

10.1007/978-3-319-70742-6_7

Paper XV: Structure from Motion by Artificial Neural Networks

Abstract:

Retrieving the 3D shape of an object from a collection of images or a video is currently realized with multiple view geometry algorithms, most commonly Structure from Motion (SfM) methods. With the aim of introducing artificial neuronal networks (ANN) into the domain of image-based 3D reconstruction of *unknown* object categories, we developed a scalable voxel-based dataset in which one can choose different training and testing subsets. We show that image-based 3D shape reconstruction by ANNs is possible, and we evaluate the aspect of scalability by examining the correlation between the complexity of the reconstructed object and the required amount of training samples. Along with our dataset, we are introducing, in this paper, a first baseline achieved by an only five-layer ANN. For capturing life's complexity, the ANNs trained on our dataset can be used a as pre-trained starting point and adapted for further investigation. Finally, we conclude with a discussion of open issues and further work empowering 3D reconstruction on real world images or video sequences by a CAD-model based ANN training data set.

Originally published as:

J. Schöning, T. Behrens, P. Faion, P. Kheiri, G. Heidemann, and U. Krumnack. Structure from Motion by Artificial Neural Networks. In: *Scandinavian Conference on Image Analysis (SCIA)*. Springer International Publishing, 2017, pp. 146–158

DOI:

10.1007/978-3-319-59126-1_13

Paper XVI: Structure from Neuronal Networks (SfN^2)

Abstract:

Multiple View Geometry (MVG) with its underlying mathematical principle is mainly used for 3D reconstruction. The most common approaches based on MVG are the Structure from Motion (SfM) methods which create 3D point clouds from a collection of images or video frames. The emerging use of artificial neural networks (ANNs) in almost every domain leads to the question if ANNs can learn the underlying mathematical geometric mappings of SfM pipelines? To answer this question, three different ANN architectures based on the three different key point matching strategies of SfM were benchmarked. Since we want to learn the mathematical, geometrical mapping of SfM approaches and not the categories or shapes of natural 3D objects, we trained and tested our ANNs on 2D projections of random 3D shapes build from small random cubes. For 3D shapes with a grid size of $3 \times 3 \times 3$ voxels, all architectures show a high prediction accuracy of the reconstructed shape. When scaling up the grid size of the 3D cubes, we recognize a significant decrease in accuracy. These initial results show that all of the different ANN architectures we considered can learn to reconstruct unknown 3D shapes from images. In a more detailed analysis of our results, we investigate how the choice of architecture influences the prediction accuracy of the 3D shape on voxel and overall shape level and if nonoccluded voxels are be predicted independently of scale. Finally, we discuss if a voxel-based representation of the 3D shape can be scaled to a useful technical resolution due to its high impact on the size of the ANN as well as the required training data.

Originally published as:

J. Schöning, G. Heidemann, and U. Krumnack. Structure from Neuronal Networks (SfN^2). In: *Journal of Computers* 13.8 (2018), pp. 988–999

DOI:

10.17706/jcp.13.8.988-999

Paper XVII: Visual Analytics for Video Applications

Abstract:

In this article, we describe the concept of video visual analytics with a special focus on the reasoning process in the sensemaking loop. To illustrate this concept with real application scenarios, two visual analytics (VA) tools are discussed in detail that cover the sensemaking process: (i) for video surveillance, and (ii) for eye-tracking data analysis. Surveillance data (i) allow discussion of key VA topics such as browsing and playback, situational awareness, and the deduction of reasoning. Using example (ii) — eye tracking data from persons watching video — we review application features such as the space-time cube, spatio-temporal clustering, and automatic comparison of multiple participants. We examine how they can support the VA process. Based on this, open challenges in video VA will be discussed.

Originally published as:

P. Tanisaro, J. Schöning, K. Kurzhals, G. Heidemann, and D. Weiskopf. Visual Analytics for Video Applications. In: *it - Information Technology* 57.1 (2015), pp. 30–36

DOI:

10.1515/itit-2014-1072

Paper XVIII: Visual Video Analytics for Interactive Video Content Analysis

Abstract:

