DGK Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 891

Ferdinand Maiwald

A window to the past through modern urban environments

– Developing a photogrammetric workflow for the orientation parameter estimation of historical images –

München 2022

Bayerische Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5303-8

Diese Arbeit ist gleichzeitig veröffentlicht:

Sächsischer Dokumenten- und Publikationsserver Qucosa unter
<https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-810852>, Dresden 2022

DGK Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 891

A window to the past through modern urban environments

– Developing a photogrammetric workflow for the orientation parameter estimation of historical images –

Von der Fakultät für Umweltwissenschaften
der Technischen Universität Dresden
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

von

M.Sc. Ferdinand Maiwald

aus Straubing

München 2022

Bayerische Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5303-8

Diese Arbeit ist gleichzeitig veröffentlicht:
Sächsischer Dokumenten- und Publikationsserver Qucosa unter
<https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-810852>, Dresden 2022

Adresse der DGK:

DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)
Alfons-Goppel-Straße 11 ● D – 80 539 München
Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100
e-mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Prof. Dr. habil. Hans-Gerd Maas, TU Dresden, Institute of Photogrammetry & Remote Sensing
Prof. Dr. Thomas P. Kersten, HCU Hamburg, Chair of Photogrammetry & Laser Scanning
Prof. Dr. habil. Pierre Grussenmeyer, INSA Strasbourg, Department of Surveying and Civil Engineering

Tag der Einreichung: 16.03.2022
Tag der Verteidigung: 23.06.2022

ISSN 0065-5325

ISBN 978-3-7696-5303-8

Statement of the Authorship

I, Ferdinand Maiwald, hereby certify that I have authored this dissertation entitled *A window to the past through modern urban environments – Developing a photogrammetric workflow for the orientation parameter estimation of historical images* and without additional resources and references than explicitly declared as such in the text by citation or footnote. I have marked both literal and accordingly adopted quotations as such. Unless otherwise stated, all graphics and illustrations were created by myself. The individual author contributions to the selected publications used in this cumulative dissertation are declared in the following section. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.

Author’s Contribution

This section declares the author’s contribution to the individual scientific articles that form the main part of the thesis. For the sake of convenience the author of the dissertation Ferdinand Maiwald is just to be referred as "the author".

Paper 1 (Chapter 2) entitled *Generation Of A Benchmark Dataset Using Historical Photographs For An Automated Evaluation Of Different Feature Matching Methods* has been published in the *ISPRS Archives*, an open access collection of conference papers published by Copernicus Publications, Göttingen (DE), on behalf of the International Society of Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS). The submitted paper has been double-blind peer reviewed. The work has been presented at the ISPRS Geospatial Week 2019 in Enschede, Netherlands. Conceptualization, description of the methodology, formal analysis, visualization, and writing of the paper have been carried out by the author alone. The paper was revised by Frank Henze. The resulting benchmark dataset uploaded on the TU Dresden Open Access Repository and Archive (OpARA) (`http://dx.doi.org/10.25532/OPARA-24`) has been technically revised by Christian Loeschen and professionally revised by Anette Eltner.

Paper 2 (Chapter 3) entitled *A 4D Information System For The Exploration Of Multitemporal Images And Maps Using Photogrammetry, Web Technologies And VR/AR* has been published in *Virtual Archaeology Review*, an international web-based, open-access, peer-reviewed scholarly journal published by the Universitat Politecnica de Valencia. Conceptualization, investigation, visualization and writing of the original draft was largely done by the author. The manuscript was co-authored by Christoph Lehmann, Jonas Bruschke, and Florian Niebling to a minor degree. The paper was revised by Frank Henze. The publication is an extension of the conference paper (Section 6.3) *Geo-information Technologies For A Multimodal Access On Historical Photographs And Maps For Research And Communication In Urban History* that has been published in the *ISPRS Archives*, a collection of peer-reviewed conference proceedings, and was presented at the 2nd International Conference of Geomatics and Restoration (GEORES 2019) in Milan, Italy. The original conference publication received the Best Paper Award.

Paper 3 (Chapter 4) entitled *An Automatic Workflow For Orientation Of Historical Images With Large Radiometric And Geometric Differences* has been published in *The Photogrammetric Record* by John Wiley and Sons, Hoboken, NJ (US). Conceptualization, description of the methodology, formal analysis, visualization, and writing of the paper have been largely done by the author. The paper was revised by Hans-Gerd Maas and Danilo Schneider. Minor linguistic corrections were made by the editor Stuart I. Granshaw.

Paper 4 (Chapter 5) entitled *Fully Automated Pose Estimation of Historical Images in the Context of 4D Geographic Information Systems Utilizing Machine Learning Methods* has been published in *ISPRS International Journal of Geo-Information* by MDPI AG, Basel. Both, the author and Christoph Lehmann contributed equally to the conceptualization, investigation, visualization and writing of the original draft, while Christoph Lehmann focused on Image Retrieval (Section 5.3.2.1) and the author on Photogrammetry (Section 5.4). Taras Lazariv contributed to the publication by implementing, evaluating and describing the second Image Retrieval approach (Section 5.3.2.2). The paper was revised by Peter Winkler and Danilo Schneider. The resulting benchmark dataset uploaded on OpARA (http://dx.doi.org/10.25532/OPARA-146) has been technically revised by Christian Loeschen and professionally revised by Anette Eltner.

Chapter 6 provides the cover pages of five associated peer-reviewed conference papers containing preliminary work. All these publication have been presented by the author on the respective conferences. The contribution of the corresponding authors is declared separately on every cover page.

Acknowledgements

My first thanks go to Prof. Dr. habil. Hans-Gerd Maas for his advice, valuable discussions and constant support throughout the constitution of this thesis. He always had an open ear for questions of any kind and gave me the academic freedom to complete my tasks.

Furthermore, I would like to express my special thanks to Prof. Dr. Thomas P. Kersten and Prof. Dr. habil. Pierre Grussenmeyer for reviewing this thesis as independent referees.

I especially want to thank all my colleagues at the Institute of Photogrammetry and Remote Sensing. In particular, I want to thank Danilo Schneider for his completely selfless guidance in the first years. This gratitude also goes to Frank Henze who unfortunately passed away too early.

I am very grateful that I was able to share my office with Melanie Elias who supported me in every phase of this thesis. The discussions with her were indispensable. My thanks also go to Anette Eltner for constant support and interest in the thesis. Further, I want to thank Hannes Sardemann and Robert Blaskow for the scientific discussions and support for the practical work done in this thesis. Similarly, I express my gratitude to Frank Liebold who could always provide a script for testing or validating my methods and to Franz Wagner who supported me in my work with neural networks and on the HPC. Therefore, I also gratefully acknowledge the GWK's support for this thesis by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden on HRSK-II. Special thanks go to Christoph Lehmann and Taras Lazariv who introduced me to the HPC and supported my research tremendously by developing a retrieval workflow according to my needs.

I am also very grateful that I was able to rely on external support. In particular, I want to thank Sander Münster for all the provided opportunities to present my research and to think out of the box. Similarly, I express my gratitude to Jonas Bruschke. Without the applications developed by him, my research would not have been as expressive. Further, I want to thank Torsten Sattler for the discussion and support with COLMAP, Dmytro Mishkin for the support with MODS and Paul-Edouard Sarlin for the support with SuperGlue. Thanks go to Denis Feurer and Robert Bischoff for providing historical datasets.

Last but not least, I want to express my gratefulness to my family who always supported me and my work and built me up when things were not going so well. My particular thanks go to my wife Sophie for her unlimited support. Especially, in stressful phases in which I often was distracted by the work. Without her I would not have been able to finish this thesis.

Abstract

The ongoing process of digitization in archives is providing access to ever-increasing historical image collections. In many of these repositories, images can typically be viewed in a list or gallery view. Due to the growing number of digitized objects, this type of visualization is becoming increasingly complex. Among other things, it is difficult to determine how many photographs show a particular object and spatial information can only be communicated via metadata.

Within the scope of this thesis, research is conducted on the automated determination and provision of this spatial data. Enhanced visualization options make this information more easily accessible to scientists as well as citizens. Different types of visualizations can be presented in three-dimensional (3D), Virtual Reality (VR) or Augmented Reality (AR) applications. However, applications of this type require the estimation of the photographer's point of view. In the photogrammetric context, this is referred to as estimating the interior and exterior orientation parameters of the camera. For determination of orientation parameters for single images, there are the established methods of Direct Linear Transformation (DLT) or photogrammetric space resection. Using these methods requires the assignment of measured object points to their homologue image points. This is feasible for single images, but quickly becomes impractical due to the large amount of images available in archives. Thus, for larger image collections, usually the Structure-from-Motion (SfM) method is chosen, which allows the simultaneous estimation of the interior as well as the exterior orientation of the cameras. While this method yields good results especially for sequential, contemporary image data, its application to unsorted historical photographs poses a major challenge.

In the context of this work, which is mainly limited to scenarios of urban terrestrial photographs, the reasons for failure of the SfM process are identified. In contrast to sequential image collections, pairs of images from different points in time or from varying viewpoints show huge differences in terms of scene representation such as deviations in the lighting situation, building state, or seasonal changes. Since homologue image points have to be found automatically in image pairs or image sequences in the feature matching procedure of SfM, these image differences pose the most complex problem.

In order to test different feature matching methods, it is necessary to use a pre-oriented historical dataset. Since such a benchmark dataset did not exist yet, eight historical image triples (corresponding to 24 image pairs) are oriented in this work by manual selection of homologue image points. This dataset allows the evaluation of frequently new published methods in feature matching. The initial methods used, which are based on algorithmic procedures for feature matching (e.g., Scale Invariant Feature Transform (SIFT)), provide satisfactory results for only few of the image pairs in this dataset. By introducing methods that use neural networks for feature detection and feature description, homologue features can be reliably found for a large fraction of image pairs in the benchmark dataset.

In addition to a successful feature matching strategy, determining camera orientation requires an initial estimate of the principal distance. Hence for historical images, the principal distance cannot be directly determined as the camera information is usually lost during the process of digitizing the analog original. A possible solution to this problem is to use three vanishing points that are automatically detected in the historical image and from which the principal distance can then be determined. The combination of principal distance estimation and robust feature matching is integrated into the SfM process and allows the determination of the interior and exterior camera orientation parameters of historical images. Based on these results, a workflow is designed that allows archives to be directly connected to 3D applications.

A search query in archives is usually performed using keywords, which have to be assigned to the corresponding object as metadata. Therefore, a keyword search for a specific building also results in hits on drawings, paintings, events, interior or detailed views directly connected to this building. However, for the successful application of SfM in an urban context, primarily the photographic exterior view of the building is of interest. While the images for a single building can be sorted by hand, this process is too time-consuming for multiple buildings.

Therefore, in collaboration with the Competence Center for Scalable Data Services and Solutions (ScaDS), an approach is developed to filter historical photographs by image similarities. This method reliably enables the search for content-similar views via the selection of one or more query images. By linking this content-based image retrieval with the SfM approach, automatic determination of camera parameters for a large number of historical photographs is possible. The developed method represents a significant improvement over commercial and open-source SfM standard solutions.

The result of this work is a complete workflow from archive to application that automatically filters images and calculates the camera parameters. The expected accuracy of a few meters for the camera position is sufficient for the presented applications in this work, but offer further potential for improvement. A connection to archives, which will automatically exchange photographs and positions via interfaces, is currently under development. This makes it possible to retrieve interior and exterior orientation parameters directly from historical photography as metadata which opens up new fields of research.

Zusammenfassung

Der andauernde Prozess der Digitalisierung in Archiven ermöglicht den Zugriff auf immer größer werdende historische Bildbestände. In vielen Repositorien können die Bilder typischerweise in einer Listen- oder Gallerieansicht betrachtet werden. Aufgrund der steigenden Zahl an digitalisierten Objekten wird diese Art der Visualisierung zunehmend unübersichtlicher. Es kann u.a. nur noch schwierig bestimmt werden, wie viele Fotografien ein bestimmtes Motiv zeigen. Des Weiteren können räumliche Informationen bisher nur über Metadaten vermittelt werden.

Im Rahmen der Arbeit wird an der automatisierten Ermittlung und Bereitstellung dieser räumlichen Daten geforscht. Erweiterterte Visualisierungsmöglichkeiten machen diese Informationen Wissenschaftlern sowie Bürgern einfacher zugänglich. Diese Visualisierungen können u.a. in drei-dimensionalen (3D), Virtual Reality (VR) oder Augmented Reality (AR) Anwendungen präsentiert werden. Allerdings erfordern Anwendungen dieser Art die Schätzung des Standpunktes des Fotografen. Im photogrammetrischen Kontext spricht man dabei von der Schätzung der inneren und äußeren Orientierungsparameter der Kamera. Zur Bestimmung der Orientierungsparameter für Einzelbilder existieren die etablierten Verfahren der direkten linearen Transformation oder des photogrammetrischen Rückwärtsschnittes. Dazu muss eine Zuordnung von gemessenen Objektpunkten zu ihren homologen Bildpunkten erfolgen. Das ist für einzelne Bilder realisierbar, wird aber aufgrund der großen Menge an Bildern in Archiven schnell nicht mehr praktikabel. Für größere Bildverbände wird im photogrammetrischen Kontext somit üblicherweise das Verfahren Structure-from-Motion (SfM) gewählt, das die simultane Schätzung der inneren sowie der äußeren Orientierung der Kameras ermöglicht. Während diese Methode vor allem für sequenzielle, gegenwärtige Bildverbände gute Ergebnisse liefert, stellt die Anwendung auf unsortierten historischen Fotografien eine große Herausforderung dar.

Im Rahmen der Arbeit, die sich größtenteils auf Szenarien stadträumlicher terrestrischer Fotografien beschränkt, werden zuerst die Gründe für das Scheitern des SfM Prozesses identifiziert. Im Gegensatz zu sequenziellen Bildverbänden zeigen Bildpaare aus unterschiedlichen zeitlichen Epochen oder von unterschiedlichen Standpunkten enorme Differenzen hinsichtlich der Szenendarstellung. Dies können u.a. Unterschiede in der Beleuchtungssituation, des Aufnahmezeitpunktes oder Schäden am originalen analogen Medium sein. Da für die Merkmalszuordnung in SfM automatisiert homologe Bildpunkte in Bildpaaren bzw. Bildsequenzen gefunden werden müssen, stellen diese Bilddifferenzen die größte Schwierigkeit dar.

Um verschiedene Verfahren der Merkmalszuordnung testen zu können, ist es notwendig einen vororientierten historischen Datensatz zu verwenden. Da solch ein Benchmark-Datensatz noch nicht existierte, werden im Rahmen der Arbeit durch manuelle Selektion homologer Bildpunkte acht historische Bildtripel (entspricht 24 Bildpaaren) orientiert, die anschließend genutzt werden, um neu publizierte Verfahren bei der Merkmalszuordnung zu evaluieren. Die ersten verwendeten Methoden, die algorithmische Verfahren zur Merkmalszuordnung nutzen (z.B. Scale Invariant Feature Transform (SIFT)), liefern nur für wenige Bildpaare des Datensatzes zufriedenstellende Ergebnisse. Erst durch die Verwendung von Verfahren, die neuronale Netze zur Merkmalsdetektion und Merkmalsbeschreibung einsetzen, können für einen großen Teil der historischen Bilder des Benchmark-Datensatzes zuverlässig homologe Bildpunkte gefunden werden.

Die Bestimmung der Kameraorientierung erfordert zusätzlich zur Merkmalszuordnung eine initiale Schätzung der Kamerakonstante, die jedoch im Zuge der Digitalisierung des analo-

gen Bildes nicht mehr direkt zu ermitteln ist. Eine mögliche Lösung dieses Problems ist die Verwendung von drei Fluchtpunkten, die automatisiert im historischen Bild detektiert werden und aus denen dann die Kamerakonstante bestimmt werden kann. Die Kombination aus Schätzung der Kamerakonstante und robuster Merkmalszuordnung wird in den SfM Prozess integriert und erlaubt die Bestimmung der Kameraorientierung historischer Bilder. Auf Grundlage dieser Ergebnisse wird ein Arbeitsablauf konzipiert, der es ermöglicht, Archive mittels dieses photogrammetrischen Verfahrens direkt an 3D-Anwendungen anzubinden.

Eine Suchanfrage in Archiven erfolgt üblicherweise über Schlagworte, die dann als Metadaten dem entsprechenden Objekt zugeordnet sein müssen. Eine Suche nach einem bestimmten Gebäude generiert deshalb u.a. Treffer zu Zeichnungen, Gemälden, Veranstaltungen, Innen- oder Detailansichten. Für die erfolgreiche Anwendung von SfM im stadträumlichen Kontext interessiert jedoch v.a. die fotografische Außenansicht des Gebäudes. Während die Bilder für ein einzelnes Gebäude von Hand sortiert werden können, ist dieser Prozess für mehrere Gebäude zu zeitaufwendig.

Daher wird in Zusammenarbeit mit dem Competence Center for Scalable Data Services and Solutions (ScaDS) ein Ansatz entwickelt, um historische Fotografien über Bildähnlichkeiten zu filtern. Dieser ermöglicht zuverlässig über die Auswahl eines oder mehrerer Suchbilder die Suche nach inhaltsähnlichen Ansichten. Durch die Verknüpfung der inhaltsbasierten Suche mit dem SfM Ansatz ist es möglich, automatisiert für eine große Anzahl historischer Fotografien die Kameraparameter zu bestimmen. Das entwickelte Verfahren stellt eine deutliche Verbesserung im Vergleich zu kommerziellen und open-source SfM Standardlösungen dar.

Das Ergebnis dieser Arbeit ist ein kompletter Arbeitsablauf vom Archiv bis zur Applikation, der automatisch Bilder filtert und diese orientiert. Die zu erwartende Genauigkeit von wenigen Metern für die Kameraposition sind ausreichend für die dargestellten Anwendungen in dieser Arbeit, bieten aber weiteres Verbesserungspotential. Eine Anbindung an Archive, die über Schnittstellen automatisch Fotografien und Positionen austauschen soll, befindet sich bereits in der Entwicklung. Dadurch ist es möglich, innere und äußere Orientierungsparameter direkt von der historischen Fotografie als Metadaten abzurufen, was neue Forschungsfelder eröffnet.

Contents

1 Introduction

The increasing digitization of historical photographs in archives allows the development of new access strategies. Usually, the digitized images are presented in a conventional two-dimensional (2D) gallery or list visualization. In contrast, a three-dimensional (3D) depiction of images comes with several advantages. Ware (2021) summarizes those in five categories: A proper visualization

- makes a large amount of data accessible and comprehensible.
- uncovers unseen details and allows the gathering of new information.
- reveals errors, artifacts and quality of the data used.
- enhances the linkage and understanding between large-scale and small-scale data.
- helps with formulating hypothesis.

In fact, all of these advantages are discovered through-out this thesis when extending 2D with a spatial (3D) and temporal (four-dimensional (4D)) component.

However, the depiction of images in Web3D, Virtual Reality (VR) and Augmented Reality (AR) applications requires the determination of the original camera's pose. While the camera pose of contemporary images can be obtained using state-of-the-art Structure-from-Motion (SfM) workflows, the precise localization of historical image collections proves more difficult.

This raises the need to identify reasons why SfM fails especially on historical photographs. Further, the identified problems need to be solved in a modified SfM workflow, preferably in a completely automatic pose estimation pipeline. Finally, it has to be tested how accurately the poses can be determined.

1.1 Thesis structure

This thesis is arranged in a cumulative manner. The individual scientific articles integrated in the thesis are published in three international peer-reviewed journals and the introductory article is a peer-reviewed conference publication. Each publication is preceded by a cover page which includes the original title, authorship, publication history, the recommended citation style, and the abstract. The thesis is written in a uniform layout which requires the adaption of the journals' formatting standards. This includes slight changes in formal text design as well as referencing style, placement and sizes of figures, tables, and equations. The original numbering of figures, tables, and equations is adapted to the related chapter numbering. Chapter headings corresponding to the individual articles have been customized providing a comprehensible context. All formatting changes are done in order to enhance the reading flow and consistent layout of the thesis.

The articles are antedated by the introduction (Chapter 1) and enclosed by the synthesis (Chapter 7). The reference list at the end of this thesis refers solely to literature citations made in these two chapters.

The thesis focuses on automatic pose estimation of cameras using an adapted SfM pipeline on historical urban images. The introduction (Chapter 1) intends to give an overview of the history of photography and highlights structure and data quality in contemporary archives. Furthermore a detailed description of the different parts of the SfM workflow is given and how these have to be adapted in order to process historical images. The introduction is concluded with consequential research objectives.

The following four chapters (Chapters 2-5) comprise the individual scientific articles. The first publication (Chapter 2) identifies the most common issues leading to incorrect or failing pose estimations of historical images when using conventional SfM software. This justifies the necessity of creating and publishing a benchmark dataset consisting of historical images exclusively. The second article (Chapter 3) presents a theoretical strategy for a completely automatic pose estimation workflow classifying Photogrammetry as a link between repository and application. First results of advanced feature matching methods and their performance on the prior developed benchmark dataset are included and compared with state-of-the-art SfM software. The third publication (Chapter 4) introduces neural networks in order to match historical image pairs in combination with Vanishing Point Detection (VPD) for principal distance estimation. The developed method is compared with state-of-the-art SfM software. The final article (Chapter 5) includes content-based image retrieval preceding the pose estimation and such enabling a complete automatic workflow from repository to local scene reconstruction. Different feature matching methods based on neural networks are tested and evaluated on their performance and accuracy. An additional benchmark dataset is created, which focuses on complete scene reconstructions using exclusively historical images.

The thesis is concluded with the synthesis (Chapter 7) which presents two final versions of the SfM workflow for the estimation of interior and exterior camera parameters of historical images. The workflows are evaluated and critically assessed in terms of accuracy and transferability. The thesis closes with recent developments and an outlook on further research opportunities.

1.2 Historical image data and archives

The history of photography starts with Niépce, Daguerre and Talbot who have been the first fixating projected light onto a material while Sir John Herschel coined the term *Photography* in 1839 (Hirsch, 2017). Shortly after that, Aimé Laussedat and Albrecht Meydenbauer can be seen as pioneers of using photographs for measurement purposes beginning in 1858 (Grimm, 2021). Following a personal correspondence with Dr. Otto Kersten, Albrecht Meydenbauer coined the term *Photogrammetry* in 1867 (Albertz, 2009) and some of his cameras and photographs can still be accessed and viewed today (Grimm, 2021). More details on the history of Photogrammetry can be found in Luhmann et al. (2019), 1.4 and in the special issue on the history of Photogrammetry in the PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, Vol. 89, No. 5 (Cramer and Kresse, 2021).

However, this thesis does not only deal with photographs taken for the purpose of measurements, but more commonly images by tourists, hobby and professional photographers. As most of the photographs have been fixated on analogue mediums like glass plates or film, image mostly refers to the digital copy of the original historical source. In the context of the presented work the term "historical" is used for images taken before approximately 2005. The oldest photographs used are from approximately 1850. An overview of time range and quality is given in Figure 1.1.

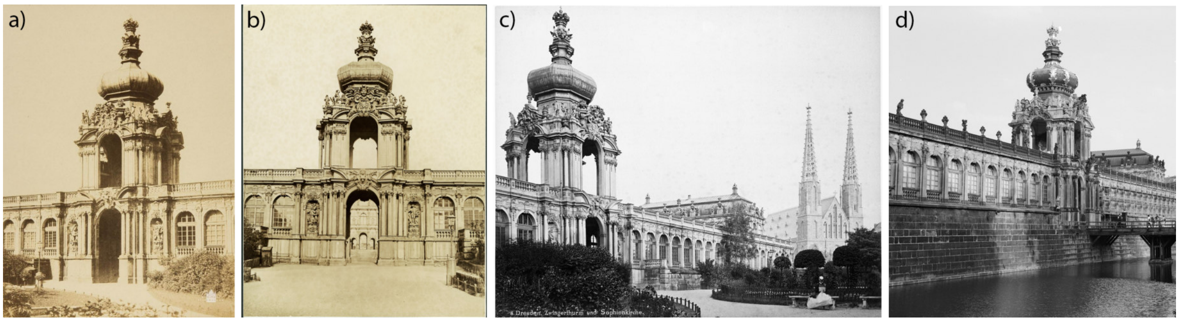


Figure (1.1): Four images of the Crowngate of the Dresden Zwinger showing the time range and quality of images obtained from Deutsche Fotothek. Image a) is dated to 1856. Image b) is dated between 1850 and 1889. Image c) is dated to 1998 (metadata - incorrect) and to 1880 (image description - correct). Image d) is dated 1998.

Most of the photographs processed in this thesis origin from one of the largest German image archive "Deutsche Fotothek" (`http://www.deutschefotothek.de/`) holding 2,193,000 images of 95 institutions (as of 03/2022). Other comparable archives are e.g., the European archive Europeana (`europeana.eu`), the Hungarian archive Fortepan (`https://fortepan.hu/`), or the Dutch archive of the Rijksmuseum (`rijksmuseum.nl`). All these repositories provide the possibility to search via metadata or using additional filters like e.g. date, depicted object, and topic. The process of digitization and metadata tagging of various data sources is not finished yet and still one of the most important topics for digital archives.

For photogrammetric data processing, digitization of historical images brings two major issues:

1. Metadata: A search query in archives is mostly done by using keywords. Thus, relevant photographs can only be found, if they are tagged with correct and meaningful metadata. This requires a prior knowledge of the expected search result *and* it implies that the uploading operator assigned the appropriate metadata. Considering that metadata standards and operators change during the ongoing digitalization the quality and correctness of metadata can not be guaranteed (see Fig. 1.1c).

2. Camera parameters: Digitization of measurement photographs is often meticulously documented and done with photogrammetric scanners. This is mostly not the case when working with amateur photographs where commercial high-resolution scanners are used commonly. While the original size and material of the analogue medium is available sometimes, information about the camera used is usually lost. That means, that especially interior camera parameters cannot be obtained and have to be estimated deliberately.

1.3 Structure-from-Motion for historical images

In order to use historical images for accurate texture projection or 3D reconstruction, the exterior orientation (= camera pose) and interior orientation of the camera has to be determined. Accurate interior and exterior orientation parameters of contemporary single images can be reliably obtained using the Direct Linear Transformation (DLT) or the non-linear space resection if spatial information about the observed scene is given.

In theory, these methods also allow the pose estimation of historical images *if* the depicted building is still intact and did not change during time. In practice, this requires the identification of the depicted object and the subsequent measurement of homologue points in object space and in the corresponding image. Considering that the "Deutsche Fotothek" alone holds 2.2 million images, this would be an enormous task which can not be tackled by a single person. Nonetheless, there exists research on specific single objects of interest where these methods are used (Wiedemann et al., 2000; Henze et al., 2009b; Bitelli et al., 2017; Bevilacqua et al., 2019; Kalinowski et al., 2021).