Reasoning is an essential processing step for any data analysis task, yet it requires semantic, contextual understanding on a high level, e.g., for the identification of entities. Developing an architecture for visual video analytics (VVA), we integrate human knowledge for highly accurate video content analysis to extract information by a tight coupling of automatic video analysis algorithms on the one hand and visualization as well as user interaction on the other hand. For accurate video content analysis, our semi-automatic VVA-architecture effectively understands and identifies *regular* and *irregular* behavior in real-world datasets. The VVA-architecture is described with both i) its *interactive information extraction and representation* and ii) its *information based reasoning* process. We give overview of existing techniques for information extraction and representation, and propose two interactive applications for reasoning. One of the applications uses 3D object representations to provide adaptive playback based on selected object parts in the 3D viewer. Another application allows the formulation of a proposition about the video by using all extracted objects and information. In case the proposition is true, the corresponding frames of the video are highlighted. Based on a user study, relevant open topics for increasing the performance of video content analysis and VVA is discussed.

Originally published as:

J. Schöning and G. Heidemann. Visual Video Analytics for Interactive Video Content Analysis. In: *Future of Information and Communication Conference (FICC)*. IEEE, 2018, pp. 183–190

ISBN:

978-1-5386-2056-4

Paper XIX: Image Based Spare Parts Reconstruction for Repairing Vital Infrastructure after Disasters

Creating or Ordering Replica of Spare Parts for Overhauling Infrastructure

Abstract:

From the very first hours after disasters such as earthquakes, hurricanes, floods, landslides, and tsunamis, survivors and rescue teams start repairing the devastated vital infrastructure. In doing so, spare parts are frequently needed, which cannot be delivered via destroyed roads or cannot be found in collapsed warehouses. In this work, we present an approach of how spare parts can be reconstructed out of images, and printed out by a normal or 3D printer. Therefore, a structure from motion algorithm is applied to a small number of images, showing the spare part from every direction, which can be captured by any kind of digital camera. Based upon the resulting dense 3D point clouds, a meshed 3D model is computed. After estimating the dimension of the real parts with a straightforward user interface, the meshed virtual model can be scaled correctly. In order to transfer the virtual 3D model back to the real world, our approach provides a 3D model format—for 3D printers—and a stacked paper format—for normal printers. In case no printers are available, the created 3D model can be transmitted via any communication network like GSM, digital radio, or internet to the next available printer or warehouse, nearby. For demonstrating our method, we experimentally reconstruct a gear wheel of a water pump. Finally, we discuss advantages, drawbacks and further steps–necessary for making our approach available.

Originally published as:

J. Schöning and G. Heidemann. Image Based Spare Parts Reconstruction for Repairing Vital Infrastructure after Disasters. Creating or Ordering Replica of Spare Parts for Overhauling Infrastructure. In: *Global Humanitarian Technology Conference (GHTC)*. IEEE, 2016, pp. 225–232

DOI:

10.1109/ghtc.2016.7857285

R
References

[1] 3DFlow. *3DF Zephyr - photogrammetry software - 3D models from photos*. November 2017. URL: https://www.3dflow.net/3df-zephyr-pro-3d-models-from-photos/.

[2] A. Açık, A. Bartel, and P. König. Real and implied motion at the center of gaze. In: *Journal of Vision* 14.2 (2014), pp. 1–19. DOI: 10.1167/14.1.2.

[3] Agisoft. *Agisoft PhotoScan*. November 2017. URL: http://www.agisoft.ru/.

[4] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. KAZE Features. In: *Computer Vision – ECCV*. Springer Berlin Heidelberg, 2012, pp. 214–227. DOI: 10.1007/978-3-642-33783-3_16.

[5] Autodesk, Inc. *Autodesk 123D Catch | 3d model from photos*. March 2017. URL: http://www.123dapp.com/catch.

[6] J. C. Balloch and S. Chernova. An RGBD segmentation model for robot vision learned from synthetic data. In: *Workshop on Spatial-Semantic Representations in Robotics*. 2017.

[7] T. Banerjee, M. Enayati, J. M. Keller, M. Skubic, M. Popescu, and M. Rantz. Monitoring patients in hospital beds using unobtrusive depth sensors. In: *Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2014, pp. 5904–5907. DOI: 10.1109/embc.2014.6944972.

[8] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In: *Computer Vision – ECCV*. Springer Berlin Heidelberg, 2006, pp. 404–417. DOI: 10.1007/11744023_32.

[9] Bentley. *3D Reality Modeling Software - ContextCapture*. November 2017. URL: https://www.bentley.com/en/products/product-line/reality-modeling-software/contextcapture.