For the pose estimation of cameras for larger non-targeted image sets and unordered image collections the term Structure-from-Motion has become widely accepted, while this method can also be seen as a concatenation of photogrammetric algorithms and methods (Luhmann et al. (2019), 5.5.2.2). It has already been analyzed in other works how to adapt the workflow to process inhomogeneous contemporary data (Snavely et al., 2006; Agarwal et al., 2011; Schönberger and Frahm, 2016) only partly transferable to historical images. Others processed exclusively historical images using additional prior information such as interior camera parameters (Grün et al., 2004; Grussenmeyer and Al Khalil, 2017; Kalinowski et al., 2021), manually selected tie points (Grün et al., 2004; Schindler and Dellaert, 2012; Zawieska and Markiewicz, 2016) or further object information (Grussenmeyer and Al Khalil, 2017; Kalinowski et al., 2021).

However, the automatic estimation of interior and exterior camera parameters using SfM for exclusively historical images without further prior information has not been a primary subject of research so far. In the following, a comprehensive analysis of every single step of the SfM workflow is given and how to customize the stages in order to process large unordered historical image datasets.

1.3.1 Terminology

The term Structure-from-Motion (also *Structure from Motion*, *structure from motion*, *structure-from-motion*) origins from the idea to generate 3D information (structure) out of multiple images or a moving camera/object (motion). It was established by Ullmann (1979) who formulated the theorem:

Theorem. *Given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views are uniquely determined.*

This implicates, that "3-D structure can be recovered from as few as four points in three views" (Ullmann, 1979). While this is the minimal configuration to solve the problem of determining 3D structures, modern approaches rely on the redundancy given by using more images and hundreds or even thousands of homologue points seen in these images. The general workflow of modern SfM solutions is depicted in Figure 1.2.

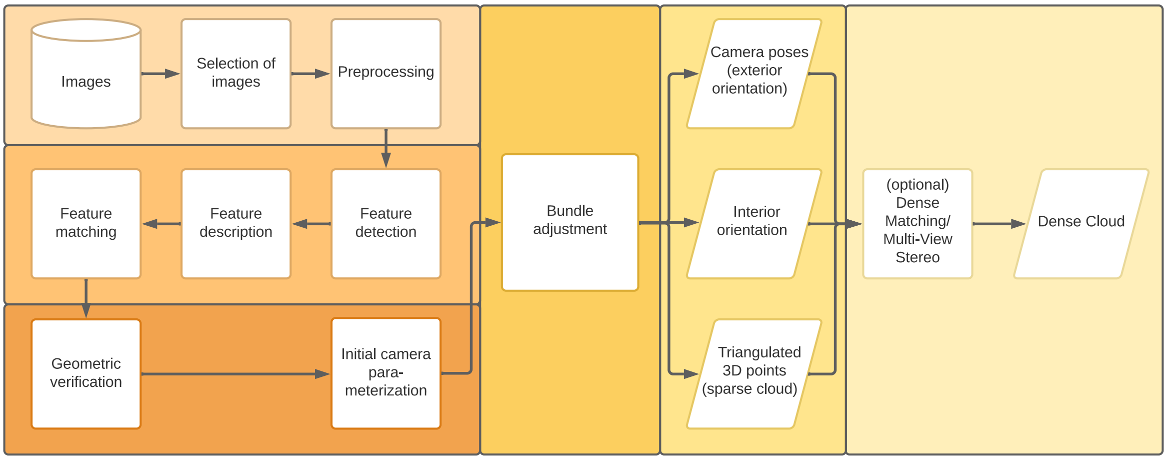


Figure (1.2): The different steps of the Structure-from-Motion workflow including an optional subsequent dense matching procedure.

In the following, every step starting from the selection of images up to the generation of dense 3D models is analyzed and it is shown how the different parts can be adapted for processing historical photographs. This thesis focuses especially on terrestrial urban images.

1.3.2 Selection of images and preprocessing

The first step to generate 3D structures is the initial selection of appropriate images. This step is often neglected as SfM is mostly used e.g., in industry (Bösemann, 2005; Luhmann, 2010), archaeology (Brutto and Meli, 2012; McCarthy, 2014), forestry (Iglhaut et al., 2019), and geomorphology (Westoby et al., 2012; Eltner and Sofia, 2020) for sequential image data. That implies, that the object of interest is photographed from different point of views with the intention to retrieve its 3D geometry. Consequently, all images are *ordered* and sequential image pairs have a *small baseline*. Additionally, these images are taken in similar conditions (lighting, temperature) with the same camera in a short time span and thus an initial filtering of images is not needed. Proprietary software like Agisoft Metashape provides some simple options to filter by image quality (blurriness, radiometry) by analyzing the sharpness of image borders.

In the case of applying the SfM workflow to archival or internet images it cannot be assumed that all images are taken sequentially or hold any specific order (Schaffalitzky and Zisserman, 2002). Image datasets of this kind are mostly referred to as unordered image/photo collections. In the past years, there have been several approaches retrieving the 3D structure of objects using increasingly large internet photo collections (Snavely et al., 2006; Agarwal et al., 2011; Radenovic et al., 2016; Heinly et al., 2015).

The images are usually pre-filtered and ordered throughout the SfM process. Features are calculated for every (down-sampled) image and the selection of valid image pairs is seen as a graph matching problem (Schaffalitzky and Zisserman, 2002; Agarwal et al., 2011). That means, that most approaches identify a small amount of features in every image and try to match these features in all other images. If matches can be found between image pairs, the images are added to an image graph and similar images should be close to each other in such a graph structure. There are various optimizations of building such an image graph/match graph (Snavely et al., 2008; Wu, 2013; Schönberger and Frahm, 2016) but these approaches were not useful for filtering historical images mainly due to four reasons:

1. Historical images are not easily describable by common feature descriptors because even very similar viewpoints can depict e.g., large radiometric variations or occlusions.

2. It is possible that the data is not evenly distributed and an image graph could not be built because of the difficulties in wide-baseline matching.

3. Metadata search in repositories yields results that are not meant to be reconstructed in a first instance (close-up images, interior views, public events).

4. It happens that metadata search yields more irrelevant results than relevant hits which is usually not the case when browsing other web collections.

Consequently, a different approach for finding relevant image data for SfM processing has been developed with the help of the Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig (ScaDS.AI) in this thesis. The aim is to filter a repository using Content-based image retrieval (CBIR) in order to find terrestrial exterior views of a specific landmark. While there is ongoing research on many different variations of image retrieval (Arandjelovic et al., 2016; Noh et al., 2017; Radenovic et al., 2018) it is not clear if these methods are applicable to historical images. The initial approach shows an overall accuracy rate around 70-80% by using the pre-trained neuronal network VGG16 (Simonyan and Zisserman, 2014) with modified dense layers (Chapter 3). These first promising results lead to an in-depth examination of the topic using and modifying implementations of Razavian et al. (2016) and Noh et al. (2017).

The neural network design, implementation and evaluation is mainly carried out by the ScaDS.AI and described in detail in Chapter 5. Thus, it is not meant to be a main part of this thesis‘ introduction. The developed approach outperforms conventional metadata search and reliably yields images of a respective landmark with a low rate of false positives. These images can be used conveniently without further manual filtering in the next steps of the SfM pipeline.

1.3.3 Feature detection, feature description and feature matching

While the conventional SfM workflow depicts feature detection, description and matching as separate steps, newly established methods merge two or all steps together by optimizing neural network structures (Dusmanu et al., 2019; Sarlin et al., 2020; Sun et al., 2021). Sometimes all three steps together are simply called *feature matching.* As this field is a main part of the thesis allowing a reconstruction of the camera pose, a comprehensive overview over the historical development of all three steps is given. Starting with Theorem 1.3.1 Ullmann (1979) defined, a need for the automatic determination of similar points in multiple images developed. This comes with the following questions:

1. How to find significant points (pixels, subpixels) in images?

2. How to describe these points distinctively?

3. How to match the points over multiple images?

1.3.3.1 Feature detection

Finding significant retrievable points in images is nowadays mostly called feature detection while in older publications the algorithm is denoted as interest operator or interest detector. In literature these points are known as interest points, image corners (corner points), keypoints (key points), or tiepoints. The most common term is features or feature points. The following part presents a timeline of the development of feature detectors with its most common principles and examples. Most of the algorithms target the processing of grayscale images because of the historical development and the simplicity of processing only one color channel.

The general idea for deriving distinctive points is to analyze the gray values and especially their gradients of the image and find patterns that are:

1. invariant to geometric and radiometric distortions
2. robust to occlusions, image noise, and image artifacts
3. uniquely describable and retrievable

About the same time as Ullmann published the SfM theorem, a simple interest operator was published that sums up gradients in four main directions of an image window with predefined size (Moravec, 1977). This approach was mainly developed to detect objects in front of a moving vehicle but can be seen as one of the first interest operators. The idea was enhanced by Förstner and Harris independently of each other who added sub-pixel accuracy (Förstner and Gülch, 1987) and rotational invariance using the autocorrelation function (Harris and Stephens, 1988). Both use non-maximum suppression to find the final corners in the respective image windows which is still a relevant method in modern computer vision tasks (Rothe et al., 2015). The Shi-Tomasi corner detector slightly changes the acceptance of relevant image windows enhancing the feature detection in consecutive frames (Shi and Tomasi, 1994). However, all these presented methods are not invariant to scale and affine transformations present in historical image pairs.

Relevant scale invariant detectors were introduced around the year 2000 using pyramid representations of the input image. The image is convolved multiple times and feature points are detected in scale space using Laplacian-of-Gaussian (LoG) (Lindeberg, 1998) or Difference-of-Gaussian (DoG) filters (Lowe, 1999). The detection of affine invariant features can be seen as the generalization of the scale invariant approach (Mikolajczyk and Schmid, 2004). Solving this problem is one of the most important steps for the pose estimation of any unordered dataset as image pairs are commonly affine or perspective transformed (Hartley and Zisserman, 2003). Thus, appropriate methods can overcome the issue of geometric distortions. The first applicable approaches use local intensity extrema in the image to find and describe image regions (Tuytelaars and Gool, 2000; Matas et al., 2002) while other works use the second moment matrix to estimate the shape of local image structures (Lindeberg, 1998; Baumberg, 2000; Mikolajczyk and Schmid, 2004).

With Maximally Stable Extremal Regions (MSER) Matas et al. (2002) propose the first method which is invariant to perspective transformations. Their approach uses the idea of region growing of local intensity extrema and outperformed various other methods in a comprehensive evaluation (Mikolajczyk et al., 2005). It has to be mentioned, that the Scale Invariant Feature Transform (SIFT) (Lowe, 2004) as the most prominent algorithm used in the whole process of feature matching, uses an efficient variant of DoG filters, but is even more known for the properties of its descriptor and the matching strategy. SIFT assigns every feature point with its image location, scale and orientation which is necessary for the calculation of its descriptor.

After the publication of SIFT and its status becoming a *gold standard*, many more methods were developed after 2004 trying to reach similar or better performance than SIFT using e.g., Hessian-based detectors (Bay et al., 2006; Alcantarilla et al., 2012), or several detectors in a row (Mishkin et al., 2015a). Others focused on improvements of the speed of feature detection (Rosten and Drummond, 2006; Lepetit and Fua, 2006; Leutenegger et al., 2011; Alcantarilla et al., 2013) or improvements in challenging situations like e.g. radiometric distortions (Li et al., 2018). Some approaches already experimented on learning feature detection e.g., by learning parameters settings for common methods (Winder and Brown, 2007) using Powell's method (Powell, 1964) or filtering keypoints by a learned classifier score (Strecha et al., 2009; Hartmann et al., 2014). An example for different feature detectors is given in Figure 1.3.

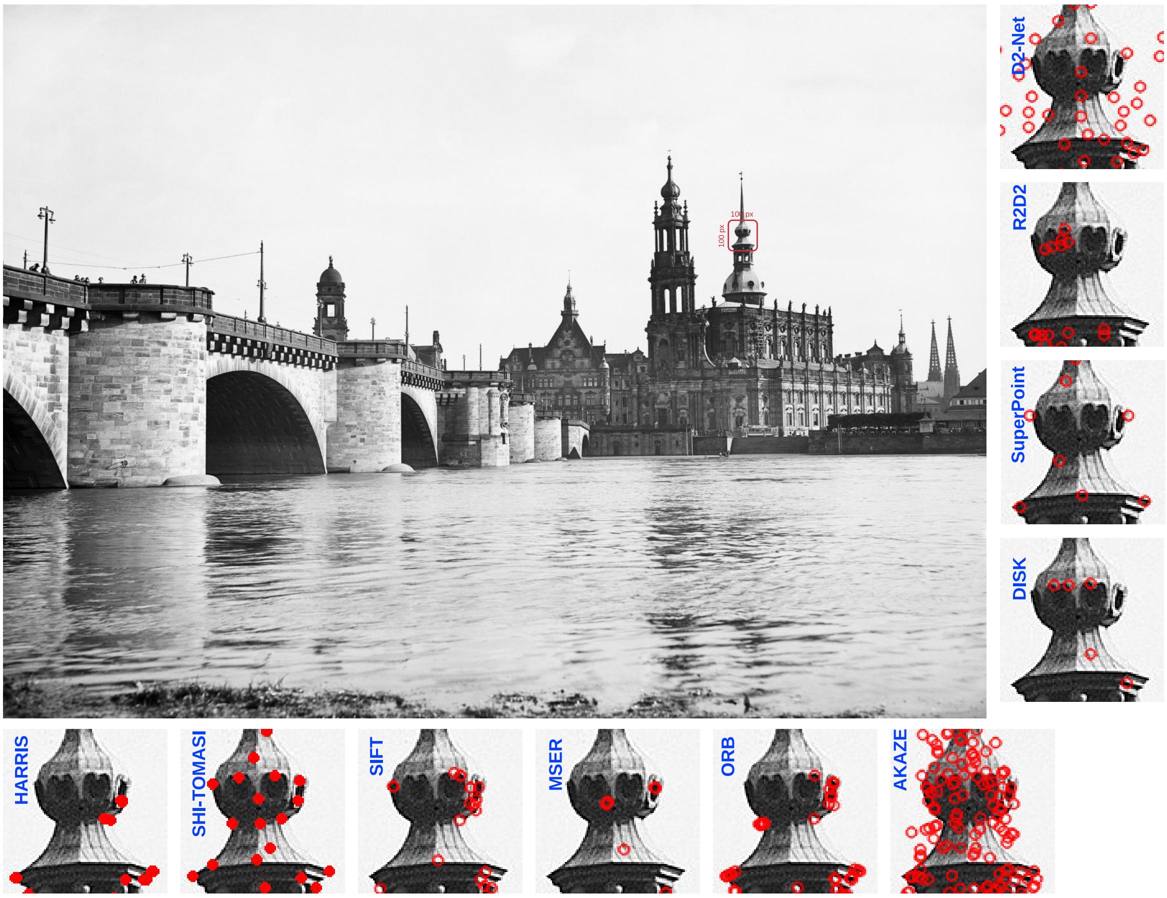


Figure (1.3): Example for different feature detection methods on a 100x100 px subpatch of an image (3543x2571 px) of the Hofkirche, Dresden. All methods use a maximum number of 4096 detected features (= standard value for SuperPoint and DISK) per image except for AKAZE and D2-Net where setting a limit is not possible. The bottom row shows algorithmic and the right column learned feature detectors. The number and position of detected tie points vary for all tested methods while the featureless sky in the image is almost never a region of interest.

With the advent of deep learning especially using Convolutional Neural Networks (CNNs) for image data, once again feature detection became the interest of different research. The detected features are often called learned features in opposite to algorithmic or "handcrafted" features. However, a significantly larger part of the research community focuses on the learning of the descriptors rather than the feature points. Therefore, many approaches use SIFT feature points and try to *describe* them distinctively using e.g., CNNs. Recent exceptions to this are e.g., Verdie et al. (2015), Savinov et al. (2017) and Laguna et al. (2019) which use neural networks exclusively for the detection of feature points. Hence, at the present time these approaches are overtaken by methods which jointly detect and describe feature points. This leads to the important step of feature description.

1.3.3.2 Feature description

Detected feature points need to be described unambiguously so that they can be found in different images. This is usually done by assigning a *descriptor* to every detected feature. Commonly, a descriptor holds information of the region around the feature point but there exist multiple variations.

While Shi and Tomasi (1994) show that the tracking of features using only few parameters is still difficult, several years later multiple works use local invariant descriptors to avoid ambiguities during feature matching (Pritchett and Zisserman, 1998; Lowe, 1999; Baumberg, 2000; Tuytelaars and Gool, 2000). These approaches use e.g., cross-correlation of intensity neighborhoods (Pritchett and Zisserman, 1998), or adaptive window-sizes (patches) with 20-40 invariants as descriptors (Baumberg, 2000). Others increase the descriptor vector from 160 (Lowe, 1999) up to 864 invariants (Matas et al., 2004) per feature point. Again, the most prominent example to mention is SIFT, which calculates orientation histograms of the image region around the keypoints and summarizes the result in a 128 element feature vector (Lowe, 2004). Until today, this is used as typical descriptor size.

Due to the structure of SIFT using different image sizes and calculating gradient magnitudes and orientations, it is invariant to scale and rotational changes. Additionally, it proved to be robust to viewpoint changes, affine distortions, noise and radiometric differences - also because of the newly proposed feature matching strategy (Lowe, 2004). The method brought forth many follow-up approaches optimizing properties of the original descriptor. Examples are Colored SIFT (CSIFT) (Abdel-Hakim and Farag, 2006), RootSIFT (Arandjelovic and Zisserman, 2012), Affine-SIFT (ASIFT) (Morel and Yu, 2009), Domain-Size Pooled SIFT (DSP-SIFT) (Dong and Soatto, 2015), or Scale-Less SIFT (Hassner et al., 2012). Other "handcrafted" descriptors that gained popularity are e.g., Speeded-Up Robust Features (SURF) which describes feature points using Haar-wavelets responses (Bay et al., 2006), or Accelerated KAZE (AKAZE) which describes the features in a nonlinear scale space (Alcantarilla et al., 2013). Another method that was used in this thesis due to its invariance to extreme radiometric distortions is the Radiation-Invariant Feature Transform (RIFT) (Li et al., 2018). The approach uses phase congruency maps to describe feature points in the frequency domain.

The descriptor vector typically consists of integer or floating point numbers but there exist also binary variants. These were mainly developed in order to speed-up the process of feature description and feature matching. As this is not of high relevance for historical images, binary descriptors shall not be part of this thesis. They are mainly used in applications like Simultaneous Localization and Mapping (SLAM) where processing time is a relevant issue (Opdenbosch et al., 2018). An example for a floating point descriptor is given in Figure 1.4.

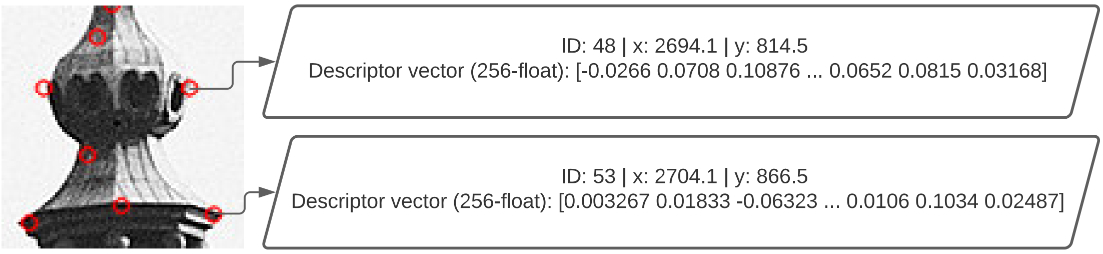

ID: 48 | x: 2694.1 | y: 814.5
Descriptor vector (256-float): [-0.0266 0.0708 0.10876 ... 0.0652 0.0815 0.03168]
ID: 53 | x: 2704.1 | y: 866.5
Descriptor vector (256-float): [0.003267 0.01833 -0.06323 ... 0.0106 0.1034 0.02487]

Figure (1.4): Example of the SuperPoint feature detector and descriptor (DeTone et al., 2018). Usually, the detected features are stored with a unique identifier, image coordinates (e.g., in pixels) and its descriptors simultaneously.

Just as with feature detection, deep learning brought new impulses for feature description. Methods that were developed before the breakthrough of neural networks which use earlier learning strategies are left out due to minor success.

One of the first approaches to learn keypoint descriptors was proposed by Jahrer et al. (2008). The network was trained on synthetically generated data to learn an optimized descriptor on DoG keypoints by minimizing the Euclidean distance (L2 distance) between the pairwise output. With a rising acceptance of neural networks for computer vision task, multiple methods were published to optimize the learning of keypoint descriptors. Some popular methods, which are still used today are e.g. DDesc (Simo-Serra et al., 2015), L2-Net (Tian et al., 2017), and HardNet (Mishchuk et al., 2017).

DDesc uses a Siamese CNN to learn discriminant patch representation by a 128-dimensional descriptor by using the L2 distance in training the network. Similarly to this approach, L2-Net also learns a 128-dimensional descriptor which can be matched by using the L2 distance but in contrast to DDesc, a progressive sampling strategy allows the training on billions of image samples. Additionally, a dedicated loss function is introduced for that approach. HardNet also outputs 128-dimensional descriptors and uses the L2-Net architecture with a novel loss. Further methods are described in detail in Chapter 5.

1.3.3.3 Feature matching

The process of feature matching is often mixed up or merged with the following step of geometric verification. This is based on earlier works in which geometric elements were used to estimate a local homography between similar image regions (Lowe, 1987; Venkateswar and Chellappa, 1995; Pritchett and Zisserman, 1998). These geometric elements (also feature groups) such as lines, vertices and edges can be seen as a predecessor to the descriptor vector. However, the term of feature matching is nowadays mainly used for the comparison and mapping of descriptor vectors in different images. The method produces so-called tentative correspondences (Matas et al., 2004) or putative matches (Jin et al., 2020) which are then used in the step of geometric verification (Section 1.3.3.4).

The procedure of finding similar descriptor vectors can be seen as a computational optimization problem. Theoretically, one could compare each descriptor vector in image 1 to each descriptor vector in image 2 and save the resulting feature point positions as a putative match if the vectors are equal. In fact, due to increasing computational power this is done in practice and called brute-force matching, exhaustive search, or exhaustive-matching strategy (Fig. 1.5).

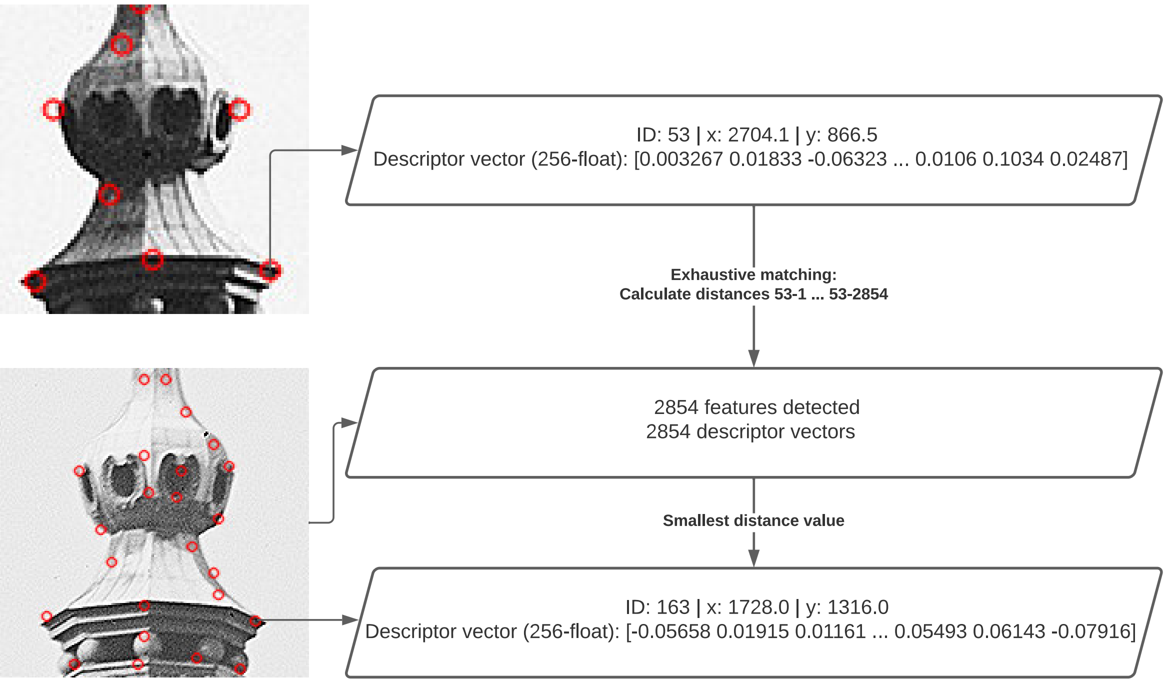

ID: 53 | x: 2704.1 | y: 866.5
Descriptor vector (256-float): [0.003267 0.01833 -0.06323 ... 0.0106 0.1034 0.02487]
Exhaustive matching:
Calculate distances 53-1 ... 53-2854
2854 features detected
2854 descriptor vectors
Smallest distance value
ID: 163 | x: 1728.0 | y: 1316.0
Descriptor vector (256-float): [-0.05658 0.01915 0.01161 ... 0.05493 0.06143 -0.07916]

Figure (1.5): Example of the SuperPoint feature detector and descriptor result in two images. Feature point 53 in image 1 is exhaustively matched with all 2854 feature points detected in image 2. Feature point 163 in image 2 is the nearest neighbor with the smallest descriptor distance value to feature point 53.

However, this approach raises the following questions:

1. Must descriptor vectors be exactly equal to produce a match?
2. What is the best and most efficient measure of equality?
3. What happens if multiple descriptor vectors are equal in one image?
4. Is there a less computationally intensive approach?

Conventional algorithms do not search for exactly the same descriptor vector because of noise and appearance of the feature point and thereby also the corresponding decriptor vector. Instead, all approaches look for the descriptor vector with the most common elements - the so-called Nearest Neighbor (NN).

The NN problem is defined by Beyer et al. (1999) as:

Theorem. *Given a collection of data points and a query point in an m-dimensional metric space, find the data point that is closest to the query point.*

An efficient method solving this problem in low-dimensional space uses k-d trees (Friedman et al., 1977). Though, it becomes inefficient in larger dimensions, e.g., for a 128-dimensional descriptor vector. An alternative and more efficient solution of the problem is provided by Muja and Lowe (2009) who create a hierarchical k-means tree using a-priori k-means clustering. While this approach speeds up the matching of high-dimensional vectors by several orders of magnitudes, it comes with slightly lower precision as not every descriptor vector is traversed (Muja and Lowe, 2009). In practice, both methods are used depending on the application. Benchmarks use an exhaustive search, while real-time applications rely on the Fast Library for Approximate Nearest Neighbors (FLANN) (Muja and Lowe, 2009).

Comparing descriptor vectors using nearest neighbor search requires a distance measure. The standard approaches use the L2 Norm (Euclidean Distance) for non-binary descriptors and the Hamming distance for binary descriptors. However, there exist works that use different measures like e.g. L1 Norm (Manhattan Distance), Hellinger Distance (Arandjelovic and Zisserman, 2012), χ^2 distance (Zhang et al., 2006), or Earth Mover's distance (Rabin et al., 2009).

Up to this point every descriptor vector in image 1 is matched to another descriptor vector in image 2 with no termination criterion. This assumes that every feature point that was found in image 1 could also be found in image 2 with almost similar descriptor vectors which is obviously never the case for two different real world image pairs.

The most common method to filter matches derived by the nearest neighbor search can once again be found in Lowe (2004) and is called Lowe's ratio test. The idea is to use the distance ratio as in equation 1.1,

$$\text{distance ratio} = distance(d_{im1}, d_{im2-NN})/distance(d_{im1}, d_{im2-secNN}) \tag{1.1}$$

where d_{im1} represents a descriptor vector d of a feature in image 1,
d_{im2-NN} is the NN in image 2,
$d_{im-secNN}$ the second NN of d_{im1} in image 2,
and $distance$ represents the chosen distance measure (e.g., L2 Norm).

If this distance ratio is larger than a certain threshold (distance ratio > distance threshold) the feature match gets rejected (Fig. 1.6).

Lowe proposes a threshold of 0.8 while other feature matching methods define and use different thresholds (Dusmanu et al., 2019; Tyszkiewicz et al., 2020) (Chapter 4). A comprehensive evaluation of the distance threshold for different methods is done by Jin et al. (2020).

Enhancing the robustness of Lowe's ratio test multiple post-processing steps can be applied. If the matches from image 1 to image 2 are regarded as a set $m_{1->2}$ there is also the option to calculate in the opposite direction using the similar matching method. The resulting set $m_{2->1}$ can be compared with the first set. It is a common approach to intersect both lists $m_{1->2} \cap m_{2->1}$ and keep the result (cf. Jin et al. (2020)).

This process is known as mutual nearest neighbor search, bipartite matching or symmetry test. Finally, there is the option to neglect matches whose descriptor distance is above a certain threshold. Usually after applying multiple of the shown filtering methods, the resulting putative matches are used and evaluated in the next step of geometric verification.

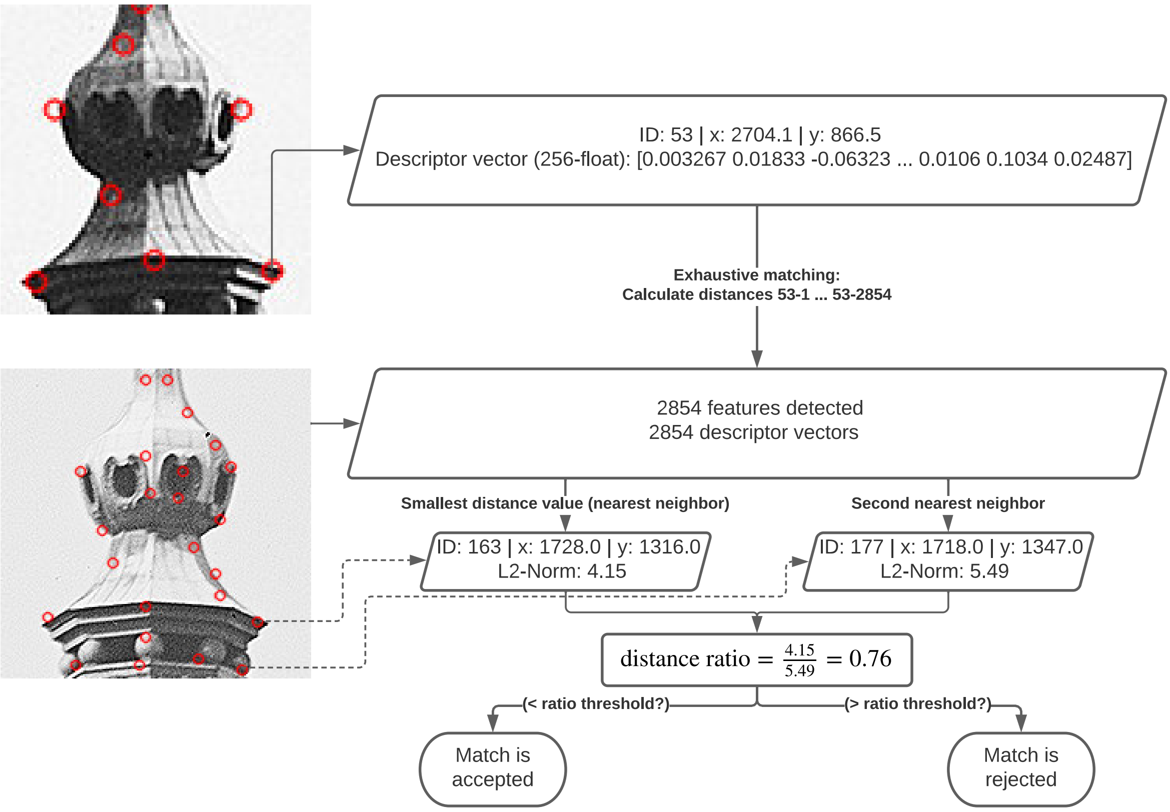

ID: 53 | x: 2704.1 | y: 866.5
Descriptor vector (256-float): [0.003267 0.01833 -0.06323 ... 0.0106 0.1034 0.02487]
Exhaustive matching:
Calculate distances 53-1 ... 53-2854
2854 features detected
2854 descriptor vectors
Smallest distance value (nearest neighbor)
Second nearest neighbor
ID: 163 | x: 1728.0 | y: 1316.0
L2-Norm: 4.15
ID: 177 | x: 1718.0 | y: 1347.0
L2-Norm: 5.49
distance ratio = $\frac{4.15}{5.49}$ = 0.76
(< ratio threshold?)
(> ratio threshold?)
Match is accepted
Match is rejected

Figure (1.6): Example of the SuperPoint result already shown in Figure 1.4. Feature point 163 is the nearest neighbor and feature point 177 the second nearest neighbor of feature point 53. For both descriptor vectors the L2-Norm is calculated. The result is used to compute the distance ratio. For a threshold of 0.8 the match would be considered good and accepted.

1.3.3.4 Geometric verification and robust estimators

The principle that feature matches cannot only be verified by their descriptor vector but also due to the geometric relationship between the correspondences is part of the field of epipolar geometry (Hartley and Zisserman, 2003). This section aims to give an overview and the mathematical foundation over two-view and three-view geometry and the associated Computer Vision terms *Homography*, *Fundamental Matrix* and *Trifocal Tensor* which are used for filtering the putative matches.

Assuming a pinhole camera model, a Homography describes the mapping of a planar surface seen in a first image to the same planar surface in a second image (Hartley and Zisserman, 2003). The mapping is defined as a (Homography) 3×3 matrix H and can be estimated using a minimum of four point correspondences $x \leftrightarrow x'$, where $x = (x, y, 1)^T$ represents a point in the first image and $x' = (x, y, 1)^T$ the corresponding point in the second image (Chum et al., 2005b). The equation for determination of H can be written as

$$x' = Hx \tag{1.2}$$

and in the following be expressed as

$$x' \times Hx = 0 \tag{1.3}$$

This enables a linear solution using least-squares techniques in the case of an over-determined set of equations and if exact four point correspondences are available Gaussian elimination can be used (Hartley and Zisserman, 2003). While this shows just the base of Homography estimation, there exist various optimizations (Chum et al., 2005b; Jawahar et al., 2006) for the determination of H.

As the Homography matrix applies only to planar surfaces, a more general solution for mapping point correspondences is the 3×3 Fundamental Matrix F. Derived by equations of epipolar geometry, for the Fundamental Matrix a similar equation as equation 1.2 exists. The Fundamental Matrix satisfies the condition that for any pair of corresponding points $x \leftrightarrow x'$ in the two images

$$x'^T F x = 0. \tag{1.4}$$

This equation is a central part of the thesis as it maps regular point correspondences in the case of *uncalibrated* cameras. In general, the equation can be solved using the linear eight-point algorithm (Longuet-Higgins, 1981; Hartley, 1997a), while the minimal configuration strictly requires only seven correspondences (Stewart, 1999).

For benchmark datasets most creators usually provide the Homography or the Fundamental Matrix for all image pairs. While this has mostly practical reasons, as corresponding points only have to be found in image pairs, there exists the possibility of degenerate configurations (Maybank, 1990; Hartley and Zisserman, 2003; Ressl, 2003). The use of the Trifocal Tensor, i.e. the geometrical relationship between three uncalibrated images, allows the robust unique determination of the relative orientation of three images, if the corresponding image points are not from a common plane in space (Ressl, 2003). The $3 \times 3 \times 3$ Trifocal Tensor T derived from point correspondences is defined as

$$[x']_\times (\sum_i x^i T_i)[x'']_\times = 0_{3\times 3} \tag{1.5}$$

where x, x', x'' are the image coordinates in the three images,
$[x]_\times$ is the 3×3 skew-symmetric matrix of the 3-vector,
i is the number of 3×3 Tensor slice,
and $0_{3\times 3}$ is the 3×3 null matrix.

While it is generally possible to determine T using six point correspondences, Ressl (2003) propose a minimum of 15 correspondences for robust estimation. The resulting Trifocal Tensor allows using point transfer to test feature matches and the direct derivation of correct Fundamental Matrices. A schematic representation of H, F and T is given in Figure 1.7. More details on the determination of T and the resulting dataset are found in Chapter 2.

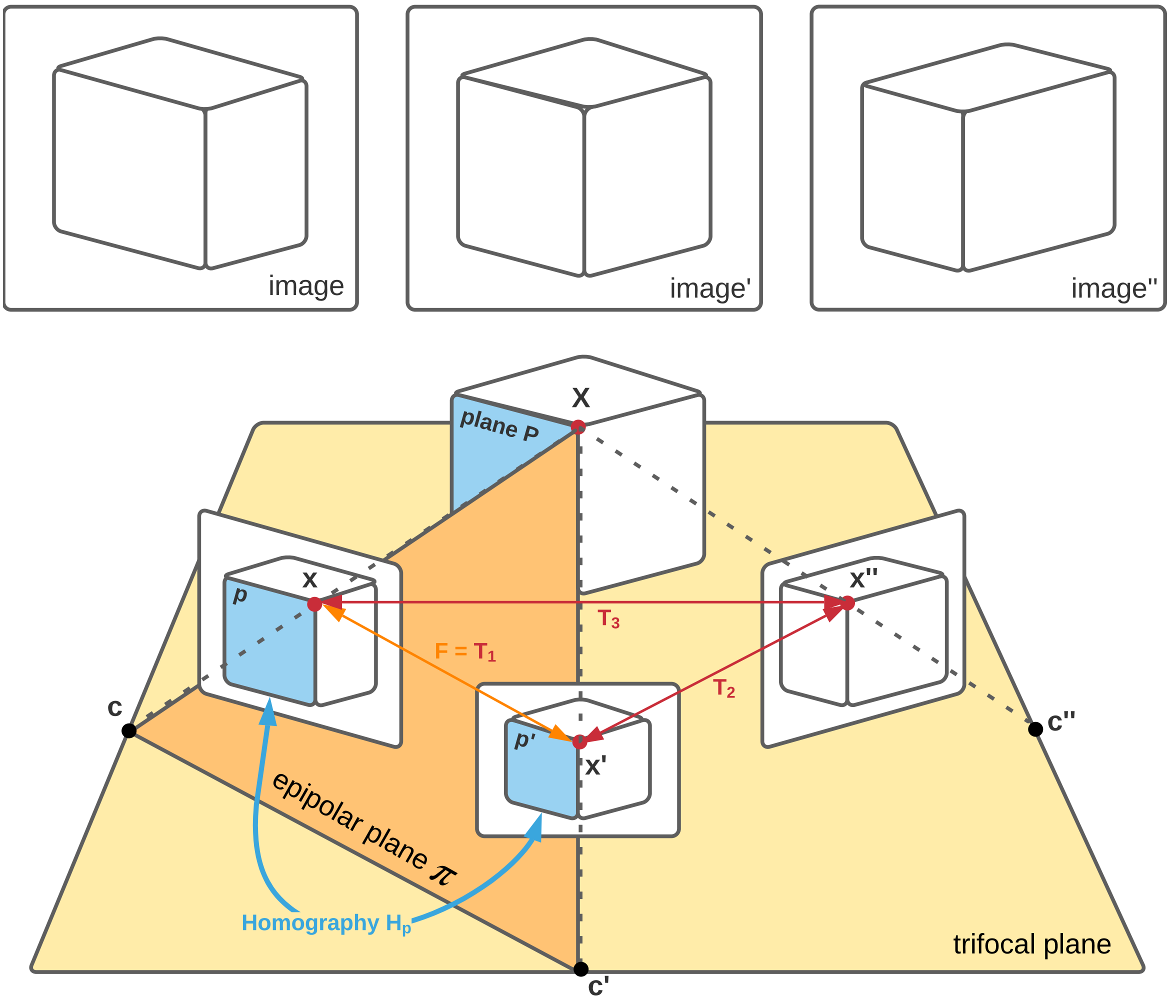


Figure (1.7): Schematic visualization of the epipolar geometry of three views. Object point X is seen by three cameras with projection centers c, c' and c'' as image point x, x' and x''. The Fundamental Matrix F describes the mapping between point x and x' lying on the epipolar plane π defined by the camera centers c and c'. Similarly, the Homography H describes the mapping between plane p and p'. The Trifocal Tensor T defines the point-point-point correspondence between x, x' and x''. For a better illustration, the depiction of epipoles which lie in the trifocal plane is omitted.

Common feature detection and matching methods do not only provide the minimum of corresponding points solving the equations 1.2, 1.4 and 1.5 above, but usually the equation system is well over-determined. This requires the use of algebraic solvers e.g., least-squares estimation with additional methods for removing coarse outliers in the putative matches. Again, the most common and widely used robust estimators for image pair matching are discussed.

With the publication of the Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981), this method became the standard for estimating the Homography and the Fundamental Matrix (Torr et al., 1995; Hartley and Zisserman, 2003; Snavely et al., 2006; Szeliski, 2010; Raguram et al., 2013). RANSAC is able to fit a mathematical model (i.e., H, F, T) to experimental data (the putative matches), while being able to deal with a significant percentage of outliers (gross errors). For example, in the case of estimating F, a random sample of eight putative matches is chosen and F is calculated. After that, every other putative match is tested and e.g., the algebraic distance $d = x'Fx$ can be calculated. If d is smaller than a threshold t for a pre-defined percentage of inliers the estimated Fundamental Matrix is kept.

Optimizations of RANSAC are Maximum Likelihood Estimation Sample Consensus (MLESAC) (Torr and Zisserman, 2000), Locally-Optimized RANSAC (LO-RANSAC) (Chum et al., 2003), Progressive Sampling Consensus (PROSAC) (Chum and Matas, 2005), Degeneracy Sample Consensus (DEGENSAC) (Chum et al., 2005a), and Marginalizing Sample Consensus (MAGSAC) (Barath et al., 2019). These methods adapt RANSAC by introducing different cost functions (Torr and Zisserman, 2000), adding more data points to the final model (Chum et al., 2003), using the matching order derived from descriptor matching (Chum and Matas, 2005), and eliminating the user input for threshold t (Barath et al., 2019). DEGENSAC is useful to eliminate degenerate configurations when estimating F, i.e. all putative matches lie on a dominant plane in the scene (Chum et al., 2005a). This is especially useful when working with images showing façades and practically eliminates the use of the Trifocal Tensor.

Recent evaluations have shown that the use of an appropriate robust estimator enhances the number of inliers and the quality of the estimated Fundamental Matrix (Barath et al., 2020; Jin et al., 2020).

1.3.3.5 Joint methods

Through the advent of deep learning, the trend in feature matching moves to complete pipelines including feature detection, feature description, matching, and geometric verification. This is made possible due to smart network architectures combining different types of neural networks in order to optimize the different steps. Another important point is the availability of ground truth data for e.g. labeled data (Deng et al., 2009), Homographies (Mikolajczyk et al., 2005), image patches (Balntas et al., 2017), depth maps (Li and Snavely, 2018), and even reconstructed 3D scenes (Snavely et al., 2006; Sattler et al., 2018). This data enables training of networks in order to fulfill specific tasks and generate the desired outputs.

An overview and explanation for many different methods is given in Chapter 4 and 5. The performance of these methods on contemporary images is mainly evaluated in the Image Matching Challenge of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) while this thesis' focus lies on the years of 2019-2021. Further comprehensive evaluations can be found in Csurka et al. (2018) and Jin et al. (2020).

Most of these approaches combine the feature detection and feature description step, in so-called detect-and-describe approaches (Yi et al., 2016; DeTone et al., 2018; Ono et al., 2018; Revaud et al., 2019; Dusmanu et al., 2019). For historical images, these methods were the first which clearly outperformed conventional algorithmic pipelines in the number of correct matches (Chapter 4). Further works combine only the description and matching step (Han et al., 2015; Zagoruyko and Komodakis, 2017) but were quickly outperformed by other approaches regarding urban image data (Yi et al., 2016).

The most recent advances were made by the SuperGlue (Sarlin et al., 2020) and Discrete Keypoints (DISK) (Tyszkiewicz et al., 2020) pipeline. Due to their promising results on very diverse historical image collections (Chapter 5), these methods are extensively used on various historical datasets and are thus explained in detail. Sarlin et al. (2020) describe the development of a Graph Neural Network exclusively for feature matching. Hence, they combine this matching strategy with SuperPoint (DeTone et al., 2018) for feature detection and description, presenting a complete end-to-end pipeline (Fig. 1.8).

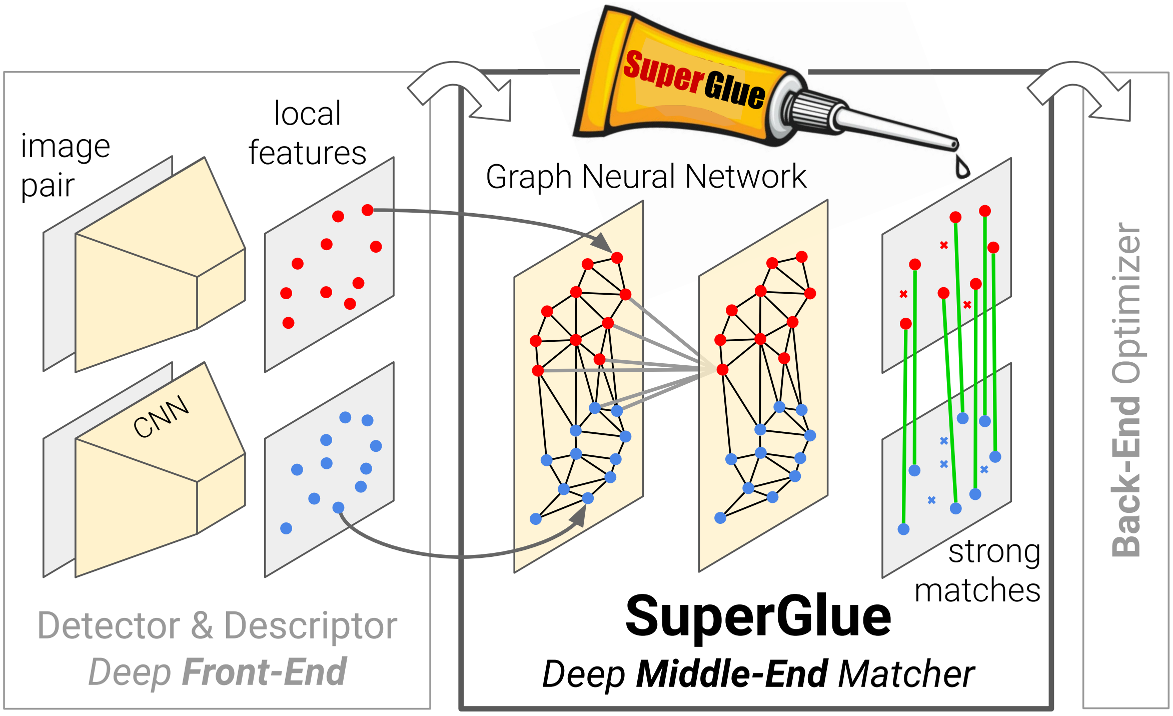


Figure (1.8): The SuperGlue approach establishes pointwise correspondences from off-the-shelf local features: it acts as a middle-end between handcrafted or learned front-end and back-end.[1]

SuperPoint is able to operate on full-sized images and reduces the dimensionality of the input image using a VGG16 encoder (Simonyan and Zisserman, 2014). The output is used in two different decoders (one for feature detection, one for description) which learn specific weights. During this process, most learned parameters are shared between the decoders what makes this a detect-and-describe approach.

As a prior to this network structure SuperPoint uses Homographic Adaption to enable self-supervised training of feature detectors. Homographic Adapation describes the change of the input image by using multiple known homographies synthetically changing viewpoint and scale. The advantages of the full pipeline are the simultaneous detection and description of features in a single network for a full resolution image in real-time (DeTone et al., 2018). Additionally, the use of Homographic Adaptation avoids the need for further training data. The keypoint positions p and their descriptors d with a size of 256 can be directly used in the SuperGlue matching.

SuperGlue solves an optimization problem whose cost is predicted by a deep neural network trained on real data. The network consists of two major components - the attentional graph neural network and the optimal matching layer (Fig. 1.9).

[1] Figure and part of the caption are taken from the publication of Sarlin et al. (2020) with the kind permission of Paul-Edouard Sarlin.

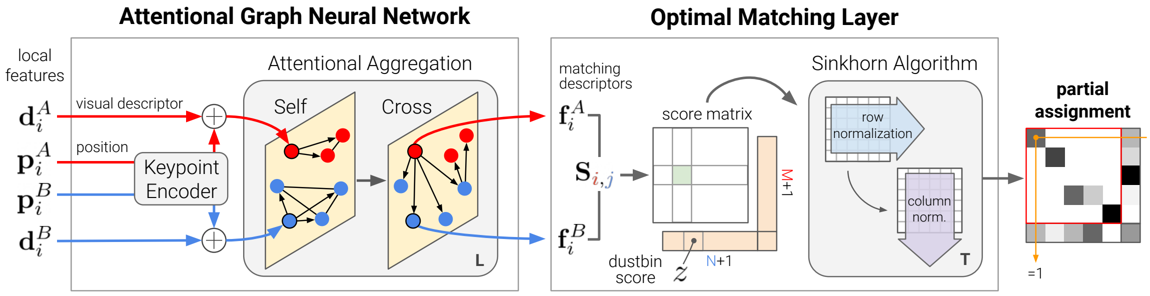


Figure (1.9): SuperGlue is made up of two major components: the *attentional graph neural network*, and the *optimal matching layer*. The first component uses a *keypoint encoder* to map keypoint positions p and their visual descriptors d into a single vector, and then uses alternating self- and cross-attention layers (repeated L times) to create more powerful representations f. The optimal matching layer creates an M by N score matrix, augments it with dustbins, then finds the optimal partial assignment using the Sinkhorn algorithm (for T iterations).[2]

The attentional graph neural network builds dynamic graphs between keypoints. Sarlin et al. (2020) divide between self-attention where the network builds a graph between keypoints in the same image, and cross-attention where a graph is built between keypoints in two images. These enables the network to focus on subsets of keypoints in multiple regions, "gluing" them together and create more powerful representation of the original descriptors d.

The optimal matching layer can be seen as a graph matching problem (Section 1.3.2) where a score matrix for all possible matches between two images is created. Because some keypoints from the first image will not be visible in the second image, Sarlin et al. (2020) use so-called dustbins to deal with e.g., occlusion. The complete problem is formulated as an optimization problem which can be solved by using the Sinkhorn algorithm (Sinkhorn and Knopp, 1967).

DISK also learns local features in an end-to-end pipeline. Therefore, Tyszkiewicz et al. (2020) use reinforcement learning. In a first step, DISK uses a U-Net (Ronneberger et al., 2015) that takes images as inputs and outputs keypoint heatmaps and dense descriptors. The feature extraction step is formulated in a probabilistic framework and is similar to SuperPoint (DeTone et al., 2018). The complete heatmap (also referred to as feature map) is then subdivided into a grid with cell size $h \times h$ and for every cell at most one feature point is selected. The feature points are associated with the corresponding descriptors at these exact locations.

All feature points are accumulated in feature sets with their corresponding descriptors which allows the calculation of the L2 distance between every descriptor combination resulting in a distance matrix. In order to match the feature points, Tyszkiewicz et al. (2020) try to optimize the ratio test (eq. 1.1) and the mutual nearest neighbor search (Section 1.3.3.3) in their neural network. While their goal is to produce more feature matches than the conventional approaches, they relax the mutual nearest neighbor constraint by using the distribution of elements in the distance matrix. Geometric ground truth is used to assign positive/negative rewards to each match and allows using a policy gradient method for training (Williams, 1992).

SuperGlue and DISK pipelines are able to reliably match difficult historical image pairs with a large number of correct feature matches and perform comparable with regards to

[2] Figure and part of the caption are taken from the publication of Sarlin et al. (2020) with the kind permission of Paul-Edouard Sarlin.

the specific dataset. SuperGlue produces more correct matches for challenging image pairs (Chapter 5). SuperGlue (+SuperPoint) brought an enormous quality improvement for all kind of pair-wise image data in the Image Matching Challenge 2020, outperforming the final SIFT pipeline from 2004 by an order of magnitude. Tests on the historical dataset developed in Chapter 2 come to the same conclusion (Fig. 1.10). The final test setup is explained in Appendix 8.1 in detail.

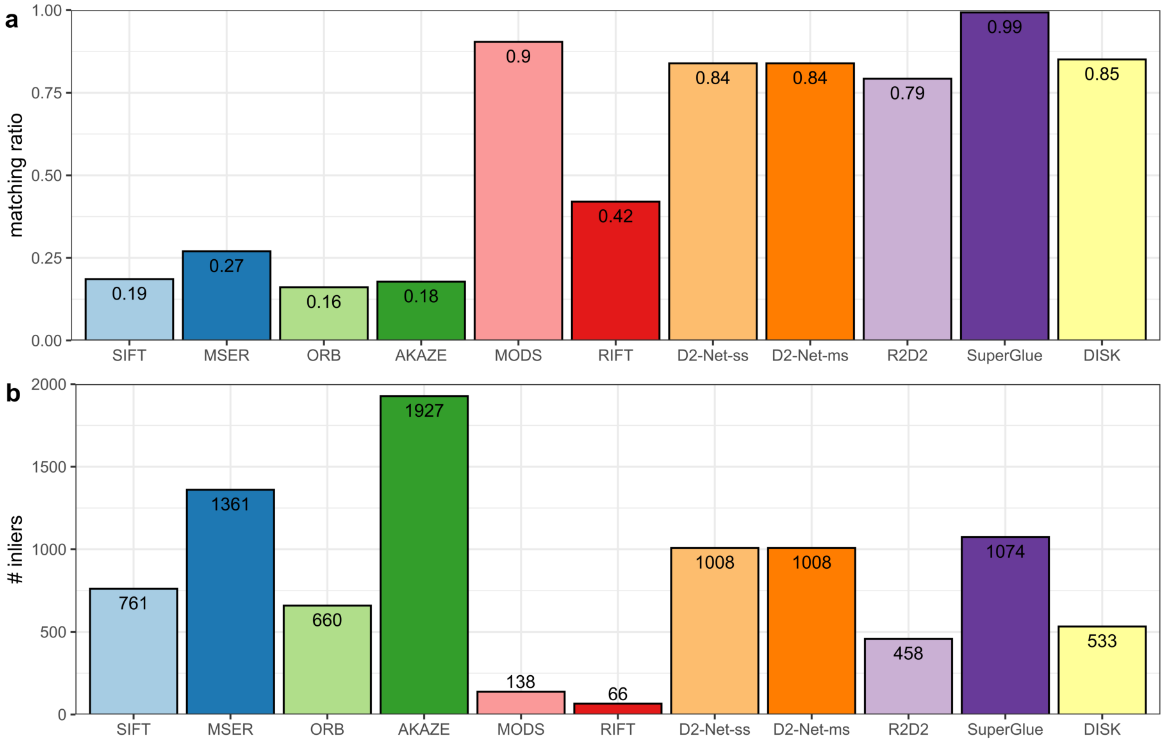


Figure (1.10): Mean of matching ratio (a) and number of inliers (b) tested on the historical benchmark dataset of Chapter 2 using different feature matching methods without geometric verification. The methods are sorted by publication year. SuperGlue and MODS produce the highest matching ratio while SuperGlue also provides a high number of correct matches (=inliers). All methods were limited to a maximum number of 4096 features detected, except for MSER and AKAZE where this is not possible.