[10] I. Biederman. Recognition-by-components: A theory of human image understanding. In: *Psychological Review* 94.2 (1987), pp. 115–147. DOI: 10.1037/0033-295x.94.2.115.

[11] H. H. Bulthoff, S. Y. Edelman, and M. J. Tarr. How Are Three-Dimensional Objects Represented in the Brain? In: *Cerebral Cortex* 5.3 (1995), pp. 247–260. DOI: 10.1093/cercor/5.3.247.

[12] S. Buoncompagni, D. Maio, D. Maltoni, and S. Papi. Saliency-Based Keypoint Reduction for Augmented-Reality Applications in Smart Cities. In: *New Trends in Image Analysis and Processing – ICIAP Workshops*. Springer International Publishing, 2015, pp. 209–217. DOI: 10.1007/978-3-319-23222-5_26.

[13] D. Caruso, J. Engel, and D. Cremers. Large-scale direct SLAM for omnidirectional cameras. In: *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 141–148. DOI: 10.1109/iros.2015.7353366.

[14] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. In: (December 9, 2015). arXiv: 1512.03012v1.

[15] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 5556–5565. DOI: 10.1109/cvpr.2015.7299195.

[16] A. Coutrot and N. Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. In: *Journal of Vision* 14.8 (2014), pp. 5–5. DOI: 10.1167/14.8.5.

[17] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3D shape scanning with a time-of-flight camera. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1173–1180. DOI: 10.1109/cvpr.2010.5540082.

[18] E. S. Dalmaijer, S. Mathôt, and S. V. der Stigchel. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. In: *Behavior Research Methods* 46.4 (2013), pp. 913–921. DOI: 10.3758/s13428-013-0422-2.

[19] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and Y. Kompatsiaris. A Survey of Semantic Image and Video Annotation Tools. In: *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*. Springer Berlin Heidelberg, 2011, pp. 196–239. DOI: 10.1007/978-3-642-20795-2_8.

[20] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs. In: *Computer graphics and interactive techniques – SIGGRAPH*. ACM Press, 1996. DOI: 10.1145/237170.237191.

[21] M. Ding, K. Lyngbaek, and A. Zakhor. Automatic registration of aerial imagery with untextured 3D LiDAR models. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8. DOI: 10.1109/cvpr.2008.4587661.

[22] U. N. O. for Disaster Risk Reduction (UNISDR) and C. for Research on the Epidemiology of Disasters (CRED). *2015 Disasters in Numbers*. November 2017. URL: http://www.unisdr.org/files/47804_2015disastertrendsinfographic.pdf.

[23] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In: *International Conference on Pattern Recognition (ICPR)*. IEEE Comput. Soc, 2000. DOI: 10.1109/icpr.2000.902888.

[24] J. Dupuis, S. Paulus, J. Behmann, L. Plumer, and H. Kuhlmann. A Multi-Resolution Approach for an Automated Fusion of Different Low-Cost 3D Sensors. In: *Sensors* 14.4 (2014), pp. 7563–7579. DOI: 10.3390/s140407563.

[25] R. A. El-laithy, J. Huang, and M. Yeh. Study on the use of Microsoft Kinect for robotics applications. In: *Position Location and Navigation Symposium (PLANS)*. IEEE, 2012, pp. 1280–1288. DOI: 10.1109/plans.2012.6236985.

[26] Elsevier B.V. *Scopus – Document search*. November 2017. URL: https://www.scopus.com/.

[27] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable Object Detection Using Deep Neural Networks. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. DOI: 10.1109/cvpr.2014.276.

[28] C. W. Eriksen and D. W. Schultz. Information processing in visual search: A continuous flow conception and experimental results. In: *Perception & Psychophysics* 25.4 (1979), pp. 249–263. DOI: 10.3758/bf03198804.

[29] F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele. A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis. In: *International Conference on Computer Vision (ICCV)*. IEEE, 2013. DOI: 10.1109/iccv.2013.438.

[30] L. Gallo, A. P. Placitelli, and M. Ciampi. Controller-free exploration of medical image data: Experiencing the Kinect. In: *Computer-Based Medical Systems (CBMS)*. IEEE, 2011. DOI: 10.1109/cbms.2011.5999138.

[31] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1231–1237. DOI: 10.1177/0278364913491297.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. DOI: 10.1109/cvpr.2014.81.

[33] Á. González. Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices. In: *Mathematical Geosciences* 42.1 (2009), pp. 49–64. DOI: 10.1007/s11004-009-9257-x.