The recent Image Matching Challenge 2021 only reported marginal improvements by changing small parts of the original SuperPoint, SuperGlue, and DISK.

1.3.4 Initial parameterization

The described joint methods for feature matching provide a set of images with their corresponding feature matches. In the following, each single image has to be linked to a camera for which the interior orientation has to be initialized. In order to optimize the alignment and orientation parameters of all cameras, bundle adjustment is performed. As bundle adjustment is well researched and implemented in various applications, it is reasonable to use an already optimized version available in every SfM software.

There are two main types of available SfM tools - proprietary and open-source software. Among the most frequently used proprietary tools are Agisoft Metashape, Pix4D, 3DF

Zephyr, PhotoModeler, and Reality Capture. All these tools allow a completely automatic generation of sparse point clouds, hence, an intervention into the workflow is not or only up to a certain degree possible. This includes especially the integration of pre-calculated feature matches and thus, in addition to their pricing, the usage is not feasible for historical image pairs. Especially, the workflow of Agisoft Metashape has been extensively tested on historical images, though it did only align a small number of these images (Chapter 3 and 4).

Open-source tools (mainly initially developed in research projects) like e.g., bundler (Snavely et al., 2006), VisualSFM (Wu, 2013), MicMac (Rupnik et al., 2017), Meshroom (AliceVision, 2018), and COLMAP (Schönberger and Frahm, 2016) allow this integration and mostly offer the tuning of further parameter settings. While the standard reconstruction workflows do not provide sufficiently good results for historical images (Chapter 3 and 4), the integration of joint feature matching methods finally allow the accurate reconstruction of multiple historical images (Chapter 4 and 5).

For processing most of the historical images the software COLMAP is chosen as it comes with several advantages:

1. completely available and documented in English language
2. open-source
3. in active development with the option to contribute
4. running on multiple operating systems (Windows, Linux, Mac)
5. database management
6. controllable via source-code, command-line, and graphical user interface

The software is mainly used for dataset creation, benchmarks, and evaluations in the Computer Vision and Photogrammetry community (Maddern et al., 2016; Remondino et al., 2017; Li and Snavely, 2018; Csurka et al., 2018; Jin et al., 2020).

When importing an image, COLMAP requires the user to define the camera type and initial parameters for the interior orientation necessary for performing bundle adjustment. Available camera types are i.a., SIMPLE_PINHOLE, SIMPLE_RADIAL, OPENCV, FULL_ OPENCV, or THIN_PRISM_FISHEYE. These models increase in complexity and provide the option to pre-calibrate and model various camera types. As the camera type is usually unknown for historical images, it is advisable to use a simpler model. Albl et al. (2016) and Polic et al. (2020) show that the calibration of many camera parameters offers too many degrees of freedom to be reliable for historical images. Experiments in Metashape performed throughout the thesis reached the same conclusions.

Thus, the initial camera type is set to SIMPLE_RADIAL, which follows the recommendation proposed by Schönberger and Frahm (2016) for unordered image collections. In addition, digitized historical images usually don't visually show a significant amount of distortion. The SIMPLE_RADIAL model requires the initialization of f, cx, cy, k.

f is mostly called focal length in the Computer Vision community, which can be seen as rather imprecise, as the term is closely related to an optical lens. A better definition is given by using the term camera constant (commonly used in the german-speaking area) or principal distance c. The principal distance as described by Luhmann et al. (2019), 3.3.2.3 is defined as follows

Definition. *"Perpendicular distance to the perspective center from the image plane in the negative z' direction. When focused at infinity, c is approximately equal to the focal length of the lens ($c \approx f'$)."*

As the digitized copy of an analogue photograph does not directly relate to an optical lens, the term principal distance with the COLMAP annotation f is used in the following. Unordered historical images are usually not digitized by accurate photogrammetric scanners which bears the following theorem:

Theorem. *For historical images the principal distance f is not uniquely determinable, since the digital copy does not mathematically relate to a physical lens anymore.*

In other words: Since the original field of view (FOV) is directly correlated to f, the photographer could theoretically have been very close to the object with a wide FOV (= small f, wide-angle lens) or very far away with a narrow FOV (= large f, telephoto lens). This is represented by equation 1.6 given in Luhmann et al. (2019), 3.4.3.2 (p. 224):

$$\tan\Omega = \frac{s'}{2f} \tag{1.6}$$

,

where 2Ω corresponds to the FOV,
s' is the diagonal of the given image format,
and f is the principal distance.

While the image format is known in pixels for the digital copy, the original sensor size is mostly unknown. Both other parameters are indeterminable as well but slight assumptions about the FOV can be made. As an empirical value, extreme large or extreme narrow FOVs can be excluded because of their rare appearance in historical photography. For instance, the first fisheye lens by Nikon was available for public photographers in 1957 (Stafford, 2004).

COLMAP follows the above assumptions by automatically setting the initial value of f (in pixels) to $1.25 \cdot max(image_{width}, image_{height})$ which corresponds to a common FOV for consumer-grade cameras. Luhmann et al. (2019), 3.4.3.2 (p. 225) proposes a rule of thumb using the diagonal of the image format as an approximation of the focal length. An example for both approaches on two common images of the Deutsche Fotothek is given in equation (1.7).

$$\begin{aligned}
&\text{Image 1:}\quad \text{width=1600 pixels, height=974 pixels}\\
&f^1_{COLMAP} = 1.25 \cdot 1600 \text{ pixels} = 2000 \text{ pixels}\\
&f^1_{Luhmann} = \sqrt{1600 \text{ pixels}^2 + 974 \text{ pixels}^2} = 1873 \text{ pixels}\\
&\text{Image 2:}\quad \text{width=1212 pixels, height=1600 pixels}\\
&f^2_{COLMAP} = 1.25 \cdot 1600 \text{ pixels} = 2000 \text{ pixels}\\
&f^2_{Luhmann} = \sqrt{1212 \text{ pixels}^2 + 1600 \text{ pixels}^2} = 2007 \text{ pixels.}
\end{aligned} \tag{1.7}$$

,

In the subsequent bundle adjustment COLMAP does not limit the range for the calibration of f but cameras with an abnormal FOV (standard values are $0.10 \cdot f$ and $10 \cdot f$ times the initial) are filtered (Schönberger and Frahm, 2016).

There are several other options for initially estimating the principal distance. Recent methods try to estimate the interior orientation of the camera by using neural networks but are specialized on wide-angle lenses (Bogdan et al., 2018) or do not provide an implementation for evaluating own images (Workman et al., 2015). Others rely on few known distances (Xiong et al., 2021), coplanar circles (Chen et al., 2004) or concentric circles (Jiang and Quan, 2005) in object space. These features and measures are commonly unavailable in historical images, so the presented approach uses vanishing points (VPs) more common in architectural scenes in order to estimate the principal distance.

Li et al. (2010) rely on prior work (Barnard, 1983; Caprile and Torre, 1990; Cipolla et al., 1999) and estimate three orthogonal VPs in a single image as seen in Figure 1.11.

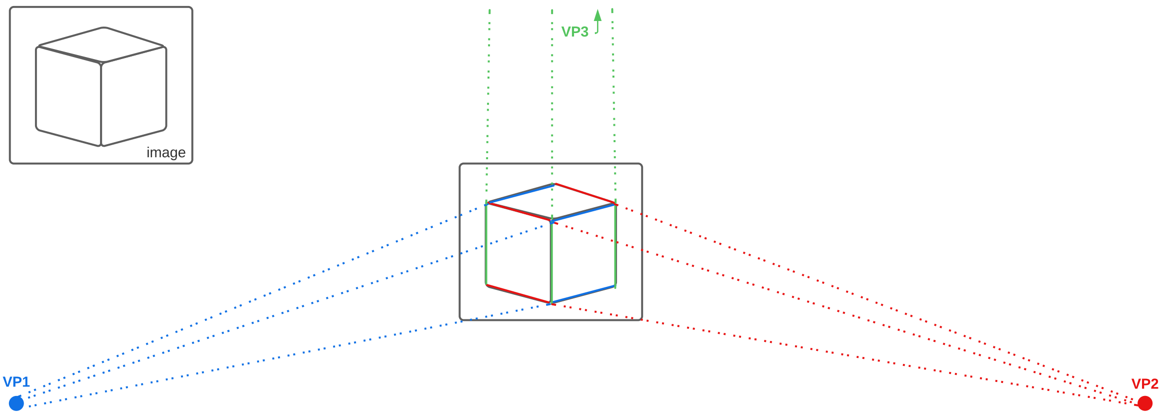


Figure (1.11): Schematic example for the detection of three VPs. VP3 is not visible in the image as it lies on the intersection of the green dotted lines.

With three VPs it is possible to estimate the principal distance (Barnard, 1983). The approach of Li et al. (2010) detects line segments and in the following all possible intersections of these line segments are calculated. Then, the intersections are plotted in a polar coordinate system with the origin at the image center. This allows the accumulation of polar angle θ of all intersection in a histogram and thus, the selection of three VPs.

While this method *always* detects three VPs, a termination criterion for historical images has to be defined, considering that not all images (especially façades) show exactly three orthogonal VPs (Chapter 4). This method becomes especially useful if there are a lot of images with three main directions. Hence, recent experiments show that for most situations the initial approximation of f in COLMAP without using VPs is good enough for the bundle adjustment to converge.

The further interior parameters cx and cy represent the principal point's coordinates. The principal point $c(x, y)$ as described by Luhmann et al. (2019) is defined as

Definition. *"foot of perpendicular from perspective centre to image plane, with image coordinates (cx, cy). For commonly used cameras approximately equal to the centre of the image [...]."*

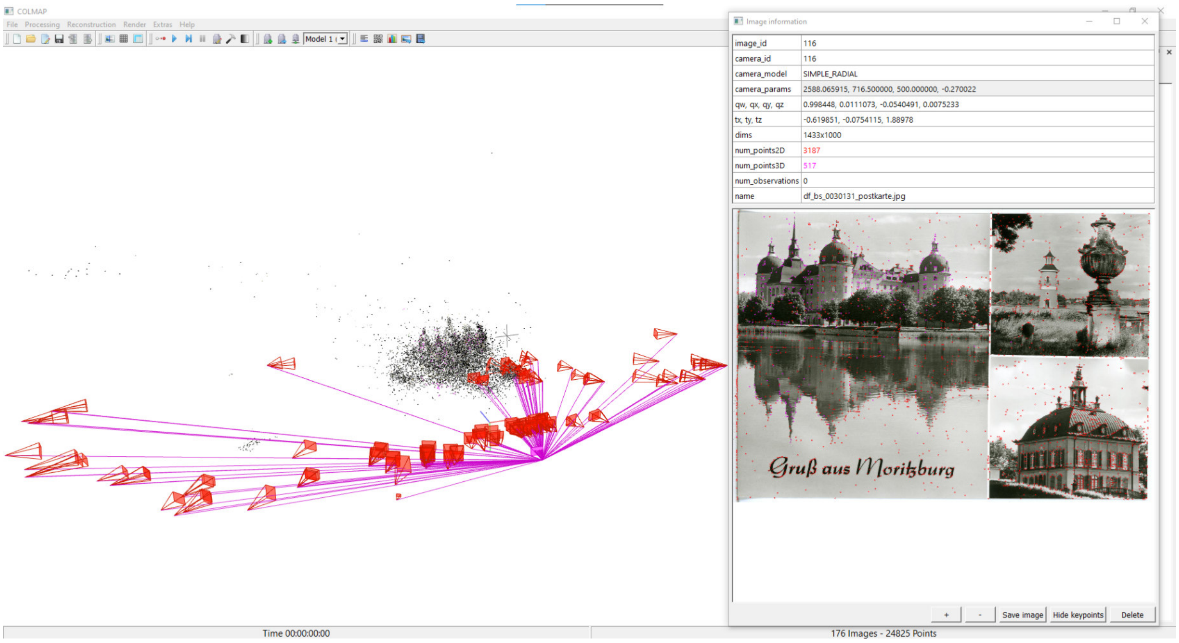
As with the principal distance, the principal point cannot be uniquely determined because the lens used mostly cannot be recovered. In real cameras the principal point is different from the image center (Luhmann et al., 2019), but as common initialization, most SfM software assume that it coincides with the image center (Snavely et al., 2006; Schönberger and Frahm, 2016).

The image coordinate system in COLMAP is defined with the origin as the top-left corner of the image with the x-axis pointing right and the y-axis towards bottom. Thus, the initial principal point's coordinates are set to $cx = 0.5 \cdot image_{width}, cy = 0.5 \cdot image_{height}$ where height and width are defined in pixels. The bundle adjustment of COLMAP does *not* refine cx and cy by default, because principal point calibration is an ill-posed problem (Schönberger and Frahm, 2016). It has been investigated that allowing cx and cy to vary in the bundle adjustment, the simultaneous estimation of f decreases in accuracy (Agapito et al., 1998). Fixing the principal point's coordinates is able to negate this effect.

Nonetheless, it has to be stated that the assumption that $c(x, y)$ lies in the image center is a critical point considering the digitization process of the analogue original. This is mainly due to two reasons:

1. If the analogue original is cropped, tilted, or shifted during digitization, $c(x, y)$ will not be near the image center.

2. Some of the original analogue photographs are framed as these were sold as postcards and redigitized, generating similar inaccuracies (Fig. 1.12).



image_id	116
camera_id	116
camera_model	SIMPLE_RADIAL
camera_params	2588.065915, 716.500000, 500.000000, -0.270022
qw, qx, qy, qz	0.998448, 0.0111073, -0.0540491, 0.0075233
tx, ty, tz	-0.619851, -0.0754115, 1.88978
dims	1433x1000
num_points2D	3187
num_points3D	517
num_observations	0
name	df_bs_0030131_postkarte.jpg

Figure (1.12): A reconstructed model of Castle Moritzburg using historical images. The overlaid photograph (camera_params in order f, cx, cy, k) with ID 116 clearly comes with a principal point that is not near the image center as Castle Moritzburg is only depicted in the left side of the postcard (which is a collage of three images). Nonetheless, the pose is reconstructed by COLMAP assuming that $c(x, y)$ lie in the image center. In order to fit into the complete model the error is propagated into f and k.

A solution to this problem could be the integration of contemporary images producing a supportive sparse cloud of the same building. In the following, the joint estimation of pose and interior orientation (with varying principal point) would be possible (Larsson et al., 2018). Another attempt is to run a subsequent bundle adjustment after a first model is created with solely refining the principal points' coordinates (Schönberger and Frahm, 2016).

The last parameter k models radial distortion in a simple way. The model is taken from VisualSFM as introduced by Wu (2013). It is recommended for the use on internet photo collections (Wu, 2013; Wu, 2014; Schönberger and Frahm, 2016) and determines the form of distortion. A negative k models barrel distortion, while a positive k models pincushion distortion (Luhmann et al., 2019). In this thesis' experiments, the calibration of more distortion coefficients lead to degenerate configurations. Consequently, k is usually set to zero, initially assuming there is no significant distortion which in fact, holds for most historical images.

The camera model and parameterization for f, cx, cy, k is initialized for every image and used in the subsequent bundle adjustment where f and k are refined while $c(x, y)$ stays initially fixed.

1.3.5 Bundle adjustment

COLMAP's bundle adjustment is used for the estimation of interior and exterior camera parameters and creation of a sparse point cloud. As described by Luhmann et al. (2019)

Definition. *bundle adjustment "estimates 3D object coordinates, image orientation parameters and any additional model parameters, together with related statistical information about accuracy and reliability. All observed (measured) values, and all unknown parameters of a photogrammetric project are taken into account within one simultaneous calculation which ensures that homologous rays optimally intersect."*

COLMAP uses an implementation which minimizes the reprojection error while jointly refining the interior and exterior orientation (Schönberger and Frahm, 2016). The non-linear equation system is solved by using the Levenberg-Marquardt algorithm efficiently, implemented by Ceres Solver (Agarwal et al., 2022). In order to produce a reconstruction up to scale, the final cost parameter of the bundle adjustment needs to converge. COLMAP provides the option to tune a large number of parameters and thresholds for its bundle adjustment, drastically increasing the chance for a convergent setup when using historical images. Parameter choices are described in Chapter 4 and 5.

Additionally, COLMAP's bundle adjustment is able to filter images, e.g., if the estimated principal distance is too large or too small, if the reprojection error is too large, or if the bundle adjustment does not converge for single images. Multiple sparse models are generated, if COLMAP is not able to globally optimize the alignment for all images. This is especially helpful for historical images, as sometimes, different building states are depicted for different points in time (Chapter 5).

1.3.6 Dense reconstruction

Following to the reconstruction of a sparse point cloud, usually a dense point cloud is calculated. This step is mostly left out for historical images due to multiple reasons but shall briefly be discussed.

The purpose of the generated 3D models is the depiction in a Web3D/VR/AR environment (Section 3). While this is easily manageable in real-time for simple block models (level of detail (LOD) 1-3), this requires a more advanced implementation (Schütz, 2016) for multiple dense point clouds with e.g., an average size of 100 megabyte (MB). While there are lower borders for an implementation in a desktop environment, a realization on mobile devices would possibly suffer from simpler hardware, mobile internet connection, and data transfer.

However, the main reason for neglecting dense reconstruction is the erroneous/inaccurate creation of dense point clouds (Maiwald et al., 2017). Some object points in the sparse cloud are e.g., created multiple times because of their difference in appearance in the corresponding images. This is invoked e.g., by temporal changes or various construction stages of the building. Consequently, the dense point clouds created by Metashape are inhomogeneous in coloring of points and their surfaces appear quite noisy. COLMAP dense reconstruction and the open Multi-View Stereo (OpenMVS) reconstruction library (Cernea, 2020) produce similar results which can be seen in Figure 1.13.

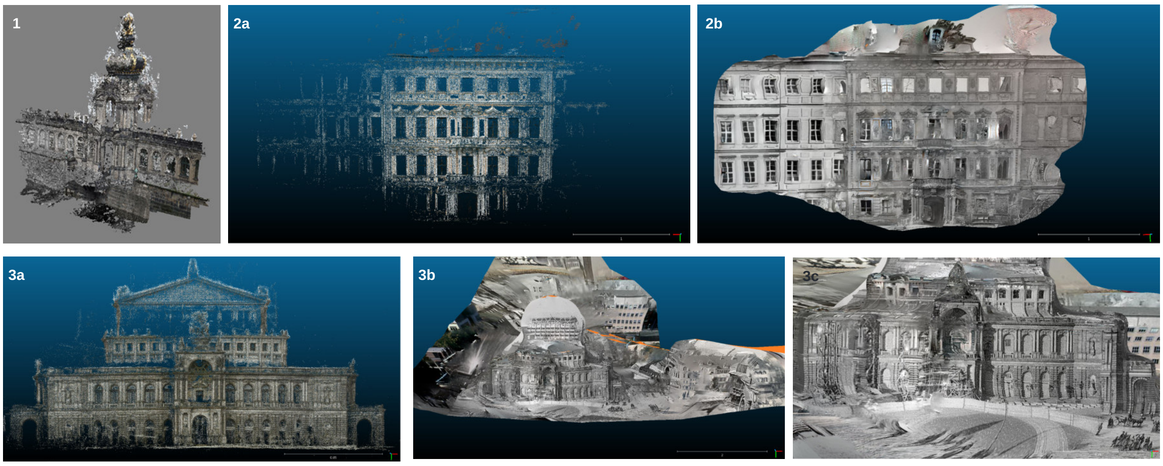


Figure (1.13): Dense reconstructions calculated by Metashape (1 by Maiwald et al. (2017)), COLMAP (2a, 3a) and OpenMVS (2b, 3b, 3c). The dense cloud generated by COLMAP and Metashape show similar properties. Both are still quite sparse on certain areas, are noisy and use varying point color information on different building parts. OpenMVS uses a different algorithm for creating dense meshes and seems to use all information contained in the images. This results in odd meshes mixing various building states.

COLMAP uses the screened Poisson surface reconstruction (Kazhdan and Hoppe, 2013) which produces similar results as the proprietary equivalent Metashape. This type of surface reconstruction uses the already existing sparse points to create more points if certain criteria are met. For few historical images with few a-priori triangulated points in the sparse point cloud the approach generates only several new points, hence very robust to noisy data and misregistration artifacts in comparison to OpenMVS (Kazhdan and Hoppe, 2013). OpenMVS uses a variation of a patch-based stereo approach (Shen, 2013). It aims at the generation of very dense point clouds but uses all images to generate depth-maps which are subsequently merged. While there is a check for consistency between neighboring depth-maps the results show that this is not an appropriate approach to create dense clouds from historical images (Fig. 1.13).

New approaches on dense cloud reconstruction (Yao et al., 2018; Chen et al., 2019) are possibly able to deal with these situations. Additionally, there exists recent research on rendering new synthesized views of a sparse cloud that are not directly seen by the cameras using neural rendering techniques (Mildenhall et al., 2022; Martin-Brualla et al., 2020). It has to be mentioned that creating a dense cloud generally does not refine the camera poses.

Consequently, the proposed workflow is terminated after bundle adjustment and the resulting camera poses are used for texturizing simpler 3D models.

1.3.7 Georeferencing

The 3D models, historical maps and digital elevation models in the 4D browser environment are located in a global coordinate system. Without further information (control points, accurate locations of the cameras), COLMAP is only able to generate camera poses and sparse cloud up to scale, i.e., in a local coordinate system. A Helmert transformation (also 7-parameter transformation, or similarity transform) can be used to transfer the local camera poses into the global coordinate system. The transformation matrix can be derived from homologue 3D point coordinates in both coordinate systems. While the determination of the global coordinates X, Y, Z of the cameras is not possible, a reliable method is the use of similar points in the sparse cloud and the related global 3D model. For a more accurate calculation of the transformation matrix, more detailed 3D models or additional measurements in the global coordinate system can be helpful. At the moment, the process of finding similar points was done in an interactive way, but when further images are added to the reconstruction the calculated transformation matrix can be used.

The transformation from COLMAP (OpenCV coordinate system) into the 4D browser environment (OpenGL coordinate system) is not a trivial one and thus explained in Appendix 8.2.

Global pose and interior orientation is transferred and stored for every historical image and can be depicted as an overlay as well as used for texturing in a 4D browser environment (Fig. 1.14).

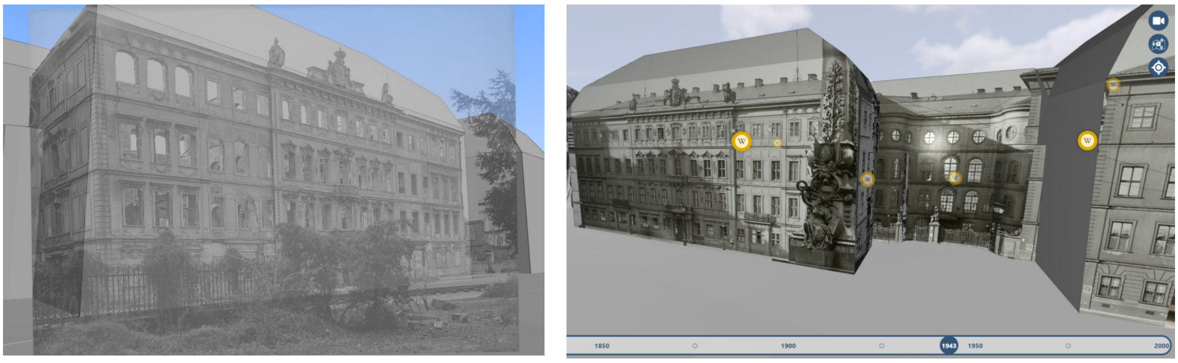


Figure (1.14): Two application scenarios for camera pose estimation of historical images. The left image shows a historical image of the Taschenbergpalais, Dresden as a semi-transparent overlay over a 3D city model (`https://4dbrowser.urbanhistory4d.org/`). The right image shows the Taschenbergpalais from a different perspective where the texture of the oriented historical image is directly projected onto the 3D model's surface (`4dcity.org`).

1.4 Research objectives

The estimation of interior and exterior parameters of historical cameras is a time-consuming process which is at the moment only done for individual buildings, manually determining corresponding image points. The ongoing digitization of historical photographs requires a strategy for accurate pose estimation in order to represent the images in a 3D space. This leads to the following research objectives:

1. Photogrammetry and Computer Vision are still uncommon fields for many researchers in architecture, art history, and architectural history. There are still only few examples where Photogrammetry is used in interdisciplinary humanistic research groups, directly assessing the needs of the future users. Therefore, it is reasonable to classify and describe the use of photogrammetric methods in this context. Further, it should be emphasized how photogrammetric methods can be used as a link between image data in archives and 3D applications.

2. As a side issue it has been identified that the data quality in archives is very inhomogeneous. Searching by metadata for a specific building, usually results in photographs of the interior, the exterior, partly related buildings, or even non-related structures. Additionally, the query yields close-up images or photographs of events and persons. Sometimes, the building is occluded or depicted on drawings. For photogrammetric purposes only exterior views (photographs) of the building are of interest as these can be processed in a geometrically correct way. Interior views could be topic of further investigations but are not part of this research. As metadata search does not yield satisfying results, a different approach based on the content of the image is investigated.

3. The joint estimation of interior and exterior camera parameters is usually done in a SfM workflow. While these are optimized for sequential image blocks, it has to be investigated whether such a workflow will converge when using exclusively historical images. It could already be shown in prior research that the reconstruction of large-scale historical datasets usually fails or requires manual effort (Maiwald et al., 2017). Consequently, the main reasons for such incomplete reconstructions have to be identified in a first step.

4. A major issue that could be repeatedly observed for all kinds of historical images, is the difficult pair-wise image matching. This is induced by large geometric and radiometric differences between historical image pairs. Thus, pair-wise registration of historical images requires a robust feature matching method including steps for feature detection, description, descriptor matching and geometric verification. For the Computer Vision community several benchmark datasets exist, for which feature matching methods can be evaluated. However, it could be proven that especially algorithmic feature matching methods were not able to transfer the high quality benchmark results to historical image pairs. This emphasizes the need for the creation of a pre-registered benchmark dataset using exclusively historical images.