[34] A. Gruber and J. Horchert. *Thomas de Maizière: Was hinter seinem Pilotprojekt zur Gesichtserkennung steckt - SPIEGEL ONLINE*. August 2017. URL: www.spiegel.de/netzwelt/netzpolitik/a-1164313.html.

[35] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: *International Conference on Robotics and Automation (ICRA)*. IEEE, 2014. DOI: 10.1109/icra.2014.6907054.

[36] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D Scene Reconstruction and Class Segmentation. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013. DOI: 10.1109/cvpr.2013.20.

[37] A. van den Hengel, A. Dick, T. Thormählen, B. Ward, and P. H. S. Torr. VideoTrace. In: *ACM Transactions on Graphics* 26.3 (2007), p. 86. DOI: 10.1145/1276377.1276485.

[38] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In: *Experimental Robotics*. Springer Berlin Heidelberg, 2014, pp. 477–491. DOI: 10.1007/978-3-642-28572-1_33.

[39] J. Heo, S. Jeong, H.-K. Park, J. Jung, S. Han, S. Hong, and H.-G. Sohn. Productive high-complexity 3D city modeling with point clouds collected from terrestrial LiDAR. In: *Computers, Environment and Urban Systems* 41 (2013), pp. 26–38. DOI: 10.1016/j.compenvurbsys.2013.04.002.

[40] B. Höferlin, M. Höferlin, G. Heidemann, and D. Weiskopf. Scalable video visual analytics. In: *Information Visualization* 14.1 (2013), pp. 10–26. DOI: 10.1177/1473871613488571.

[41] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

[42] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy. Visual odometry and mapping for autonomous flight using an RGB-D camera. In: *Springer Tracts in Advanced Robotics*. Springer International Publishing, 2016, pp. 235–252. DOI: 10.1007/978-3-319-29363-9_14.

[43] J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. In: *Psychological Review* 99.3 (1992), pp. 480–517. DOI: 10.1037/0033-295x.99.3.480.

[44] ISO/IEC. *Information technology—Coding of audio-visual objects—Part 14: MP4 file format (ISO/IEC 14496-14:2003)*. Tech. rep. 14496-14. International Organization for Standardization, 2003.

[45] ISO/IEC. *Information technology—Multimedia content description interface—Part 3: Visual (ISO/IEC 15938-3:2001)*. Tech. rep. 15938-3. International Organization for Standardization, 2001.

[46] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual Analytics: Scope and Challenges. In: Springer Berlin Heidelberg, 2008, pp. 76–90. DOI: 10.1007/978-3-540-71080-6_6.

[47] P. J. Kellman and T. F. Shipley. A theory of visual interpolation in object perception. In: *Cognitive Psychology* 23.2 (1991), pp. 141–221. DOI: 10.1016/0010-0285(91)90009-d.

[48] A. Kowdle, Y.-J. Chang, A. Gallagher, D. Batra, and T. Chen. Putting the User in the Loop for Image-Based Modeling. In: *International Journal of Computer Vision* 108.1-2 (2014), pp. 30–48. DOI: 10.1007/s11263-014-0704-x.

[49] A. Kowdle, Y.-J. Chang, D. Batra, and T. Chen. Scribble based interactive 3D reconstruction via scene co-segmentation. In: *Image Processing (ICIP)*. IEEE, 2011, pp. 2577–2580. DOI: 10.1109/icip.2011.6116190.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In: *Communications of the ACM* 60.6 (2017), pp. 84–90. DOI: 10.1145/3065386.

[51] K.-D. Kuhnert and M. Stommel. Fusion of Stereo-Camera and PMD-Camera Data for Real-Time Suited Precise 3D Environment Reconstruction. In: *Intelligent Robots and Systems (IROS)*. IEEE, 2006, pp. 4780–4785. DOI: 10.1109/iros.2006.282349.

[52] K. Kurzhals, C. F. Bopp, J. Bässler, F. Ebinger, and D. Weiskopf. Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli. In: *Workshop on Beyond Time and Errors Novel Evaluation Methods for Visualization (BELIV)*. ACM Press, 2014, pp. 54–60. DOI: 10.1145/2669557.2669558.

[53] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In: *International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 1817–1824. DOI: 10.1109/icra.2011.5980382.