5. Evaluating different feature matching methods on a benchmark dataset allows the selection of a suitable algorithm for processing manifold historical images. The following SfM reconstruction of the camera poses results in a local coordinate frame and the interior camera parameters are estimated. Usually, there exists no reference data or ground truth for exterior and interior camera parameters of historical photographs. Consequently, a strategy has to be developed that allows the investigation of the accuracy of the derived SfM results.

6. In a last step the local SfM reconstruction needs to be georeferenced for depiction in applications which use e.g., georeferenced 3D models, points of interest (POIs), or map data. A strategy has to be developed which allows this integration even if the depicted building is no longer present. In order to process data of various libraries, archives, and repositories a completely automatic pipeline for all above mentioned steps is a preferable research objective.

2 Generation of a benchmark dataset using historical photographs for the evaluation of feature matching methods

Chapter 2 published under CC-BY license in "The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences" (eISSN: 2194-9034) as:

Generation Of A Benchmark Dataset Using Historical Photographs For An Automated Evaluation Of Different Feature Matching Methods

Authors: Ferdinand Maiwald[1]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany

Publication history: submitted January 2019, accepted April 2019, published June 2019
Full reference: Maiwald, F.: GENERATION OF A BENCHMARK DATASET USING HISTORICAL PHOTOGRAPHS FOR AN AUTOMATED EVALUATION OF DIFFERENT FEATURE MATCHING METHODS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W13, 87–94, https://doi.org/10.5194/isprs-archives-XLII-2-W13-87-2019, 2019.

Abstract: This contribution shows the generation of a benchmark dataset using historical images. The difficulties when working with historical images are pointed out and structured in three categories. Especially large viewpoint differences, image artifacts and radiometric differences lead to weak matching results with classical feature matching approaches. The necessity of publishing an own benchmark dataset is emphasized when comparing to existing datasets which are partly using synthetic data, well-known orientation or strictly categorized image differences. The presented image dataset consists at the moment of 24 images which are oriented in image triples using the properties of the Trifocal Tensor as a more stable image geometry. In the following, three different feature detectors and descriptors that have already been proven well on historical images (MSER, ORB, RIFT) are evaluated using the new benchmark dataset. Then, several outlier removal methods were applied on the detected features. The tests show that for the entirety of image pairs RIFT performs slightly better than the other two methods. Nonetheless, for some image pairs MSER significantly improves the matching score but even so, historical image pairs are difficult to be matched with the presented methods due to challenging outlier removal. Still, the estimated projective relative orientation could be used in an autocalibration approach to place the images in a metric scene.

Keywords: benchmark, image dataset, historical images, image orientation, feature matching

2.1 Introduction

This contribution presents the generation of a benchmark dataset for the evaluation of different feature matching methods on historical images. The work is placed in the context of a 4D web application (3D models and related historical images and data) of the city of Dresden as an alternative media repository for e.g., art historians. Oriented images and methods to match historical images provide the basis for the placement of the images in such a 3D space. The presented images originate from the photo library of the Saxon State and University Library Dresden (SLUB), which contains about 1.8 million images of 80 institutions at this point in time. The majority of images in this archive was taken between 1940 and 1990 (`http://deutschefotothek.de`). The images for the benchmark dataset were redigitized for this purpose and show various buildings. While the absolute orientation of these historical photographs is neither given nor easy to define, this approach focuses on the determination of the relative orientation between different historical images.

This leads to diverse issues considering that extrinsic and especially intrinsic camera parameters are mostly unknown. Additionally, the images are taken by different camera types which vary in exposure and acquisition time. Consequently, the presented dataset is relatively oriented in a projective frame using a more stable triple image geometry (Hartley, 1997b). The matches between the three images of one building view and additionally the relating Trifocal Tensor T are determined and given. This orientation data can then be used to evaluate different feature detectors, descriptors and feature matching methods on historical images. In the following, it may be possible that an oriented image mosaic can be metrically spatialized in a three-dimensional environment with the appropriate scale using autocalibration (Faugeras et al., 1992). The dataset consists of 24 images (2 image triples respectively for 4 buildings) and could be extended in the future. The images have different properties ranging from small viewpoint and radiometric changes to large differences. These properties can be summarized in the following three categories.

2.1.1 Image differences based on digitization and image medium

Even if an image would have been taken twice at the same moment in time, some differences concerning the digitized copy could occur during and even before digitization. This happens because historical images are mainly archived on photographic plates or photographic film. Any change on this original data is preserved during digitization. Especially, the conservative emulsion on the glass plates can deteriorate and additionally a glass plate is fragile and any crack will be pictured in the digitized image (Gillet et al., 1986). Scratches, dust and fingerprints may also be visible in the digital copy.

Similarly, photographic film is vulnerable to damage e.g. by mold, photo-oxidation, air pollutants and improper handling (Slate, 2001). All of these image artifacts are transferred using digitization techniques and will interfere with the process of feature detection. One further image difference that may appear and is relevant for photogrammetry is the change of the principal point in the digital copy. It does not have to be necessarily the middle of the digital copy but it can shift, if only a part of the original image is digitized or if the original data has been cropped. It may be even possible that the principal point is not pictured on the digital copy. Additionally, when the digitization information (sensor, resolution, dynamic range, working area, accuracy, filters) is not available every metric data is lost in the process.

2.1.2 Image differences based on different cameras and acquisition technique

When comparing various historical images, the main difference between them is the strongly changing representation of the depicted object. Photographs of the same object are taken in summer and in winter, in daylight and in nighttime and thus, the radiometric properties change. The historical images may be blurred, noisy, under- and overexposed and different light spots, reflections and shadows can appear in the same photographic scene and interfere with the feature detection. Sometimes, people, cars or other objects are in front of the depicted building and influence the feature matching.

Additionally, on the one hand it is possible that there are extreme viewpoint changes between the images and on the other hand sometimes one building is solely photographed from similar perspectives, which makes a 3D reconstruction difficult. Since the camera types are mostly unknown and undocumented the inner orientation important for the reconstruction is not available and has to be estimated.

2.1.3 Object differences based on different dates of acquisition

A difficult topic is the dealing with object differences shown in the photographs. Building differences can vary between very small changes like on claddings, window frames or small statues to large ones considering destroyed or reconstructed buildings. It is not possible to assume that a historical building that is represented on various images did not change over time. Nonetheless, some valuable orientation information can even be determined using these destroyed or changed buildings. It will be difficult to decide whether an object changed so much that any metric information generated with photogrammetric methods is invalid. Furthermore, it is still discussed how to represent this error-prone data (Apollonio, 2016; Kensek et al., 2004). It could be possible in a first step to categorize historical images using content-based image-retrieval on a very accurate scale and only use feature matching methods on image pairs of clearly the same building in the same state.

2.2 Related work

There exists already a numerous variety of image datasets in computer vision for different purposes like (people-)detection, classification, recognition, tracking, segmentation, multiview and many more. Famous datasets are e.g. the Caltech 256 dataset for classification purposes (Griffin et al., 2007) or the KITTI dataset used in autonomous driving and SLAM research (Geiger et al., 2013). The presented dataset could be integrated in the multiview category and closes a gap between different existing datasets. In contrast to datasets with a lot of images and their inner orientations (Moreels and Perona, 2005) it is not or only hardly possible to provide that many historical images including the proper inner orientation since the camera types are mostly unknown.

Similar to the Affine Covariant Regions dataset (Mikolajczyk et al., 2005) the presented benchmark dataset consists of real data (= not synthetic data) with changes in illumination, viewpoint, blur and rotation. Some of the historical images even have large viewpoint or illumination changes like in the Extreme View Dataset or the Ultra Wide Baseline Dataset (Mishkin et al., 2015a). These existing datasets are using the fact that "the images are either of planar scenes or the camera position is fixed during acquisition, so that in all cases the images are related by homographies [..] and this mapping is used to determine ground truth matches [..].” (Mikolajczyk et al., 2005). This is not (always) possible when using historical data, so the presented benchmark dataset is described by the predefined corresponding points and the Trifocal Tensor determining the relative orientation between image triples. At the time of this research no other freely available benchmark dataset with oriented images older than 40 years used for feature detection and matching could be found.

However, many people are working with historical images and further data to reconstruct mostly buildings and sights. This includes e.g., the reconstruction of the great Buddha of Bamiyan (Grün et al., 2004), dinosaur tracks (Falkingham et al., 2014) or the orientation of historical images of Atlanta, GA (Schindler and Dellaert, 2012). But, also recent research is done with historical data e.g., in combination with terrestrial laser scanning (Bitelli et al., 2017), using old film negatives (Rodríguez Miranda and Valle Melón, 2017) or aerial images (Giordano et al., 2018). Though, those projects show a developing degree of automation in image processing a lot of work is still done manually in this field of research (Henze et al., 2009a; Gouveia et al., 2015). An oriented historical image dataset could help to improve automated approaches in image classification, image matching and image orientation.

2.3 The image dataset

Examples for the historical image dataset are shown below (Fig. 2.1).

The whole published dataset consists of 24 images with a maximum side length of 3543 pixels. It is mostly unclear, whether the original data is originated from photographic plates or film negatives. The images are grouped in two triplets respectively for 4 buildings ($2 \times 3 \times 4 = 24$). Images were chosen with respect to their possible matching quality. The images show combined differences in illumination, field of view, viewpoints, blurring and slight rotation. Some of the images show building reflections in water or extreme shadowing. Thus, a very challenging dataset when using a single feature matching method is provided.

Since the relative orientation of the image pairs cannot be easily described through a homography as explained before, the first step would be the description of the image pairs using a Fundamental Matrix F calculated out of at least 7 point correspondences, where F is defined by equation 2.1,

$$x'^T F x = 0 \tag{2.1}$$

where x' and x are at least 7 image correspondences in homogeneous coordinates.

One must say, that this equation can hardly be used to test correspondences determined with feature matching methods because an estimated (e.g. using the Random Sample Consensus (RANSAC)) Fundamental Matrix F is only a projective map taking a point to a line. That means a point $x(x, y, z)$ in the first image defines a line (the corresponding epipolar line $l' = Fx$) in the second image (Hartley and Zisserman, 2003). Additionally, the point transfer from image 1 to image 2 using the epipolar line can lead to false positives considering matches that lie randomly on the epipolar line but are no true matches.

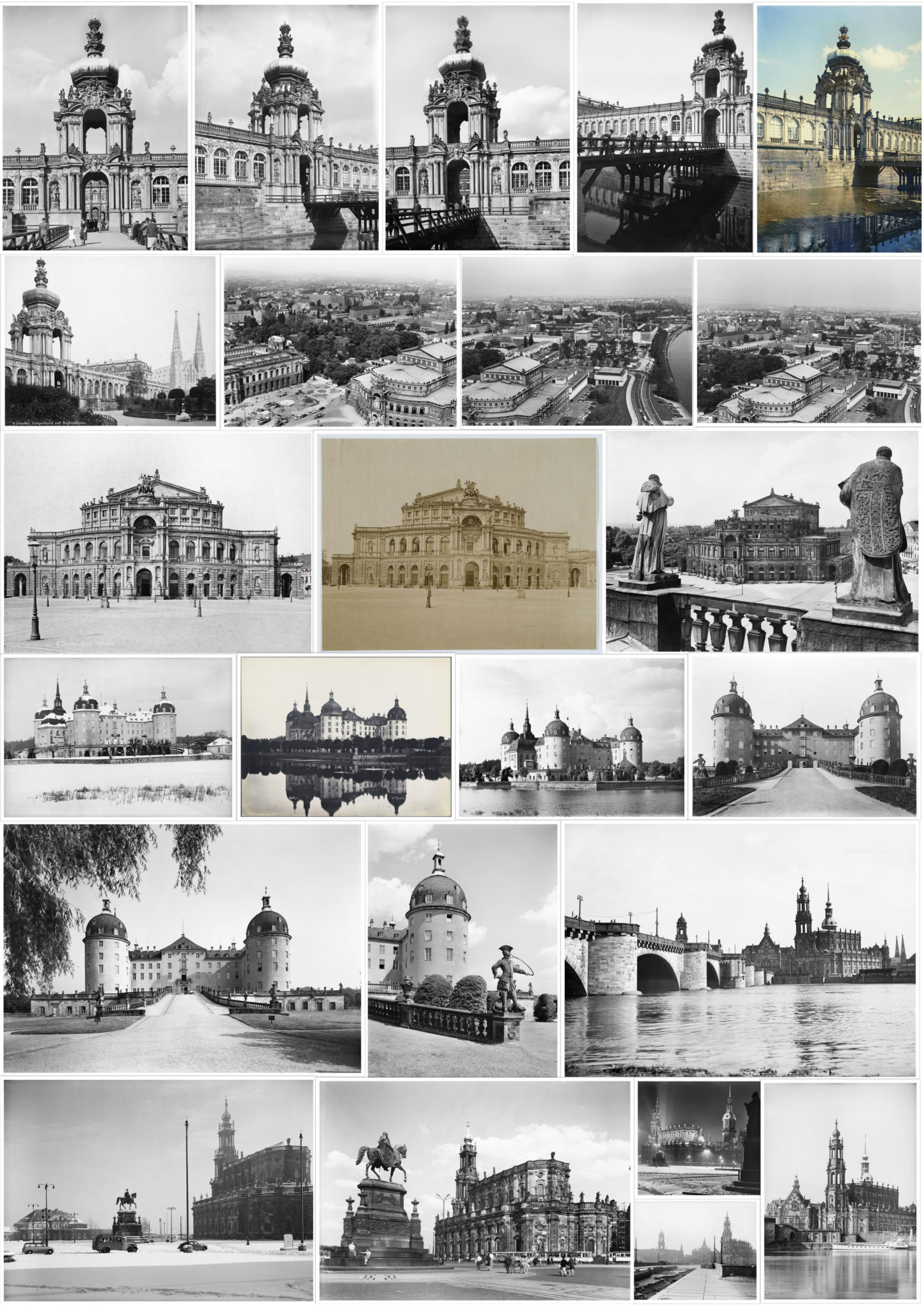


Figure (2.1): All current images of the benchmark dataset showing the variety of historical images

This leads to a more stable image configuration when using three images, because e.g. the epipolar lines from image 1 to image 3 and from image 2 to image 3 of the same feature point intersect in image 3 in the homologue feature point (Maas, 1997). The matching can be simplified using the $3 \times 3 \times 3$ Trifocal Tensor T and its properties for a point-point-point correspondence (eq. 2.2) (Hartley and Zisserman, 2003).

$$[x']_\times (\sum_i x^i T_i)[x'']_\times = 0_{3\times3} \tag{2.2}$$

where x, x', x'' = image coordinates in the three images
$[x]_\times = 3 \times 3$ skew-symmetric matrix of 3-vector
i = number of 3×3 Tensor slice
T = Trifocal Tensor of the three images
$0_{3\times3} = 3 \times 3$ null matrix

A point transfer from e.g. the first view to the third view can then be realized using equation 2.3 and the corrected Fundamental Matrices F_{12}, F_{13} and F_{23} extracted from the Trifocal Tensor (Hartley and Zisserman, 2003).

$$x''^k = x' l'_j T_i^{jk} \tag{2.3}$$

where i,j,k = indices that correspond to the entities
in the first, second and third views respectively

Since the Trifocal Tensor is not that easy to determine like a homography or the Fundamental Matrix, it is provided for every benchmark image triple. Additionally, the calculation of T is explained in the following. There are various methods that are used for the computation of the Trifocal Tensor namely e.g., the minimal parameterization by Faugeras and Papadopoulo (Faugeras and Papadopoulo, 1998) and by Nordberg (Nordberg, 2009) or the constrained solutions by Ponce and Hebert (Ponce and Hebert, 2014) as well as Ressl (Ressl, 2002). Most approaches have already been tested and the constrained solution by Ressl has shown the most robust results leading to the smallest reprojection errors (Julià and Monasse, 2018).

Using this computation method requires approximation values for the Trifocal Tensor and the Projection Matrices of the three images. These can be found by solving $At = 0$ in a linear way. The matrix A is the Jacobian of the trilinearities and consists of the ($n = 4$) row-wise ordered sub matrices A_c where c is the number of point correspondences (eq. 2.4) (Ressl, 2003).

$$A_c = (S^{\text{red}}(x'') \otimes S^{\text{red}}(x'''))(x'^T \otimes I_9) \tag{2.4}$$

where S^{red} = reduced axiator for point coordinates
$\otimes$ = Kronecker product
for 4 linearly independent equations

Afterwards, the approximation values can be calculated by minimizing the algebraic error using a singular value decompositon (SVD). It is recommended to use at least 10 normalized point correspondences in all three images with a pixel noise of 1 to minimize the reprojection error with the subsequent constrained solution (Ressl, 2003). For the benchmark dataset at least 15 manual point correspondences were used in the image triples and a pixel noise < 1 was targeted. The verified results for the different matching strategies show that this goal could be accomplished. The detailed description, the images, the matched points and the corresponding Trifocal Tensor are available on the website[1].

Since the Trifocal Tensor provides geometric relations between three views only in a projective frame independent of scene structure (Hartley and Zisserman, 2003) the resulting camera matrices P, P', P'' retrieved by equation 2.5 could be introduced as a prior relative orientation into an autocalibration algorithm (Heinrich et al., 2011) allowing the estimation of inner and exterior orientation and in the following, the generation of simple structures in euclidean metric 3D space.

$$\begin{aligned} P &= [I \mid 0] \\ P' &= [[T_1, T_2, T_3]e'' \mid e'] \\ P'' &= [(e''e''^T - I)[T_1^T, T_2^T, T_3^T]e' \mid e''] \end{aligned} \tag{2.5}$$

where e', e'' = respective normalized epipoles
T_1, T_2, T_3 = Tensor slices

2.4 Comparison of different feature detection and description methods

In the following, the different feature detection methods used on the benchmark image dataset are briefly explained. Three distinct algorithms were chosen to process the images in full resolution and to find point features. The comparison is done between image pairs but can be evaluated using the Trifocal Tensor. Thus, the number of correct matches in relation to the sum of all matches (= matching score) could be determined. Some of the common methods have already been tested on historical image data and a combination of the Oriented FAST and Rotated BRIEF (ORB) feature detector and the Speeded-Up Robust Features (SURF) feature descriptor produced decent results (Ali and Whitehead, 2014). Another approach that generated a good matching ratio was the Maximally Stable Extremal Regions (MSER) feature detector and descriptor (Wolfe, 2013). Additionally, those results are compared with a newer method called Radiation-Invariant Feature Transform (RIFT), that neglects radiometric differences in images and thus, can be a good addition to existing approaches.

For the first and second test the standard implementations of ORB, SURF and MSER in OpenCV were used. The third test used the implementation of RIFT in Matlab (Li et al., 2018). The results are presented without outlier removal using brute force matching, outlier removal using a symmetry test and as a third approach outlier removal using Fundamental Matrix calculation with RANSAC (Fischler and Bolles, 1981). Additionally, for RIFT the native calculation using the Fast Sample Consensus (FSC) (Wu et al., 2015) is shown.

[1] `https://dx.doi.org/10.25532/OPARA-24`

2.4.1 Oriented FAST and Rotated BRIEF (ORB)

ORB is a common alternative to SIFT and uses an intensity oriented FAST (Rosten and Drummond, 2006) for feature detection and an in-plane rotation invariant version of BRIEF (Calonder et al., 2010) for feature description (Rublee et al., 2011). Since the hybrid version using the ORB detector and the SURF descriptor achieved better results on historical images (Ali and Whitehead, 2014), the presented approach chooses this as a first method for feature detection and description. The oriented FAST detects keypoints using the intensity threshold between the center pixel and a circular ring around that center. The orientation of the keypoints is done using an intensity centroid (Rosin, 1999). The standard maximum value for the number of features retained was set from a maximum of 500 to 2500 to allow a better comparison with the other methods. In the following, ORB is used for the description of the features since it outperforms other descriptors by repeatability, distinctiveness and robustness (Bay et al., 2006).

2.4.2 Maximally Stable Extremal Region Detector (MSER)

As a second method the presented approach uses MSER (Matas et al., 2004). This algorithm is usually applied on image pairs with a wide baseline. Classical feature points are replaced by regions which are closed under projective transformation of image coordinates and monotonic transformation of image intensities (Matas et al., 2004). Those properties can be especially useful for historical images because of the already explained image differences. Regions described by a connected number of pixels are chosen by the property that all pixels inside one extremal region have either a higher or a lower intensity than all the pixels on its outer boundary (Mikolajczyk et al., 2005). Again, SURF is used for the description of the regions consisting of feature point sets.

2.4.3 Radiation-invariant Feature Transform (RIFT)

The third method used is called RIFT. The radiation-invariant feature transform is chosen because of its invariance to nonlinear radiation distortions (NRD) (Li et al., 2018) and the use of edge features in addition to corner features. Both effects can support the feature detection in historical images. The approach uses the Fourier transform to generate phase congruency maps. Independent maps for each orientation of a 2D log-Gabor filter are created and used for the detection of corner features as well as edge features. In the following, those features are described by a 216-dimensional feature vector calculated through a maximum index map based on a log-Gabor convolution sequence (Li et al., 2018). RIFT is currently not scale invariant and so it should perform bad on large scale-changes. However, it has been observed that feature points in image pairs with small scale-changes can still be matched correctly.

2.4.4 Feature matching and outlier removal

For the comparison of all methods, the presented approach uses a brute force matching (_bf) for all detected feature points, i.e. all feature points with their particular descriptors are matched (so every descriptor in image 1 is compared with every descriptor in image 2). In the following, two different outlier removal methods are evaluated. The first approach uses a symmetry test (_sym). So matches from image 1 to image 2 are only kept if these are also matches from image 2 to image 1. In the second approach the calculation of a Fundamental Matrix between both images based on the feature matching result using brute force matching is used to eliminate outliers. Therefore, the RANSAC algorithm (_RANSAC) was chosen

(Fischler and Bolles, 1981). Additionally, for RIFT the already implemented outlier removal (_native) using the Fast Sample Consensus (FSC) (Wu et al., 2015) is shown.

2.5 Results

The results of the different feature detection and matching methods are shown in Table 2.1.

Table (2.1): Results for different feature matching methods for 8 different image triples (=24 image pairs). Matching results are shown respectively for every dataset for the image pairs 1__2, 1__3, and 2__3 as ratio in % between all found matches and correct matches (matching score). Good results are highlighted in green whereas bad results are shown in red

Dataset	Mb_1			Mb_2			Zw_1			Zw_2			So_1			So_2			Hk_1			Hk_2		
Imagepair	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3	1_2	1_3	2_3
Serial Number #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MSER_bf	0,74	1,17	0,33	0,40	0,58	0,24	9,04	3,01	10,33	0,18	0,39	0,47	3,20	10,24	7,07	0,77	3,30	0,26	0,33	0,26	0,07	1,15	0,61	1,31
MSER_RANSAC	0,00	1,12	0,00	0,00	2,88	16,22	59,82	0,00	54,17	0,00	0,00	0,00	23,85	29,20	63,02	0,00	13,33	0,00	4,95	2,11	0,00	0,00	0,00	0,00
MSER_sym	0,81	1,38	0,40	0,60	0,41	0,20	17,05	5,00	20,89	0,23	0,44	0,77	5,75	18,10	17,79	0,74	8,19	0,62	0,61	0,24	0,06	1,12	1,46	2,06
ORB_bf	2,00	1,28	0,72	1,24	0,52	1,12	6,88	2,64	10,56	0,40	1,28	2,40	2,80	7,88	9,04	1,00	4,44	0,76	0,64	0,08	0,00	4,24	0,44	0,76
ORB_RANSAC	44,83	9,09	0,00	3,45	0,00	0,00	10,71	3,23	57,72	0,00	0,00	22,45	17,31	15,74	42,42	0,00	13,64	5,26	0,00	0,00	0,00	42,55	0,00	0,00
ORB_sym	2,38	1,79	0,58	2,19	1,19	2,37	17,36	6,34	22,34	0,48	2,42	3,89	6,73	14,67	20,97	2,46	7,30	1,39	1,11	0,00	0,00	9,28	0,39	0,70
RIFT_bf	6,44	7,80	4,88	0,92	1,92	1,76	4,92	1,56	16,45	1,40	4,24	4,52	3,12	12,53	9,68	0,80	7,48	0,40	3,72	0,96	0,48	0,84	1,96	3,56
RIFT_RANSAC	0,00	1,01	0,00	0,00	0,00	0,00	3,85	0,00	80,00	0,00	18,18	0,00	0,00	30,95	0,83	0,00	12,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RIFT_sym	18,51	26,71	18,18	1,65	5,77	5,15	15,24	4,34	36,49	3,71	10,86	17,80	16,09	27,07	34,30	1,13	16,33	1,15	17,09	2,80	2,94	2,42	10,00	11,46
RIFT_native	77,78	83,33	80,43	0,00	0,00	72,73	47,62	0,00	50,00	15,38	46,94	62,50	42,50	0,00	0,00	0,00	8,82	25,00	66,67	25,00	25,00	0,00	69,23	84,21

For every image triple the matches are shown respectively at first between image 1 and image 2, secondly between image 1 and image 3 and at last for image 2 and image 3. Therefore, the number of correct matches (determined using the point transfer with the Trifocal Tensor) is compared with the absolute number of matches and given as ratio in (also referred to as matching score) with respect to the feature matching method and the applied outlier removal. Good results (> 40%) are highlighted in green whereas bad results (< 40%) are highlighted in red. The transition from red to green around 40% is coloured in white. Additionally, Table 2.2 shows the number of all correct matches determined by the different approaches for the 24 image pairs.

Table (2.2): Results for different feature matching methods for 8 different image triples (=24 image pairs). The total number of correct matches is shown respectively for every dataset for the image pairs 1__2, 1__3, and 2__3. Good results are highlighted in green whereas bad results are shown in red

Serial Number #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MSER_bf	60	95	16	60	86	42	1064	354	1742	24	51	63	707	2266	1904	123	530	43	14	11	9	95	50	102
MSER_RANSAC	0	1	0	0	3	18	329	0	942	0	0	0	150	2239	1554	0	28	0	5	2	0	0	0	0
MSER_sym	7	22	4	11	6	3	282	75	448	4	9	15	213	747	679	15	87	6	5	2	1	11	9	13
ORB_bf	50	32	18	31	13	28	172	66	264	10	32	60	70	197	226	25	111	19	16	2	0	106	11	19
ORB_RANSAC	26	4	0	1	0	0	12	1	142	0	0	11	9	172	154	0	9	2	0	0	0	20	0	0
ORB_sym	15	12	3	12	7	16	117	41	170	3	15	22	44	151	156	14	53	8	7	0	0	57	2	4
RIFT_bf	161	195	122	23	48	44	123	39	411	35	106	113	78	313	242	20	187	10	93	24	12	21	49	89
RIFT_RANSAC	0	1	0	0	0	0	1	0	20	0	2	0	0	39	1	0	1	0	0	0	0	0	0	0
RIFT_sym	87	125	72	4	18	17	80	16	316	15	63	89	60	281	213	3	104	3	61	7	7	7	28	33
RIFT_native	42	35	37	0	0	8	10	0	46	4	23	20	17	0	0	0	3	1	4	2	2	0	9	16

The total number of feature points is provided within the dataset. The image pairs are serially numbered from 1 to 24. It is easy to see that for example, the image pair 16 could not be matched by any of the algorithms with a good result in opposite to e.g., the image pair 9 where every approach shows better results highlighted in the respective column in green. For an easier comparison the matching scores (ratio) and the number of correct matches of the four best approaches (MSER_RANSAC, ORB_RANSAC, RIFT_RANSAC, RIFT_native) are shown in two different diagrams (Fig. 2.2, 2.3).