[54] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[55] S. Lee, D. Jang, E. Kim, S. Hong, and J. Han. A real-time 3D workspace modeling with stereo camera. In: *Intelligent Robots and Systems (IROS)*. IEEE, 2005, pp. 2140–2147. DOI: 10.1109/iros.2005.1545105.

[56] M. Li, P. Wonka, and L. Nan. Manhattan-World Urban Reconstruction from Point Clouds. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 54–69. DOI: 10.1007/978-3-319-46493-0_4.

[57] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition Using Prioritized Feature Matching. In: *Computer Vision – ECCV 2010*. Springer Berlin Heidelberg, 2010, pp. 791–804. DOI: 10.1007/978-3-642-15552-9_57.

[58] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110. DOI: 10.1023/b:visi.0000029664.99615.94.

[59] Matroska. *Matroska Media Container*. November 2017. URL: https://www.matroska.org/.

[60] J. Moehrmann and G. Heidemann. FOREST - A Flexible Object Recognition System. In: *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. SCITEPRESS - Science, and Technology Publications, 2015. DOI: 10.5220/0005175901190127.

[61] T. V. Nguyen, J. Feng, and S. Yan. Seeing Human Weight from a Single RGB-D Image. In: *Journal of Computer Science and Technology* 29.5 (2014), pp. 777–784. DOI: 10.1007/s11390-014-1467-0.

[62] I. Nikolov and C. Madsen. Benchmarking Close-range Structure from Motion 3D Reconstruction Software Under Varying Capturing Conditions. In: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. Springer International Publishing, 2016, pp. 15–26. DOI: 10.1007/978-3-319-48496-9_2.

[63] Q. Pan, G. Reitmayr, and T. Drummond. ProFORMA: Probabilistic Feature-based On-line Rapid Model Acquisition. In: *British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2009. DOI: 10.5244/c.23.112.

[64] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI. 2016, pp. 3839–3845.

[65] C. Paris, L. Vialle, and U. Hammer. *TitleVision - USF specs*. November 2017. URL: http://www.titlevision.dk/usf.htm.

[66] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. In: *Computer Vision and Image Understanding* 150 (2016), pp. 109–125. DOI: 10.1016/j.cviu.2016.03.013.

[67] G. Pillai. *Caught on Camera: You are Filmed on CCTV 300 Times a Day in London*. November 2017. URL: www.ibtimes.co.uk/britain-cctv-camera-surveillance-312382.

[68] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: *International Conference on Intelligence Analysis*. 2005.

[69] Pupil Labs UG. *Mobile Eye Tracking Headset - Technical Specifications*. November 2017. URL: https://pupil-labs.com/pupil/#technical-specs.

[70] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit. Dynamic Saliency Models and Human Attention: A Comparative Study on Videos. In: *Computer Vision – ACCV*. Springer Berlin Heidelberg, 2012, pp. 586–598. DOI: 10.1007/978-3-642-37431-9_45.

[71] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol. Scalable logo recognition in real-world images. In: *International Conference on Multimedia Retrieval (ICMR)*. ACM Press, 2011, 25:1–25:8. DOI: 10.1145/1991996.1992021.

[72] C. Rother, V. Kolmogorov, and A. Blake. “GrabCut” interactive foreground extraction using iterated graph cuts. In: *ACM Transactions on Graphics* 23.3 (2004), pp. 309–314. DOI: 10.1145/1015706.1015720.

[73] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In: *International Conference on Computer Vision (ICCV)*. IEEE, 2011. DOI: 10.1109/iccv.2011.6126544.

[74] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[75] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In: *SIGCHI conference on Human factors in computing systems (CHI)*. ACM Press, 1993, pp. 269–276. DOI: 10.1145/169059.169209.

[76] SR Research Ltd. *EyeLink Portable Duo*. November 2017. URL: http://www.sr-research.com/eyelinkportableduo.html.

[77] J. L. Schonberger and J.-M. Frahm. Structure-from-Motion Revisited. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. DOI: 10.1109/cvpr.2016.445.

[78] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. DOI: 10.1109/cvpr.2017.272.

[79] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, pp. 519–528. DOI: 10.1109/cvpr.2006.19.

[80] B. Shneiderman. Response time and display rate in human performance with computers. In: *ACM Computing Surveys* 16.3 (1984), pp. 265–285. DOI: 10.1145/2514.2517.

[81] P. Simard, D. Steinkraus, and J. Platt. Best practices for convolutional neural networks applied to visual document analysis. In: *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE Comput. Soc, 2003, pp. 958–962. DOI: 10.1109/icdar.2003.1227801.