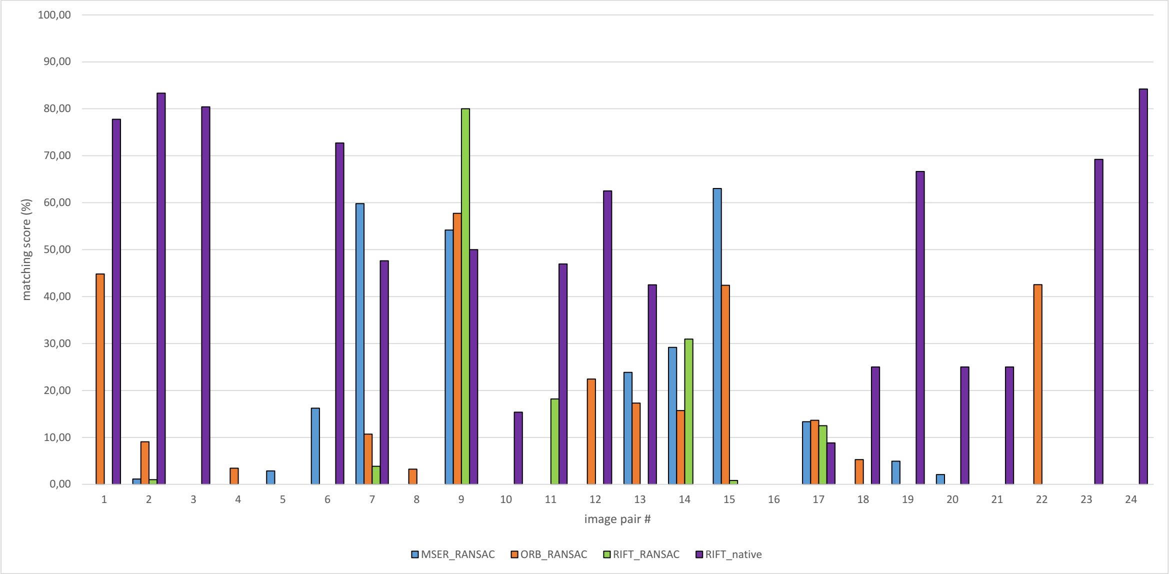


Figure (2.2): Matching scores of every image pair for the four best performing algorithms MSER_RANSAC, ORB_RANSAC, RIFT_RANSAC, and RIFT_native

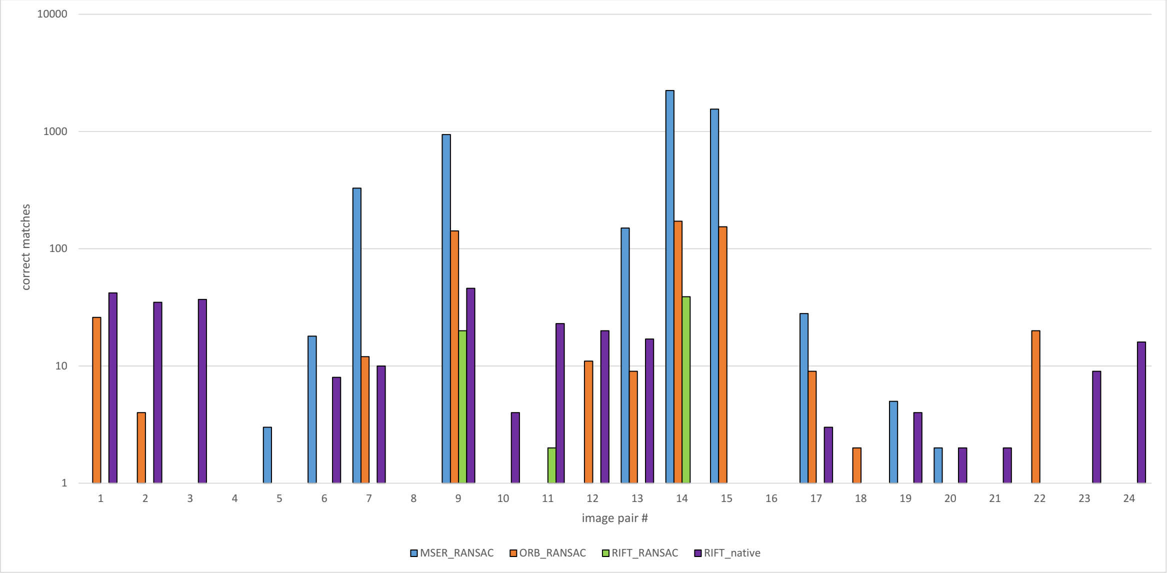


Figure (2.3): Total number of correct matches shown on a logarithmic scale of every image pair for the four best performing algorithms MSER_RANSAC, ORB_RANSAC, RIFT_RANSAC, and RIFT_native

The tables as well as the diagrams demonstrate that all three methods fall short of expectations. A small number of correct matches can almost always be found for every image pair with the brute force attempt but it is hardly possible to filter those out. Some exceptions exist like e.g., for image pair 7 and MSER only around 9% of the initial matches are correct but with the outlier removal method RANSAC it is achievable to reach a matching score of around 60%.

However, the native version of RIFT using the Fast Sample Consensus produces better results than the other approaches. For a lot of image pairs, a matching score $> 60\%$ could be attained. But, regarding the total number of correct matches (Tab. 2.2, Fig. 2.3) these are very low compared to e.g., MSER but most of the times still enough to perform an estimation of a Fundamental Matrix. The combination of ORB and SURF doesn't outperform any of the other algorithms and is therefore not suitable for the feature matching of the depicted historical images.

100% correct matches could not be reached with the presented approaches. Consequently, it can be said that the crucial point when working with historical images is the outlier removal step. Since there always is a small number of feature points in the image pairs that could be matched it will be the objective to filter those correctly. The symmetry test only slightly improved the results of the brute force matching. RANSAC performs better but most of the times the exact Fundamental Matrix could not be found. It seems that a refined RANSAC algorithm like FSC that is used in the RIFT approach could improve the matching scores.

A combination of all methods could result in higher scores and will be tested in the future. Multiple iterations when calculating the Fundamental Matrix or improved RANSAC algorithms like FSC, PROSAC (Chum and Matas, 2005) or MLESAC (Torr and Zisserman, 2000) could improve the matching scores for all approaches.

Summarizing, for all image triples of the benchmark dataset it is possible to find homologue points and match them almost only using RIFT. MSER generally finds the most feature points but most of the times RIFT shows the highest matching scores in combination with FSC. For some special image constellations, the other approaches could be more appropriate and a combination of methods could lead to better results (Mishkin et al., 2015a). Historical images are still a challenge for classical feature detection and matching algorithms, thus a cautious outlier removal is inevitable.

2.6 Conclusions and future work

The contribution shows the generation and evaluation of a dataset consisting of 24 historical images. Difficulties determining the relative orientation of the data arise due to large image differences and unknown camera parameters. Thus, a more stable image configuration using three images described by the Trifocal Tensor T has been established. Therefore, T is given for every image triple in the dataset. The Trifocal Tensor can be used to evaluate different feature detectors and matching methods on historical images and the dataset can be used as a benchmark set. In this research MSER, ORB and RIFT were used since these algorithms have already shown good results in other publications. For the presented dataset RIFT produced better results than the other two methods. FSC performed better in outlier removal than the symmetry test or RANSAC.

It is planned to establish a more reliable workflow for historical image matching using multiple methods consecutively. Also other already developed approaches will be tested on the dataset in the future (Maiwald et al., 2018). Different outlier removal methods could still improve the matching scores. Additional oriented historical images will be added to the dataset to provide a challenging base for other researchers.

Since the images are oriented with the Trifocal Tensor only in a projective space it is planned to use this estimated relative orientation as a base for a metric solution and calculate the inner and exterior orientation of the historical images. In the following, these images could be placed in the 3D/4D web application. Furthermore, simple features like single lines or planes could be generated in 3D space to create generalized historical 3D models.

Acknowledgments

The research upon which this paper is based is part of the junior research group UrbanHistory4D's activities which has received funding from the German Federal Ministry of Education and Research under grant agreement No 01UG1630. The image dataset licensed under CC-BY-SA-4.0 is available at `https://doi.org/10.25532/OPARA-24`.

References

Ali, H. K. and A. Whitehead (2014). "Feature Matching for Aligning Historical and Modern Images". In: *Int. J. Comput. Appl.* 21, 3, pp. 188–201.

Apollonio, F. I. (2016). "Classification Schemes for Visualization of Uncertainty in Digital Hypothetical Reconstruction". In: *3D Research Challenges in Cultural Heritage II.* Springer International Publishing, pp. 173–197. DOI: 10.1007/978-3-319-47647-6_9.

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

Bitelli, G., M. Dellapasqua, V. A. Girelli, S. Sbaraglia and M. A. Tinia (2017). "Historical Photogrammetry and Terrestrial Laser Scanning for the 3d Virtual Reconstruction of Destroyed Structures: A Case Study in Italy". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5/W1, pp. 113–119. DOI: 10.5194/isprs-archives-xlii-5-w1-113-2017.

Calonder, M., V. Lepetit, C. Strecha and P. Fua (2010). "BRIEF: Binary Robust Independent Elementary Features". In: *Computer Vision – ECCV 2010.* Springer Berlin Heidelberg, pp. 778–792. DOI: 10.1007/978-3-642-15561-1_56.

Chum, O. and J. Matas (2005). "Matching with PROSAC — Progressive Sample Consensus". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* IEEE, pp. 220–226. DOI: 10.1109/cvpr.2005.221.

Falkingham, P. L., K. T. Bates and J. O. Farlow (2014). "Historical Photogrammetry: Bird's Paluxy River Dinosaur Chase Sequence Digitally Reconstructed as It Was prior to Excavation 70 Years Ago". In: *PLoS ONE* 9, 4, e93247. DOI: 10.1371/journal.pone.0093247.

Faugeras, O. and T. Papadopoulo (1998). "A nonlinear method for estimating the projective geometry of 3 views". In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271).* Narosa Publishing House, pp. 477–484. DOI: 10.1109/iccv.1998.710761.

Faugeras, O. D., Q. .-. Luong and S. J. Maybank (1992). "Camera self-calibration: Theory and experiments". In: *Computer Vision — ECCV'92.* Springer Berlin Heidelberg, pp. 321–334. DOI: 10.1007/3-540-55426-2_37.

Fischler, M. A. and R. C. Bolles (1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24, 6, pp. 381–395. DOI: 10.1145/358669.358692.

Geiger, A., P. Lenz, C. Stiller and R. Urtasun (2013). "Vision meets robotics: The KITTI dataset". In: *The International Journal of Robotics Research* 32, 11, pp. 1231–1237. DOI: 10.1177/0278364913491297.

Gillet, M., C. Garnier and F. Flieder (1986). "Glass Plate Negatives. Preservation and Restoration". In: *Restaurator* 7, 2, pp. 49–80. DOI: 10.1515/rest.1986.7.2.49.

Giordano, S., A. L. Bris and C. Mallet (2018). "Toward Automatic Georeferencing of Archival Aerial Photogrammetric Surveys". In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2, pp. 105–112. DOI: 10.5194/isprs-annals-iv-2-105-2018.

Gouveia, J., F. Branco, A. Rodrigues and N. Correia (2015). "Travelling through space and time in Lisbon's religious buildings". In: *2015 Digital Heritage.* IEEE, pp. 407–408. DOI: 10.1109/digitalheritage.2015.7413916.

Griffin, G., A. Holub and P. Perona (2007). *Caltech-256 object category dataset (Unpublished).* Report. California Institute of Technology. URL: http://www.vision.caltech.edu/Image_Datasets/Caltech256/.

Grün, A., F. Remondino and L. Zhang (2004). "Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan". In: *The Photogrammetric Record* 19, 107, pp. 177–199. DOI: 10.1111/j.0031-868x.2004.00278.x.

Hartley, R. and A. Zisserman (2003). *Multiple view geometry in computer vision.* Cambridge: Cambridge university press. ISBN: 0521540518. DOI: https://doi.org/10.1017/CBO9780511811685.

Hartley, R. I. (1997b). "Lines and Points in Three Views and the Trifocal Tensor". In: *International Journal of Computer Vision* 22, 2, pp. 125–140. DOI: 10.1023/a:1007936012022.

Heinrich, S. B., W. E. Snyder and J.-M. Frahm (2011). "Maximum likelihood autocalibration". In: *Image and Vision Computing* 29, 10, pp. 653–665. DOI: 10.1016/j.imavis.2011.07.003.

Henze, F., H. Lehmann and B. Bruschke (2009a). "Nutzung historischer Pläne und Bilder für die Stadtforschungen in Baalbek/Libanon". In: *Photogrammetrie Fernerkundung Geoinformation 3/2009* 3, pp. 221–234.

Julià, L. F. and P. Monasse (2018). "A Critical Review of the Trifocal Tensor Estimation". In: *Image and Video Technology.* Springer International Publishing, pp. 337–349. DOI: 10.1007/978-3-319-75786-5_28.

Kensek, K. M., L. S. Dodd and N. Cipolla (2004). "Fantastic reconstructions or reconstructions of the fantastic? Tracking and presenting ambiguity, alternatives, and documentation in virtual worlds". In: *Automation in Construction* 13, 2, pp. 175–186. DOI: 10.1016/j.autcon.2003.09.010.

Li, J., Q. Hu and M. Ai (2018). "RIFT: Multi-modal Image Matching Based on Radiation-invariant Feature Transform". In: *arXiv:1804.09493*, pp. 1–14.

Maas, H.-G. (1997). "Mehrbildtechniken in der digitalen Photogrammetrie". Dissertation. ETH Zurich. DOI: 10.3929/ETHZ-A-001865074.

Maiwald, F., D. Schneider, F. Henze, S. Münster and F. Niebling (2018). "Feature Matching of Historical Images Based on Geometry of Quadrilaterals". In: *ISPRS - International*

Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2, pp. 643–650. ISSN: 2194-9034. DOI: https://doi.org/10.5194/isprs-archives-XLII-2-643-2018.

Matas, J., O. Chum, M. Urban and T. Pajdla (2004). “Robust wide-baseline stereo from maximally stable extremal regions”. In: *Image and vision computing* 22, 10, pp. 761–767. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2004.02.006.

Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. V. Gool (2005). “A Comparison of Affine Region Detectors”. In: *International Journal of Computer Vision* 65, 1-2, pp. 43–72. DOI: 10.1007/s11263-005-3848-x.

Mishkin, D., J. Matas and M. Perdoch (2015a). “MODS: Fast and robust method for two-view matching”. In: *Computer Vision and Image Understanding* 141, pp. 81–93. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2015.08.005.

Moreels, P. and P. Perona (2005). “Evaluation of features detectors and descriptors based on 3D objects”. In: *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1.* Vol. 1. IEEE, pp. 800–807. DOI: 10.1109/iccv.2005.89.

Nordberg, K. (2009). “A minimal parameterization of the trifocal tensor”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 1224–1230. DOI: 10.1109/cvpr.2009.5206829.

Ponce, J. and M. Hebert (2014). “Trinocular Geometry Revisited”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 17–24. DOI: 10.1109/cvpr.2014.10.

Ressl, C. (2002). “A minimal set of constraints and a minimal parameterization for the trifocal tensor”. In: *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34, 3/A, pp. 277–282. ISSN: 1682-1750.

Ressl, C. (2003). “Geometry, constraints and computation of the trifocal tensor”. Dissertation. TU Wien.

Rodríguez Miranda, A. and J. M. Valle Melón (2017). “Recovering Old Stereoscopic Negatives and Producing Digital 3d Models of Former Appearances of Historic Buildings”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W3, pp. 601–608. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-2-W3-601-2017.

Rosin, P. L. (1999). “Measuring Corner Properties”. In: *Computer Vision and Image Understanding* 73, 2, pp. 291–307. DOI: 10.1006/cviu.1998.0719.

Rosten, E. and T. Drummond (2006). “Machine Learning for High-Speed Corner Detection”. In: *Computer Vision – ECCV 2006.* Springer Berlin Heidelberg, pp. 430–443. DOI: 10.1007/11744023_34.

Rublee, E., V. Rabaud, K. Konolige and G. Bradski (2011). “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International Conference on Computer Vision.* IEEE, pp. 2564–2571. DOI: 10.1109/iccv.2011.6126544.

Schindler, G. and F. Dellaert (2012). “4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections”. In: *Journal of Multimedia* 7, 2, pp. 124–131. ISSN: 1796-2048. DOI: https://doi.org/10.4304/jmm.7.2.124-131.

Slate, J. H. (2001). “Not Fade Away: Understanding the Definition, Preservation and Conservation Issues of Visual Ephemera”. In: *Collection Management* 25, 4, pp. 51–59. DOI: `10.1300/j105v25n04_06`.

Torr, P. and A. Zisserman (2000). “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry”. In: *Computer Vision and Image Understanding* 78, 1, pp. 138–156. DOI: `10.1006/cviu.1999.0832`.

Wolfe, R. (2013). “Modern to historical image feature matching”. Honours Project. Published online: http://robbiewolfe. ca/programming/honoursproject/report.pdf.

Wu, Y., W. Ma, M. Gong, L. Su and L. Jiao (2015). “A Novel Point-Matching Algorithm Based on Fast Sample Consensus for Image Registration”. In: *IEEE Geosci. Remote Sensing Lett.* 12, 1, pp. 43–47. DOI: `https://doi.org/10.1109/LGRS.2014.2325970`.

3 Photogrammetry as a link between image repository and 4D applications

Chapter 3 published under CC-BY-NC-ND license in "Virtual Archaeology Review" (eISSN: 1989-9947) as:

A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR

Authors: Ferdinand Maiwald[1], Jonas Bruschke[2], Christoph Lehmann[3], Florian Niebling[2]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]Human-Computer Interaction, Universität Würzburg, Germany
[3]Centre for Information Services and High Performance Computing, TU Dresden, Germany

Publication history: submitted May 2019, accepted July 2019, published July 2019
Full reference: MAIWALD, Ferdinand et al. A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR. **Virtual Archaeology Review**, [S.l.], v. 10, n. 21, p. 1-13, july 2019. `https://doi.org/10.4995/var.2019.11867`.

Abstract: This contribution shows the comparison, investigation, and implementation of different access strategies on multimodal data. The first part of the research is structured as a theoretical part opposing and explaining the terms of conventional access, virtual archival access, and virtual museums while additionally referencing related work. Especially, issues that still persist in repositories like the ambiguity or missing of metadata is pointed out. The second part explains the practical implementation of a workflow from a large image repository to various four-dimensional (4D) applications. Mainly, the filtering of images and in the following, the orientation of images is explained. Selection of the relevant images is partly carried out manually but also with the use of deep convolutional neural networks for image classification. In the following, photogrammetric methods are used for finding the relative orientation between image pairs in a projective frame. For this purpose, an adapted Structure-from-Motion (SfM) workflow is presented, in which the step of feature detection and matching is replaced by the Radiation-Invariant Feature Transform (RIFT) and Matching On Demand with View Synthesis (MODS). Both methods have been evaluated on a benchmark dataset and performed superior to other approaches. Subsequently, the oriented images are placed interactively and in the future automatically in a 4D browser application showing images, maps, and building models. Further usage scenarios are presented in several Virtual Reality (VR) and Augmented Reality (AR) applications. The new representation of the archival data enables spatial and temporal browsing of repositories allowing the research of innovative perspectives and the uncovering of historical details.

Keywords: historical photographs and maps; photogrammetry; repository; documentation; Geographic Information System (GIS); orientation and matching

3.1 Introduction

The access to extensive repositories of historical photographs and maps using classical database systems presupposes a comprehensive knowledge about structure and semantics of the respective database objects and is aimed above all at experienced users. Without knowledge of the depicted buildings as well as the concrete structuring of data, like keywords, ontologies and metadata, search queries often do not lead to the desired results. Consequently, new representation strategies and the usage of informative models and systems become more and more accepted by the users.

Especially, Geographic Information Systems (GIS) tools combined with three-dimensional (3D) models of buildings or heritage sites enable access to a much broader community. Technologies such as Web3D, Augmented Reality (AR) or Virtual Reality (VR) can then be used to enhance the understanding of the virtual 3D environment.

This research aims to investigate and prototypically develop access to such repositories via a spatiotemporal localization of historical photographs and maps within a 3D model of the city center of Dresden/Germany. The project is conducted by nine researchers with various disciplinary backgrounds such as information science, education technology, art and architectural history, photogrammetry, media informatics, and computer science. The central component is a web-based four-dimensional (4D) browser, which includes temporal data in addition to the 3D spatial information of the media and building objects. The spatial visualization of location-based media and objects via a 3D city model is intended to enable access to urban history information as simply and intuitively as possible to an extended circle of users. That means, that browsing the data is not limited to a search bar and its output, but gets expanded into four dimensions using 3D navigation methods and a responsive time slider. A tackled scenario is to provide the interface for largescale collections of urban images. Backed by spatial information, diverse visualization methods enable investigation of, e.g., the statistical distribution of images in respect to their orientation or the depicted objects.

If all objects are located in a higher-level coordinate reference system, the 3D model should serve as the basis for a location-dependent AR representation on mobile devices. Our research group aims to provide tools for Cultural Heritage education in different settings using AR technology. The focus is on the communication of research results concerning historical photography of architecture to the general public through exhibitions, as well as guided and unguided tours. Bringing geo-located photography and historic models into the current cityscape through hand-held AR, tourists are enabled to engage with the historical situation of the photographer. In addition to location-based AR, tools for the exploration of large amounts of photographic documents in exhibitions are developed, combining 3D printed models of architecture with geo-located photographs in handheld as well as see-through AR settings. Using the computed spatial orientation of photographs with respect to the respective 3D models of buildings, 3D printed models can be augmented with historical images to provide historical textures.

Hence, one main aspect for the different presented access strategies is the precise selection of relevant images and their spatialization in a 3D space. The historical photographs origin from the photo library of the Saxon State and University Library Dresden (SLUB), which contains about 1.8 million images of over 80 institutions at this point in time. Access to the images is granted via the website (`http://deutschefotothek.de`), which provides keyword search and additional content filters like photographer, depicted people and topic.

Our research group wants to extend these conventional search possibilities using machine learning methods regarding the visual content of the images. In cooperation with the Competence Center for Scalable Data Services and Solutions (ScaDS) different deep learning approaches were tested for the filtering of historical images.

Once filtered, the images have to be oriented in 3D space in relation to the depicted 3D city models. For this purpose, different photogrammetric methods were applied on the images and models. An interactive Direct Linear Transformation (DLT) and manual placing of images in 3D space have already been tested in the prototype application. Since a completely automated process is preferable, the photogrammetric work focuses on the relative orientation and thus an optimized Structure-from-Motion (SfM) workflow for historical images. Difficulties arise due to different camera types, acquisition techniques, image media, digitization artefacts, and many more. Especially for feature detection and matching, the common used methods have to be adapted.

This article aims to provide an overview of parts of current research and development work in the ongoing interdisciplinary research project. It shows how multimodal access especially on image repositories has been realized in the past and how it will be implemented in the future. The workflow from conventional repositories as databases over filtering and spatializing the images to presentation and interactive browsing in a web application as well as in a VR/AR environment are presented. The main focus is on the photogrammetric works, since these are the key part for the orientation and positioning of images in the different applications.

Some approaches and ideas have already been described in Maiwald et al. (2019a) but are now updated and extended by more recent work results.

3.2 Multimodal access on repositories

There are a lot of different ways to get access to multimodal data resources such as plans, maps, images, etc., from different points in time. Conventional interactive access to repositories with the requirement of being on site stands in contrast to the virtual access using online collections. Listed under the collective term *virtual museums*, 3D web applications and technologies such as AR, VR, and Mixed Reality (MR) are considered and compared.

3.2.1 Conventional access

When working with specific historical data, it is often necessary to crawl archives and find source documents piece by piece, especially, since neither the digitization nor the annotation of huge data archives are fully completed. However, this requires an excellent knowledge of the topic and some information may not be found at all. Especially, the following four issues still persist in many libraries and archives (Evens and Hauttekeete, 2011):

- Complex process of digitization.
- Identification of correct metadata.
- Business models/financing.
- Intellectual property rights management.

This leads to the two main problems of finding and afterward publishing the relevant data. Even in recent research, it is often necessary to collect analog data and process it. For the purpose of reconstruction, the archive documentation research is often considered as the first step (Bitelli et al., 2017). One example is the modeling of the Larz Anderson Estate using mainly photographs, journals, and drawings scattered in several American archives (Ackerman and Glekas, 2017). It was also possible to reconstruct and texturize the collapsed dome of the San Pietro Church in Tuscania using an automatic SfM approach with 23 images (Beltrami et al., 2019), while the Great Mosque of Aleppo could be reconstructed with image data on 16 CD-ROMs (Grussenmeyer and Al Khalil, 2017). For the automatic 3D modeling of the Zwinger in Dresden, a dataset of over 800 archival images had to be reduced to around 50 images manually (Maiwald et al., 2017) for the subsequent SfM approach, and recently, parts of the Fortezza Vecchia could be recreated using archive images and plans from different analog sources (Bevilacqua et al., 2019).

Consequently, the traditional process of archive work from exploring and finding data to digitizing and publishing the sources is still present nowadays.

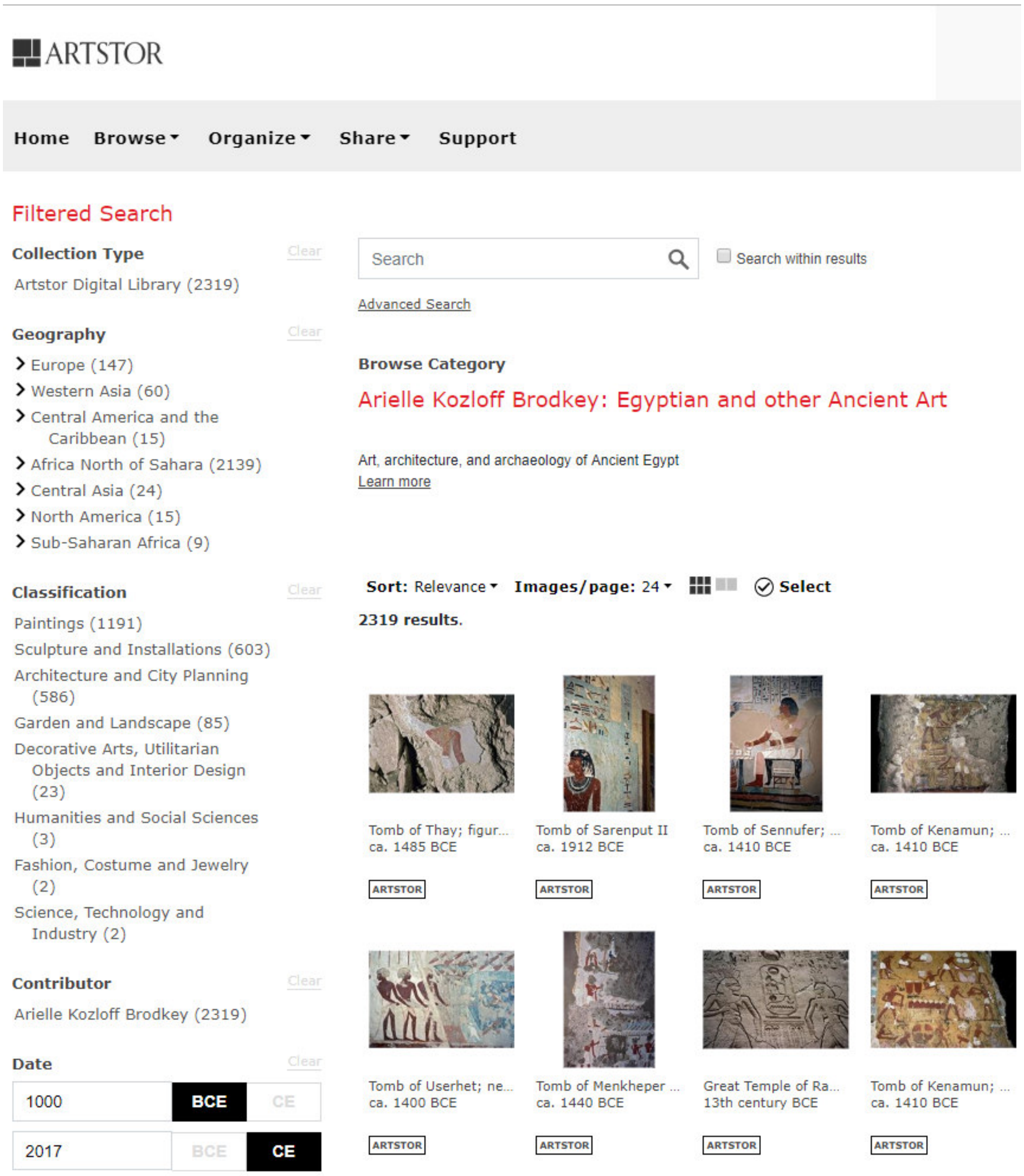
3.2.2 Virtual access using online collections

As already stated, digital information is and will be essential for automated workflows as well as the dissemination of information and knowledge in society. Therefore, Ross and Hedstrom (2005) proposed six reasons why digital preservation plays an important role in cultural heritage institutions. These are in brief:

- protection and conservation of cultural memory is a societal good;
- international scientific collaborations benefit from the availability of data repositories;
- accountability of those institutions;
- re-use of digital information as an economic benefit;
- effective and affordable strategies as a move from an industrial to a knowledge economy;
- sustainable digital libraries depend upon the availability of preservation tool and services.

When looking at today's archives, data is usually presented in one or two dimensions. Hence, objects, books, photographies, and other information is displayed as text or can be depicted as images. Examples for comprehensive online collections are Europeana (`https://www.europeana.eu`), Artstor (`https://library.artstor.org`) and Arachne (`https://arachne.dainst.org`), but there exist many more.

Usually, the websites providing access to these repositories show a search bar in which users can type keywords. Those lead to a resulting page (Fig. 3.1) where numerous filters (e.g. media classification, contributor, license, etc.) can be applied.



ARTSTOR

Home Browse ▾ Organize ▾ Share ▾ Support

Filtered Search

Collection Type Clear

Artstor Digital Library (2319)

Geography Clear

- Europe (147)
- Western Asia (60)
- Central America and the Caribbean (15)
- Africa North of Sahara (2139)
- Central Asia (24)
- North America (15)
- Sub-Saharan Africa (9)

Classification Clear

Paintings (1191)
Sculpture and Installations (603)
Architecture and City Planning (586)
Garden and Landscape (85)
Decorative Arts, Utilitarian Objects and Interior Design (23)
Humanities and Social Sciences (3)
Fashion, Costume and Jewelry (2)
Science, Technology and Industry (2)

Contributor Clear

Arielle Kozloff Brodkey (2319)

Date Clear

1000 BCE CE
2017 BCE CE

Search □ Search within results

Advanced Search

Browse Category

Arielle Kozloff Brodkey: Egyptian and other Ancient Art

Art, architecture, and archaeology of Ancient Egypt
Learn more

Sort: Relevance ▾ **Images/page:** 24 ▾ ✓ Select

2319 results.

Tomb of Thay; figur... ca. 1485 BCE ARTSTOR
Tomb of Sarenput II ca. 1912 BCE ARTSTOR
Tomb of Sennufer; ... ca. 1410 BCE ARTSTOR
Tomb of Kenamun; ... ca. 1410 BCE ARTSTOR
Tomb of Userhet; ne... ca. 1400 BCE ARTSTOR
Tomb of Menkheper ... ca. 1440 BCE ARTSTOR
Great Temple of Ra... 13th century BCE ARTSTOR
Tomb of Kenamun; ... ca. 1410 BCE ARTSTOR

Figure (3.1): Example search result for the category "Arielle Kozloff Brodkey: Egyptian and other Ancient Art" on `https://library.artstor.org`.

Nonetheless, challenges such as low data quality, difficult access, usability problems, and poor availability of information in certain areas of interest still persist today (Münster et al., 2018). Considering the search bar, metadata plays an important role in finding the relevant object or data file but also limits the search possibilities. Problems arise when the metadata is missing, incorrect, or insufficient. Then, it will not be possible for a user to find the specific object. Filters can help to narrow the search query but are often not enough to simplify the results.

Additionally, different types of users are following diverse search strategies and for most of the users, Google is still the main entrance for search queries (Beaudoin and Brady, 2011). It has been investigated by different research groups that a beginner uses far more simple terms than an expert for metadata-based search queries and that online data retrieval and access is still a challenge for many people (Beaudoin and Brady, 2011; Fleming-May and Green, 2016; Münster et al., 2018; Yoon and Chung, 2011). In the presented research, access to historical images is granted via the photo library of the SLUB, where similar issues are present. These issues are meant to be investigated and improved during the project. In addition, the mentioned repository holds maps, drawings, and scaled plans, which are also valuable for diverse multimodal representation strategies in informative models and systems.

3.2.3 Virtual museums

Since the junior research group develops various applications resulting out of multimodal online repositories, this chapter presents related work in the field of cultural heritage summarized under the term *virtual museum.*

While there is no standard definition for this term, the Virtual Multimodal Museum (`https://www.vi-mm.eu`) states that "a virtual museum (VM) is a digital entity that draws on the characteristics of a museum, in order to complement, enhance, or augment the museum through personalization, interactivity, user experience and richness of content". Styliani et al. (2009) identified four exhibition types as:

- Web3D exhibitions;
- VR exhibitions;
- AR exhibitions;
- MR exhibitions;

which shall be defined and presented with recent examples in the field of cultural heritage in the following. Considering the fact that also a two-dimensional (2D) representation of an exhibit can be regarded as a virtual museum the research in this chapter focuses on examples in at least 3D.

A Web3D exhibition could be defined as a virtual museum that is made accessible through the World Wide Web platform and browser independent. Advantages of such a representation, especially concerning the field of cultural heritage, are (Slater and Sanchez-Vives, 2016):

- Worldwide access and exploration of cultural sites.
- Preservation by digitization.
- Restoration and experience of cultural sites.
- Modeling of cultural sites under different future conditions.

Originally defined for VR applications, these advantages also apply to Web3D exhibitions of cultural sites and objects.

Web3D emerged especially due to the rise and distribution of fast Internet connections and additionally the development and maintenance of standards. Starting with the Virtual Reality Modeling Language (VRML) in 1997 by the Web3D Consortium (http://www.web3d.org), this standard has been replaced by the Extensible 3D (X3D) graphics and Humanoid Animation (H-Anim) standards, which are open, free, extensible, and interoperable.

This allows the creation of different applications such as 3D WebGIS for analysing the ancient Maya kingdom of Copan, Honduras (Girardi et al., 2013) or the visualization of high resolution 3D models from Andalusian universities' cultural heritage in the project Atalaya3D

(Melero et al., 2018). In contrast to the depicted technology, the 4D cities project used Java and a Java applet to grant web-based access on time-varying city models (Schindler and Dellaert, 2012) related to the presented research.

In addition, the digitization and visualization of cultural heritage play an important role in the field of Web3D exhibitions. Examples are various projects of CultLab3D (`https://www.cultlab3d.de`), models of the 3D-ICONS project (Koutsoudis et al., 2015), and of the INCEPTION project (Maietti et al., 2018). While it is easier to distinguish Web3D and extended reality applications, the terms VR, AR, and MR are more difficult to be separated. A lot of researchers see VR as the possibility to navigate and interact (selecting and moving objects) in a 3D environment (Gutierrez et al., 2008; Guttentag, 2010), with the addition that the user is immersed in a completely virtual environment (Milgram et al., 1995). AR is often seen as a type of VR (Burdea and Coiffet, 2003; Guttentag, 2010), but should be in fact classified as an independent term. The user sees the real environment through, e.g. a see-through head-mounted display (HMD) and additional virtual information are added to the real world (Milgram et al., 1995). MR in the scheme of the Reality-Virtuality Continuum (Fig. 3.2) can be seen as everything between the two extrema, completely real environment, and completely virtual environment (Milgram et al., 1995; Styliani et al., 2009).

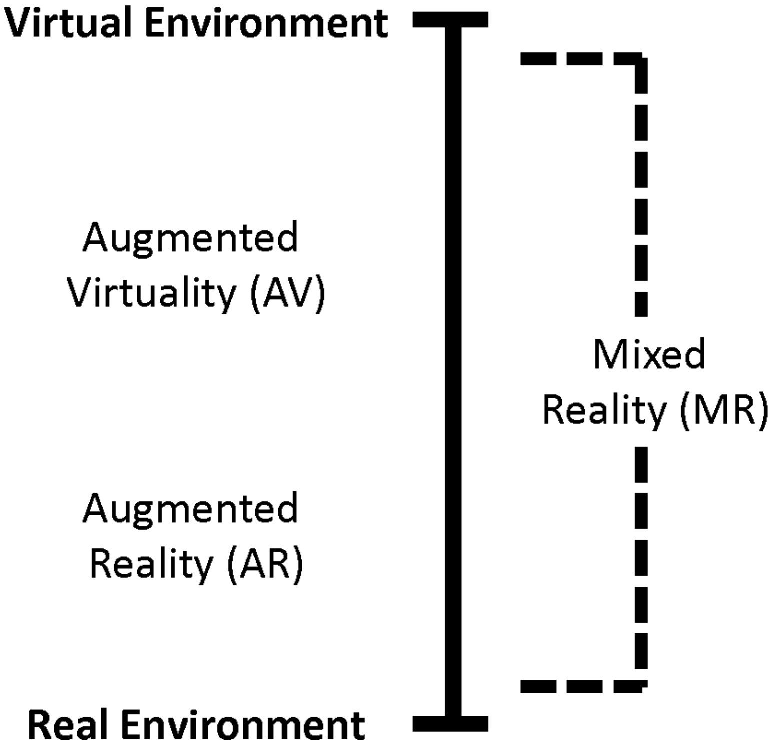


Figure (3.2): Representation of the Reality-Virtuality Continuum according to Milgram et al. (1995).

Some recent examples are a VR application of an Ottoman fortress (Tschirschwitz et al., 2019), research concerning the effectiveness of the AR application in Augusta Raurica (Armingeon et al., 2019), or the MR application of the reconstruction of one building of the Bergen-Belsen concentration camp (Oliva et al., 2015).

3.3 Workflow and access strategies

3.3.1 Overview

This section presents strategies for a completely automatic workflow from a repository of historical images up to different 4D applications using web technologies and VR/AR. Some of the methods are already implemented, while others still have to be evaluated for achieving superior results (Fig. 3.3).

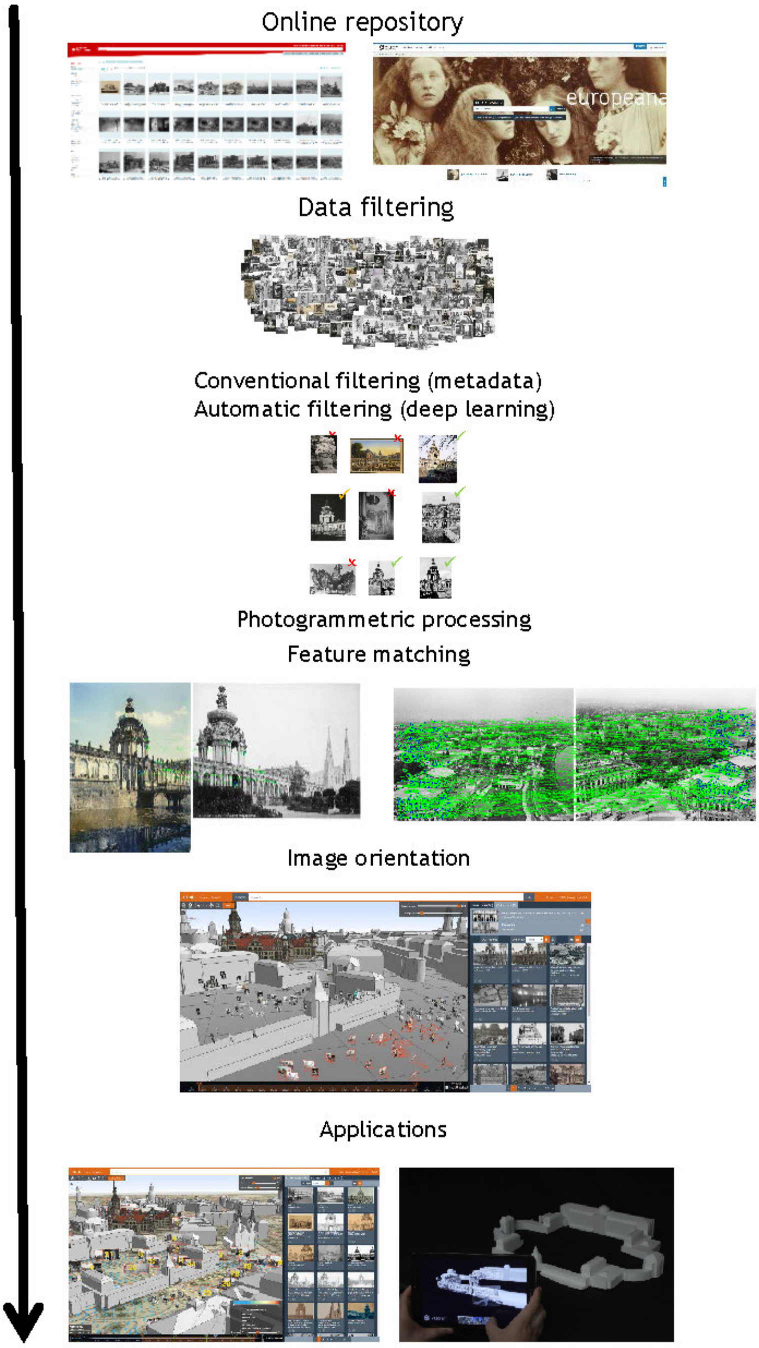

Online repository
europeana
Data filtering
Conventional filtering (metadata)
Automatic filtering (deep learning)
Photogrammetric processing
Feature matching
Image orientation
Applications

Figure (3.3): Overview of the presented workflow from an online repository to applications presenting informative models.

The starting point of the applications is an online repository containing mainly historical images but also drawings, maps, and plans. The workflow proceeds with the filtering of the relevant data. Images that are not desirable are, e.g., stereo images, close range and indoor scenes, and images, where people are the main motive.

Filtering can be carried out manually by keyword search and a further selection of several images, but due to the already stated problems when using metadata (see 2.2), a completely automated image retrieval is preferable. Thus, machine learning approaches are also used for filtering.

The resulting images are oriented using a broad range of photogrammetric methods. The main goal is the relative orientation of a large image mosaic which is already in the same (global) coordinate system as the application or has to be transformed using automatically retrieved parameters for a Helmert transformation. Therefore, a historical image data set has been created and is used for the evaluation of different feature detection and matching methods.

Oriented images, 3D city models, and maps are spatialized and made accessible in the 4D web browser application (`http://4dbrowser.urbanhistory4d.org`) of the research group where the user can jump into the perspective of the photographer. Additionally, different points in time with changing building models and maps can be visited. Thus, filtering of images can be carried out spatially and temporally leading to representation in four dimensions.

Different AR and VR applications are developed in the project where the oriented images and the 3D city models are used. HMD-based fully immersive VR with gamification aspects as well as handheld AR is employed to allow users free exploration of historical photography in a spatial setting.

3.3.2 Filtering

When working in large image repositories, a first step is to filter the images which are suitable for the use of photogrammetric methods. Especially close-up, indoor or portrait photographs should be neglected.

Filtering online collections is still considered a difficult task, because training a deep learning network often requires a large amount (500-1000 images per category) of training data (Deng et al., 2009). Whereas, for general object categories such as buildings a lot of training images exist, for a more concrete category like a specific building in a city, there is mostly no annotated data. A second problem is the changing nature and style of illustrations and pictures over different epochs, e.g. as drawings, paintings, photos, etc. Note that from the perspective of a neural net there is already a difference between pictures from a mobile phone and a normal camera.

Recent experiments with pre-trained networks lead to promising results, as these have to be slightly modified only in the upper layers. The advantage is three-fold: the training time is reduced drastically, the neural net has a structure that is already proven for the task and training is possible with smaller amounts of data. Thereby, the crucial requirement is to find an appropriate pre-trained network. For the application at hand, the pre-trained networks VGG16 and VGG19, which are deep convolutional neural nets for image classification, were used (Simonyan and Zisserman, 2014).

Only the dense layers on top of the network are modified and retrained, while the convolutional layers are used with their pre-trained weights. More precisely, on top of the pre-trained convolutional part of the network two dense layers were added: one consisting of 256 units (with a ReLU activation function) and the top layer consisting of one unit (with a sigmoid activation function). The next step could be a hyperparameter optimization, e.g. targeting the number of units in the penultimate layer.

The overall accuracy rate was approximately 70-80%. Nevertheless, one issue still is a high false-negative rate (approximately 40%), i.e. pictures of interest were classified as irrelevant. The current work in progress tackles especially this issue by dividing the whole pipeline into single modules that are optimized separately. For example, one module classifies a picture as “building/architecture” or “other”. Within these pre-filtered pictures, another module classifies as “building of interest” or “irrelevant building”.

One further important aspect is the assessment and quantification of uncertainty of the estimated accuracy and error-rates, especially as the input datasets are small. This is realized with statistical methods such as resampling methods, e.g. bootstrapping (Efron and Tibshiran, 1994).

Through the whole optimization, it is important to make the decision process of the convolutional neural networks visible. This is realized by means of so-called heat maps, that identify areas of most relevance for the final classification decision. More precisely, classification activation maps are used (Selvaraju et al., 2017).

At the moment, the output of the neural network helps to classify the images, but a manual selection has to be performed afterward to remove the outliers, which are not useful for the following methods.

3.3.3 Photogrammetry

The filtered images need to be oriented and georeferenced in order to be placed in the applications. While it was possible to run a complete SfM procedure using Agisoft PhotoScan v. 1.3.1 for a small subset (11 of 499) of images, the major number of images from the repository could not be integrated into such a workflow (Maiwald et al., 2017).

Figure (3.4): Historical imagery of urban sites in Dresden, (Germany): a) Dinglinger House, b) Georges’s Gate, and c) Crown Gate of the Zwinger.

Most of the times the first step of orienting the images failed due to a lack of homolog points found. Other existing applications such as VisualSFM (Wu, 2013), Meshroom (Moulon et al., 2013), and Agisoft Metashape produced similar results for various historical image sets (Fig. 3.4) as shown in Table 3.1.

Table (3.1): Comparison of three different SfM softwares used on three historical image data sets. The table shows respectively the used modes, average feature points found, tie points used for the creation of a sparse cloud and finally the number of oriented images in comparison to total images.

	George's Gate	Zwinger	Dinglinger House
Software		**Agisoft Metashape (proprietary)**	
Mode		Highest Accuracy, Generic Preselection	
Average feature points found	39127	214919	39354
Tie points (Sparse cloud)	3109	3027	1968
Oriented images	17/32	3/34	5/14
Software		**Meshroom (open source)**	
Mode		Ultra Accuracy, SIFT descriptor	
Average feature points found	4251	26642	3877
Tie points (Sparse cloud)	3651	4282	232
Oriented images	19/32	14/34	3/14
Software		**VisualSfM (open source)**	
Mode		Default Accuracy, SIFT descriptor	
Average feature points found	8966	8893	9187
Tie points (Sparse cloud)	272	128	49
Oriented images	2/32	4/34	2/32

Even though Agisoft Metashape and Meshroom were able to find a solution for the orientation of around half the images of the George's Gate a visual control of the cameras revealed that the calculated orientations were most of the times incorrect. A significantly smaller number of correct orientations (around 5) has to be assumed for this building, similar to the others. Sometimes tie points for multiple models have been created but could not be merged into a single building model.

Several reasons negatively effecting the keypoint detection and matching could be determined amongst others as image differences due to (Maiwald, 2019):

- original image medium;
- digitization;
- different camera models;
- acquisition time;
- acquisition technique,

resulting in image artefacts, scene occlusion, and radiometric variations.

Additionally, image pairs may have, on the one hand, very similar perspectives inconvenient for a photogrammetric reconstruction, or, on the other hand, large viewpoint changes hampering the common feature matching approach based on the Scale Invariant Feature Transform

(SIFT) (Lowe, 2004) and Speeded-Up Robust Features (SURF) (Bay et al., 2006). One approach for buildings, which still exist, could be the generation of a strong image network by using contemporary images. This increases the reliability and the number of matching points between the historical images and the recent images (Barazzetti et al., 2013; Maiwald et al., 2017).

For destroyed or lost buildings, experiments using exclusively geometrical features show that orientation is possible for single image pairs if the feature detection and matching step is individually observed (Maiwald et al., 2018).

Thus, the feature detection and feature matching step of the end-to-end SfM tools has to be broken up, since most of the times SIFT or a variation of it is used. Testing different feature matching methods requires a dataset with ground truth orientation. Since no dataset using exclusively historical images was available, an open source image dataset (`https://dx.doi.org/10.25532/OPARA-24`) oriented using the properties of the Trifocal Tensor was created (Maiwald, 2019) (Fig. 3.5).

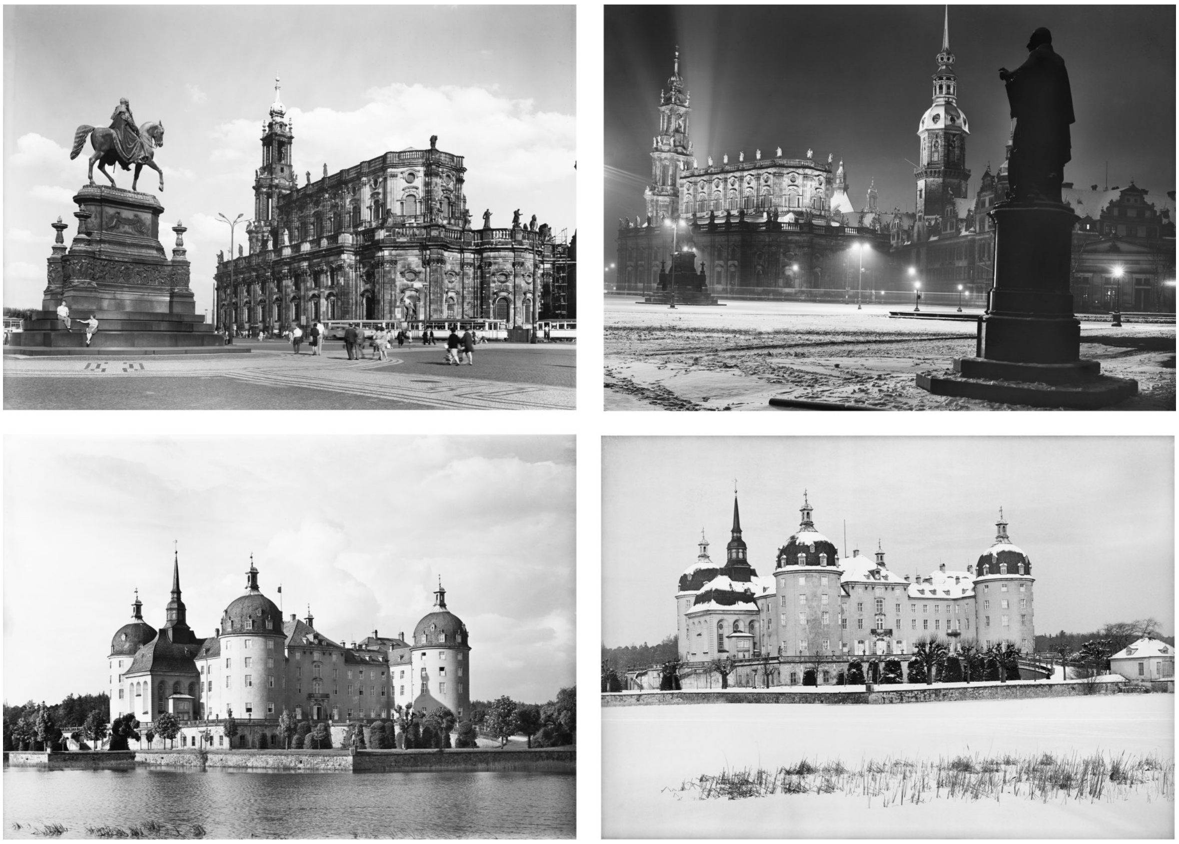


Figure (3.5): Examples of image pairs of the proposed oriented dataset. The image pair in the top row (Hk_1/2_3) was very difficult to be matched, while the bottom pair (Mb_1/1_3) had 100% correct matches by two different methods.

Evaluating some already successful feature matching methods such as Maximally Stable Extremal Regions (MSER) (Matas et al., 2004) and Radiation-Invariant Feature Transform (RIFT) (Li et al., 2018) showed that feature matching between historical image pairs is difficult but sometimes possible. It has been identified that most of the times the outlier removal method seems to be the crucial point. Especially, the outlier removal method Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981) can be improved if there is only a small number of correct matches out of total matches (also referred to as matching score). Alternatives are, e.g., Progressive Sampling Consensus (PROSAC) (Chum and Matas, 2005) or Fast Sample Consensus (FSC) (Wu et al., 2015). The number of correct matches can be determined using the properties of three-view geometry and the already given Trifocal Tensor.

Recent results using Matching On Demand with View Synthesis (MODS) (Mishkin et al., 2015a) and RIFT with advanced outlier removal could improve the matching score significantly and thus, could be integrated into the image orientation of a SfM workflow (Fig. 3.6).