[82] E. S. Spelke. Principles of Object Perception. In: *Cognitive Science* 14.1 (1990), pp. 29–56. DOI: 10.1207/s15516709cog1401_3.

[83] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008. DOI: 10.1109/cvpr.2008.4587706.

[84] Sub Station Alpha. *ASS v4.00+ Script Format*. November 2017. URL: moodub.free.fr/video/ass-specs.doc.

[85] C. Szegedy, A. Toshev, and D. Erhan. Deep Neural Networks for Object Detection. In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 2553–2561.

[86] K. Thoeni, A. Giacomini, R. Murtagh, and E. Kniest. A comparison of multi-view 3D reconstruction of a rock wall using several cameras and a laser scanner. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-5 (2014), pp. 573–580. DOI: 10.5194/isprsarchives-xl-5-573-2014.

[87] Tobii AB. *Tobii EyeX Controller*. November 2017. URL: http://www.tobii.com/xperience/products/.

[88] L. M. Trick and J. T. Enns. Lifespan changes in attention: The visual search task. In: *Cognitive Development* 13.3 (1998), pp. 369–386. DOI: 10.1016/s0885-2014(98)90016-8.

[89] Y. Uh and H. Byun. Multi-view 3D reconstruction by random-search and propagation with view-dependent patch maps. In: *Multimedia Tools and Applications* 75.23 (2016), pp. 16597–16614. DOI: 10.1007/s11042-016-3621-x.

[90] Y. Uh, Y. Matsushita, and H. Byun. Efficient Multiview Stereo by Random-Search and Propagation. In: *Conference on 3D Vision (3DV)*. IEEE, 2014. DOI: 10.1109/3dv.2014.35.

[91] S. Ullman. *High-level vision: Object recognition and visual cognition*. 2nd ed. MIT press Cambridge, MA, 1997.

[92] A. O. Ulusoy, M. J. Black, and A. Geiger. Semantic Multi-view Stereo: Jointly Estimating Objects and Voxels. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. DOI: 10.1109/cvpr.2017.482.

[93] L. G. Ungerleider. "What" and "where" in the human brain. In: *Current Opinion in Neurobiology* 4.2 (1994), pp. 157–165. DOI: 10.1016/0959-4388(94)90066-3.

[94] L. G. Ungerleider and M. Mishkin. Two Cortical Visual Systems. In: *Analysis of Visual Behavior*. MIT Press, Boston, 1982, pp. 549–586.

[95] M. Vergauwen and L. V. Gool. Web-based 3D Reconstruction Service. In: *Machine Vision and Applications* 17.6 (2006), pp. 411–426. DOI: 10.1007/s00138-006-0027-1.

[96] Visual Geometry Group, University of Oxford. *Multi-view and Oxford Colleges building reconstruction - Dinosaur*. November 2017. URL: http://www.robots.ox.ac.uk/~vgg/data/data-mview.html.

[97] R. Wang, M. Schwörer, and D. Cremers. Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras. In: (August 25, 2017). arXiv: 1708.07878v1 [cs.CV].

[98] C. Wu. *SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT)*. November 2017. URL: https://github.com/pitzer/SiftGPU.

[99] C. Wu. Towards Linear-Time Incremental Structure from Motion. In: *International Conference on 3D Vision (3DV)*. IEEE, 2013, pp. 127–134. DOI: 10.1109/3dv.2013.25.

[100] C. Wu. *VisualSFM: A Visual Structure from Motion System*. November 2017. URL: http://ccwu.me/vsfm/.

[101] S. Wu, S. Zheng, H. Yang, Y. Fan, L. Liang, and H. Su. SAGTA: Semi-automatic Ground Truth Annotation in crowd scenes. In: *International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2014. DOI: 10.1109/icmew.2014.6890539.

[102] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. DOI: 10.1109/cvpr.2015.7298801.

[103] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In: *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2014, pp. 75–82. DOI: 10.1109/wacv.2014.6836101.

[104] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. SUN Database: Exploring a Large Collection of Scene Categories. In: *International Journal of Computer Vision* 119.1 (2014), pp. 3–22. DOI: 10.1007/s11263-014-0748-y.

[105] Xiph.org. *Ogg*. November 2017. URL: https://www.xiph.org/ogg/.

[106] H. M. Yip, K. K. Ho, M. H. A. Chu, and K. W. C. Lai. Development of an omnidirectional mobile robot using a RGB-D sensor for indoor navigation. In: *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 2014, pp. 162–167. DOI: 10.1109/cyber.2014.6917454.

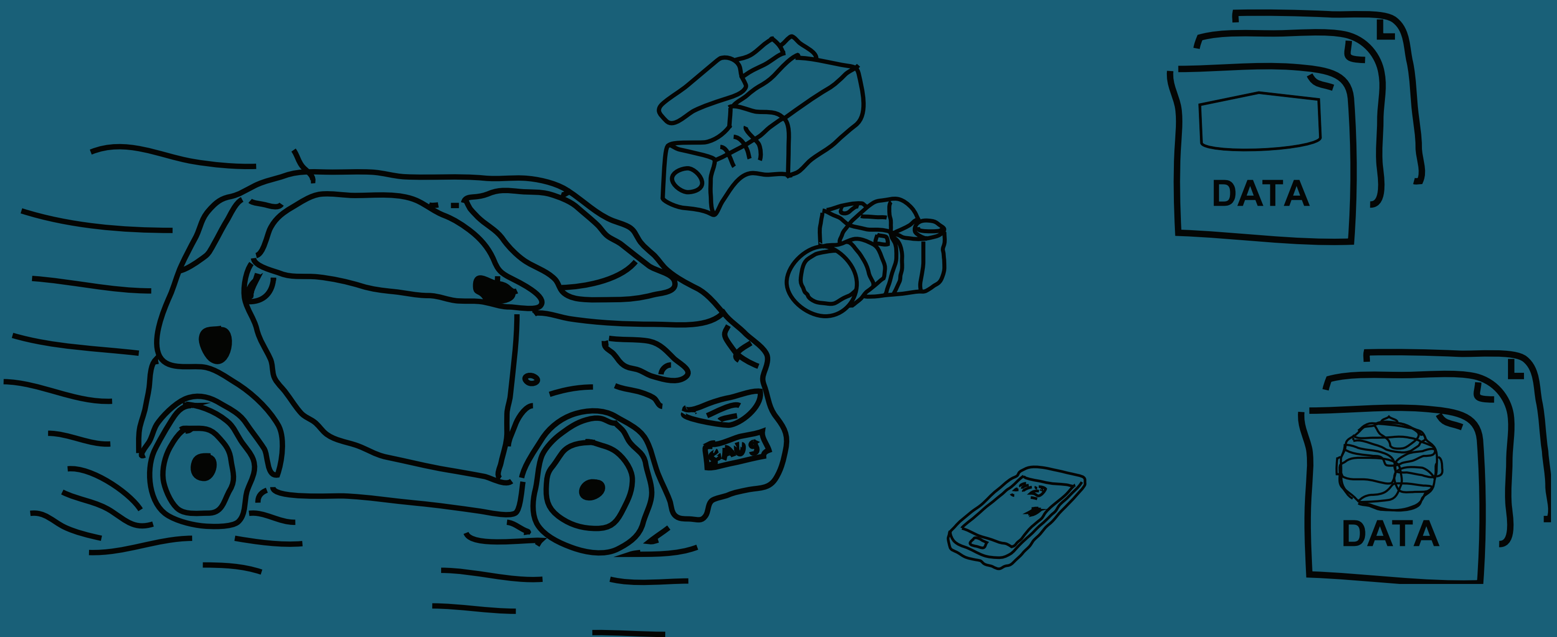
[107] YouTube. *Statistics - YouTube*. May 2015. URL: https://www.youtube.com/yt/press/statistics.html.

[108] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun. Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localization. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2222–2230. DOI: 10.1109/cvpr.2017.146.

[109] Z. Zhang, T. Tan, K. Huang, and Y. Wang. Three-Dimensional Deformable-Model-Based Localization and Recognition of Road Vehicles. In: *IEEE Transactions on Image Processing* 21.1 (2012), pp. 1–13. DOI: 10.1109/tip.2011.2160954.

[110] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene Parsing through ADE20K Dataset. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. DOI: 10.1109/cvpr.2017.544.

[111] M. Zollhöfer, C. Theobalt, M. Stamminger, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, and et al. Real-time non-rigid reconstruction using an RGB-D camera. In: *ACM Transactions on Graphics* 33.4 (2014), pp. 1–12. DOI: 10.1145/2601097.2601165.


GAUS
DATA
DATA