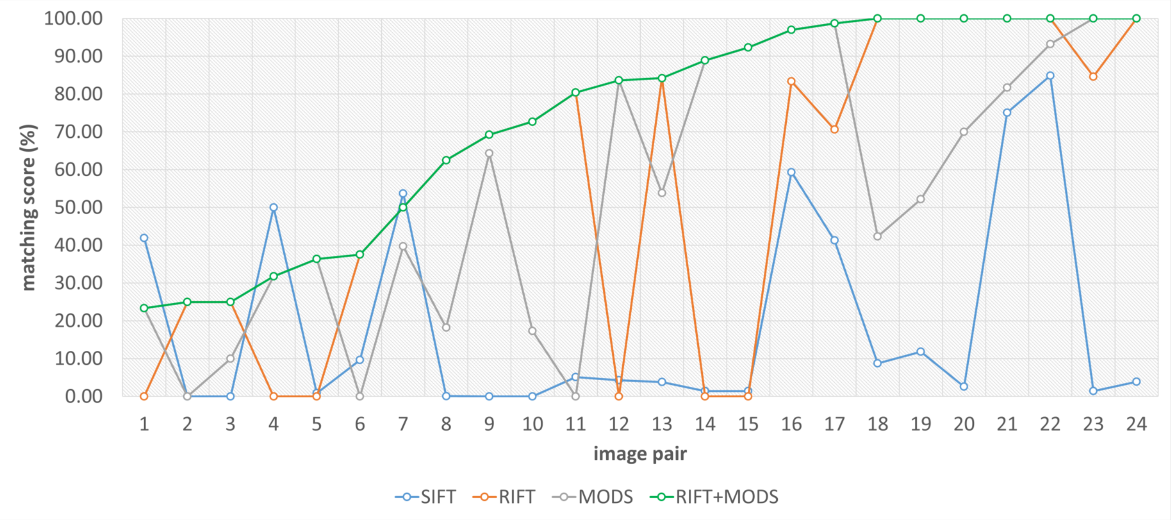


Figure (3.6): Ascending matching score for three different feature matching methods used on 24 image pairs of a benchmark dataset. The combination of RIFT and MODS showed the best results on the historical images.

Especially, the combination of both methods can be useful to find matches between a lot of different historical image pairs (Tab. 3.2).

It is planned to add single historical images consecutively, compare them to the already existing oriented images and try to find matches using the shown methods. The accuracy of the image orientation is improved by running a subsequent bundle adjustment. Since SfM is a scale-invariant method the retrieved image mosaics have to be scaled, generally performed with scale bars or field measurements. Usually, this is not possible when using historical images. Thus, in the web application, the relatively oriented image mosaic has to be adjusted to fit into the scene.

One possible implementation is an interactive orientation using homolog points between a detailed 3D building model (possibly with an underlying historical map) and a single image. The orientation of the image can be estimated using a minimum of six points and the DLT (Hartley and Zisserman, 2003). In the following, the relative orientations can be used for placing the whole image mosaic.

Another approach could be the use of gamification elements implemented in the VR application (see Section 3.3.5). Users are engaged to locate the position of the photographer of

Table (3.2): Matching score (%) for image pairs of the proposed image data set processed using the different feature matching methods SIFT, RIFT and MODS.

Image pair	SIFT	RIFT	MODS	RIFT+MODS
So_1/1_3	41.89	0.00	23.36	23.36
Hk_1/2_3	0.00	25.00	0.00	25.00
So_2/2_3	0.00	25.00	10.00	25.00
So_1/2_3	50.01	0.00	31.76	31.76
So_2/1_2	0.88	0.00	36.36	36.36
Hk_1/1_3	9.62	37.50	0.00	37.50
Zw_1/2_3	53.72	50.00	39.71	50.00
Zw_2/2_3	0.08	62.50	18.18	62.50
Zw_2/1_2	0.00	69.26	46.39	69.23
Mb_2/2_3	0.00	72.70	17.31	72.70
Mb_1/2_3	5.10	80.43	0.00	80.43
Hk_2/1_2	4.30	0.00	83.64	83.64
Hk_2/2_3	3.80	84.21	53.85	84.21
Mb_2/1_2	1.40	0.00	88.88	88.88
Mb_2/1_3	1.40	0.00	92.31	92.31
Zw_1/1_3	59.35	83.33	96.97	96.97
So_2/1_3	41.30	70.59	98.68	98.68
Hk_1/1_2	8.77	100.00	42.31	100.00
Zw_2/1_3	11.85	100.00	52.17	100.00
Mb_1/1_2	2.60	100.00	70.00	100.00
So_1/1_2	75.06	100.00	81.73	100.00
Zw_1/1_2	84.87	100.00	93.22	100.00
Hk_2/1_3	1.40	84.62	100.00	100.00
Mb_1/1_3	3.90	100.00	100.00	100.00

a historical image in relation to a 3D model competitively against other users. The resulting image orientations could be averaged (including outlier removal) and used as a prior orientation for completely automatic workflow.

Approaches for automatic registration of 3D models and images are already investigated in many different studies (Callieri et al., 2008; Franken et al., 2005; Paudel et al., 2015) and will be tested for reliability in the future.

3.3.4 Browser access

The registered images can be viewed in a 4D browser-based prototype application (`http://4dbrowser.urbanhistory4d.org`). As an entry point, the user is lead to the starting page including information on the project, the data and navigation through the platform. Additionally, a help site is accessible explaining all the available tools and features. When starting to explore the data, a contemporary 3D city model and a historical map are displayed in a 3D viewport that forms the main part of the graphical user interface (GUI). The viewport also shows all the images hierarchically clustered with respect to their location using single-linkage clustering (Fig. 3.7).

Observing a single image allows jumping into the perspective of the photographer using the predefined orientations. This enhances amongst others the understanding of the position of the photographer and allows the uncovering of historical details (Schindler and Dellaert, 2012).

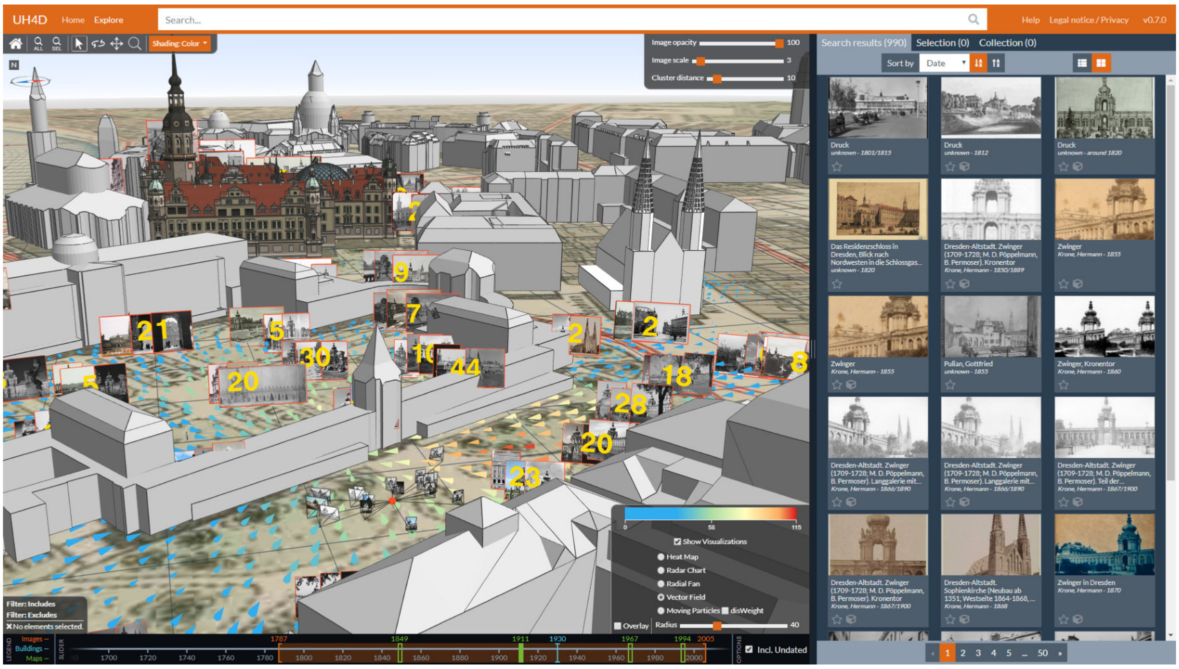


Image opacity 100
Image scale 3
Cluster distance 10

☑ Show Visualizations
○ Heat Map
○ Radar Chart
○ Radial Fan
● Vector Field
○ Moving Particles □ disWeight
□ Overlay Radius 40
☑ Incl. Undated

Druck
unknown - 1801/1815

Druck
unknown - 1812

Druck
unknown - around 1820

Das Residenzschloss in Dresden, Blick nach Nordwesten in die Schlossgas...
unknown - 1820

Dresden-Altstadt. Zwinger (1709-1728; M. D. Pöppelmann, B. Permoser). Kronentor
Krone, Hermann - 1850/1889

Zwinger
Krone, Hermann - 1855

Zwinger
Krone, Hermann - 1855

Pulian, Gottfried
unknown - 1855

Zwinger, Kronentor
Krone, Hermann - 1860

Dresden-Altstadt. Zwinger (1709-1728; M. D. Pöppelmann, B. Permoser). Langgalerie mit...
Krone, Hermann - 1866/1890

Dresden-Altstadt. Zwinger (1709-1728; M. D. Pöppelmann, B. Permoser). Langgalerie mit...
Krone, Hermann - 1866/1890

Dresden-Altstadt. Zwinger (1709-1728; M. D. Pöppelmann, B. Permoser). Teil der...
Krone, Hermann - 1867/1900

Dresden-Altstadt. Zwinger (1709-1728; M. D. Pöppelmann, B. Permoser). Kronentor
Krone, Hermann - 1867/1900

Dresden-Altstadt. Sophienkirche (Neubau ab 1351; Westseite 1864-1868, ...
Krone, Hermann - 1868

Zwinger in Dresden
Krone, Hermann - 1870

Figure (3.7): Graphical user interface of the 4D browser prototype showing image clusters, a historical map, 3D models, a visualization of orientations and the listed search results.

The opacity of the images can be regulated revealing the underlying 3D building model and eventually detect changes in the object geometry (Bruschke et al., 2017). Further information (author, date, owner, etc.) belonging to the respective image can be displayed in an additional window. Multiple images can be added to a collection allowing comparison and, in the future, the user dependent recording of scientific workflows.

Searching for special images can be done using a conventional search bar and the respective metadata. In addition, the browsing for images is enhanced by directly showing the output of the search request in the 3D viewport. Thus, the user is not only limited to a single output list used in classical online repositories, but is able to directly verify the request using the 3D view.

Furthermore, one can narrow the image search using the timeline on the bottom of the application adding the 4D aspect. The timeline is not only applicable for historical images, but also for the 3D models and for currently four underlying historical maps. Another search possibility is the filtering of images by depicted buildings by selecting the corresponding 3D objects. It is planned to generate additional 3D building models for different points in time manually and by using photogrammetric methods.

In regard to the usage by experts such as art and architectural historians, advanced statistical visualization methods can be applied to the images to support answering specific research questions (Bruschke et al., 2018b). This includes conventional heat maps but also innovative visualization methods that consider the orientation of the images (Fig. 3.8).

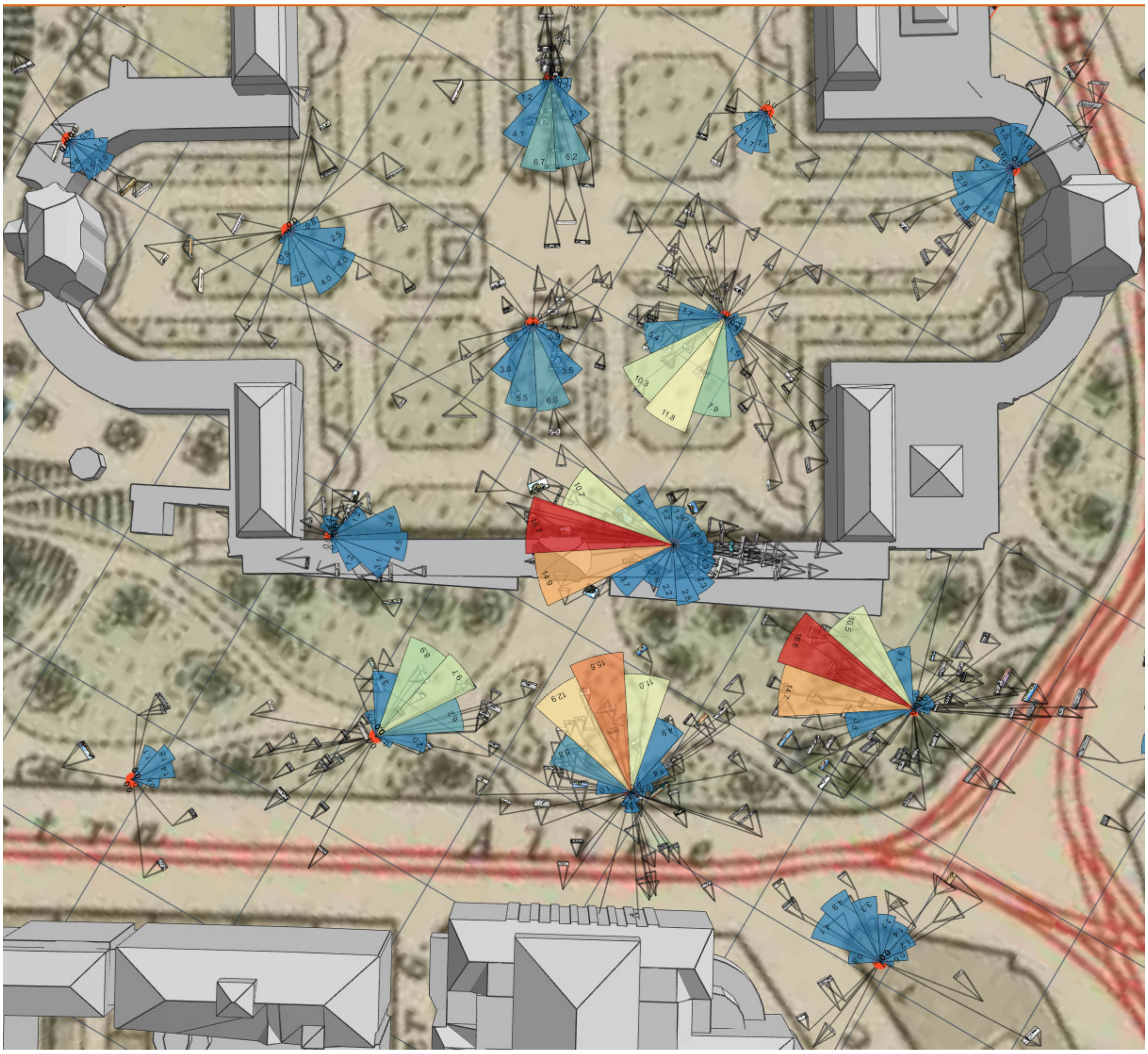


Figure (3.8): “Fan” visualization of image orientations in the 4D browser.

The browser application utilizes AngularJS as a basic front end web framework to build single-page applications (SPAs). Regarding 3D graphics, three.js is used as a higher-level API for WebGL. The 3D models are stored in the OpenCTM format that features a high compression rate and enables fast transmission. Owing to the SPA approach, all loaded 3D content is kept in the background when the user switches to another subpage.

When switching back to the main interactive 3D view, the 3D content has not to be loaded again and can be displayed instantly. The back end is a REST API on the basis of Node.js and Express.js. It handles not only all database requests, but also routines to automatically retrieve metadata from the original image repository. The data itself is stored compliant to the CIDOC Conceptual Reference Model (CRM). This results in strongly connected data that is predestined to be stored not in a relational database but in a graph database, in this case Neo4j (Bruschke and Wacker, 2014).

3.3.5 VR and AR access

The registered images are also used in different prototypical AR and VR applications. In a HMD-based fully immersive VR environment users are challenged to find the specific locations and orientations of multiple images. One scenario for the dissemination of Cultural Heritage research results includes the use of gamification elements, where several users are encouraged to replicate the position and orientation of photographers of historical images faster than their opponents. After finding the positions of some already spatialized images with respect to a full-scale 3D model, the users could be motivated to help to find and improve the location of unregistered as well as only roughly registered images (Fig. 3.9).

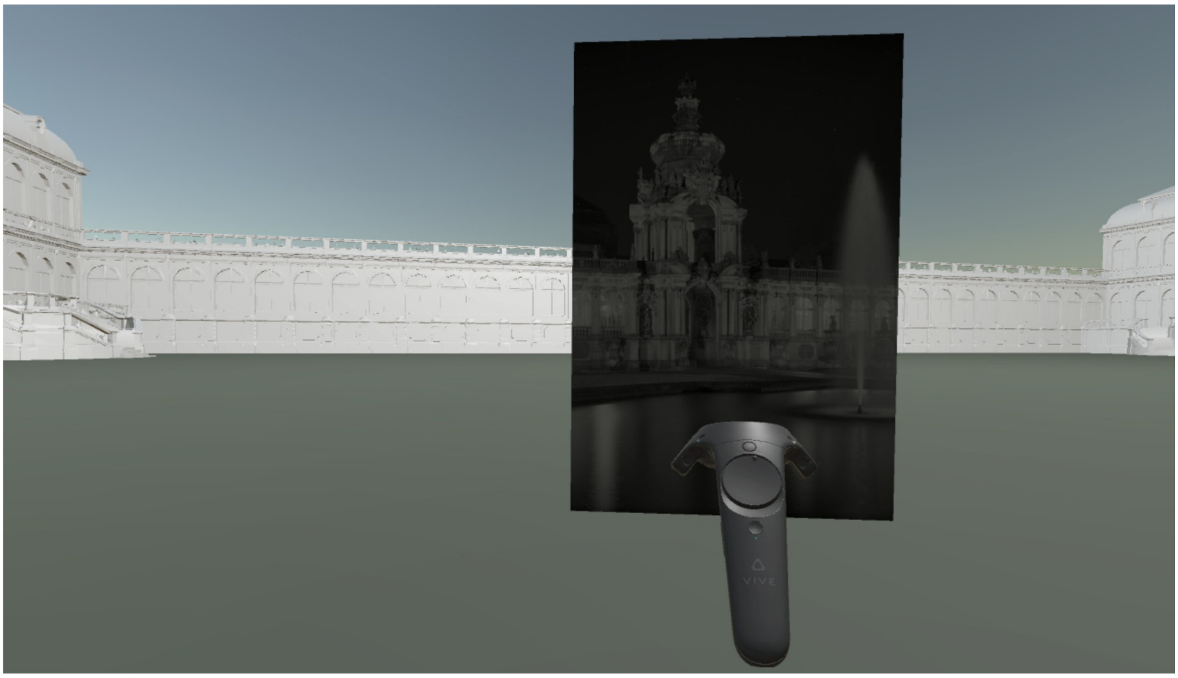


Figure (3.9): Fully immersive VR environment where users are challenged to find the location and orientation of historical images.

In addition, historic textures for registered models of architecture are created by UV mapping vertices of the respective 3D model to the registered historical photography in both VR and AR. The usage of photos as textures enables us to use coarse 3D models of buildings acquired from the city municipality, as details are provided by historical high-resolution photos.

The UV coordinates are calculated by casting rays from the known position P of the photographer to each vertex V in the model, through the projective plane of the currently processed image I. To verify visibility of vertices in the geometry of buildings from position P, we perform collision detection of rays with models in the scene. Parts of a building that are not contained in I, because they are either outside of the view frustum of the camera or hidden by other structures, cannot be textured with information contained within I and are left blank.

Employing photos as textures allow for the usage of very coarse 3D models of buildings, as details are provided by comparably high resolution photos.

In addition to the usage in VR environments, we use the introduced photo projection method to create AR installations combining 3D printed models of architecture with historical photography using 10.1” Android tablets (Niebling et al., 2018). Tracking of the model is provided by Vuforia Model Targets with Advanced Recognition in a Unity application combining the digital model of the building with a media repository of spatially registered historical photography (Fig. 3.10).

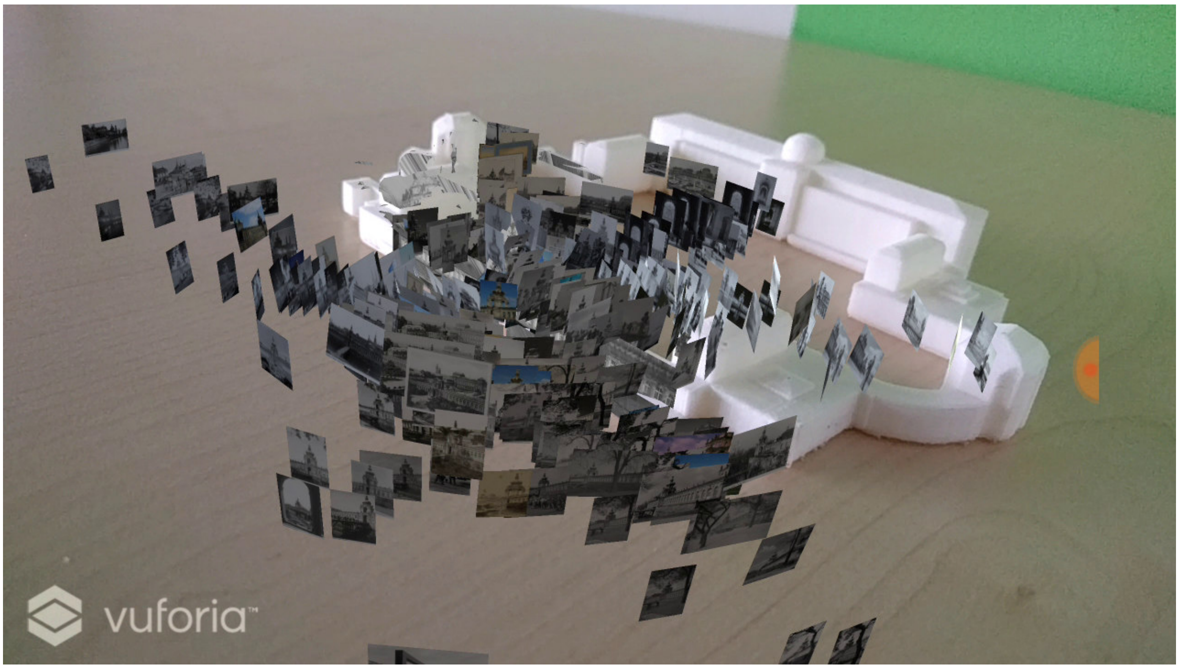


Figure (3.10): AR application blending registered images in front of a 3D printed model of the Zwinger.

In the AR installation, the textured digital model is rendered on top of the video of the tracked physical model, discarding parts that are not contained in the respective photo selected by the user. Spatial movement of the tablet can be performed to view buildings from a perspective diverging from the perspective of the photo. The user is then still able to perceive the historical appearance of the visible parts of the building depicted in the historical image. In addition, the handheld AR devices can be used to spatially access and browse the catalog of images in AR allowing selection of photography according to the positions of the spatially registered documents. Pedagogical approaches to enhance user's experiences and knowledge transfer are currently investigated (Niebling et al., 2017).

3.4 Conclusions

The contribution shows in the first part that there are still a lot of different ways to access archival data. While digital methods are on the rise, the conventional analog access to repositories is nonetheless used for recent research. Especially, the issue that only a fraction of all museum objects is digitized at the moment enables the searching success only for experts of the respective topic.

But, access will expand in the future due to increasing digitization and the use of digital technologies. Hence, different access strategies from the digital online repository up to the virtual museum are presented and compared. An increased amount of Web3D and AR/VR exhibitions could be demonstrated. Recent examples show that advantages in these technologies are, e.g., preservation, protection, accessibility, sustainability, and reusability of the digitized data. 3D technologies and extended reality enhance the understanding of objects not only in cultural heritage.

The steps from a repository to these access methods are not only theoretically discussed. The contribution shows how a fully automatic workflow can look like and how it is prototypically developed in the project. Filtering a large amount of archival data is considered as a first step. While usually a manual selection of appropriate images is performed, this contribution also shows new technologies based on deep learning for the automatic choice of photographs. Historical images are very challenging for these technologies and thus, these methods have to be evaluated further but can be a suitable addition for completely automatic workflow.

Photogrammetric evaluation can be seen as the core of this research. The precise orientation of images is necessary for the different applications. For some image pairs, the orientation can be calculated automatically using end-to-end software packages such as Agisoft Metashape. Though, for many image combinations, the first step of the orientation fails due to the lack of homolog points in image pairs. Therefore, the use of advanced feature matching methods particularly MODS and RIFT is proposed. These methods performed superior in comparison to other approaches and serve as a basis for the perspective relative orientation of the images.

In the following, the global registration in relation to the 3D models can be carried out manually or using semiautomatic (DLT) and in the future, automatic methods.

The oriented images can be accessed in different applications. A 4D web browser is introduced as an alternative repository holding historical images, plans, buildings, and maps. It is possible to take the perspective of the photographer and view temporal changes of the city using 3D models and historical maps. The depicted visualization of orientations can be a useful tool for art and architectural historians.

Several examples on VR and AR applications provide different access methods for the images using a printed and a virtual 3D model enhancing the understanding of the scene. Ray casting can be used to texture the model using the historical image material and users are able to support the orientation of images in a prototypical gamification VR application.

Concluding, a completely automatic workflow from an image repository over filtering and photogrammetry to various access methods is outlined. In the future, the different parts should be connected in a single working environment or even attached to an existing repository to provide the different strategies to a broader range of users. While most of the examples were executed with archival data of the city of Dresden, the proposed methods could be integrated into other urban areas or even in rural sites combined with GIS.

Acknowledgments

The research upon which this paper is based is part of the junior research group UrbanHistory4D's activities which has received funding from the German Federal Ministry of Education and Research under grant agreement No 01UG1630.

This work was supported by the German Federal Ministry of Education and Research (BMBF, 01IS18026BA-F) by funding the competence center for Big Data "ScaDS Dresden/Leipzig".

References

Ackerman, A. and E. Glekas (2017). “Digital Capture and Fabrication Tools for Interpretation of Historic Sites”. In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W2, pp. 107–114. DOI: 10.5194/isprs-annals-iv-2-w2-107-2017.

Armingeon, M., P. Komani, T. Zanwar, S. Korkut and R. Dornberger (2019). “A Case Study: Assessing Effectiveness of the Augmented Reality Application in Augusta Raurica”. In: *Augmented Reality and Virtual Reality*. Springer International Publishing, pp. 99–111. DOI: 10.1007/978-3-030-06246-0_8.

Barazzetti, L., S. Erba, M. Previtali, E. Rosina and M. Scaioni (2013). “Mosaicking thermal images of buildings”. In: *Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection*. SPIE, pp. 1–8. DOI: 10.1117/12.2019985.

Bay, H., T. Tuytelaars and L. V. Gool (2006). “SURF: Speeded Up Robust Features”. In: *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, pp. 404–417. DOI: 10.1007/11744023_32.

Beaudoin, J. E. and J. E. Brady (2011). “Finding Visual Information: A Study of Image Resources Used by Archaeologists, Architects, Art Historians, and Artists”. In: *Art Documentation: Journal of the Art Libraries Society of North America* 30, 2, pp. 24–36. DOI: 10.1086/adx.30.2.41244062.

Beltrami, C., D. Cavezzali, F. Chiabrando, A. I. Idelson, G. Patrucco and F. Rinaudo (2019). “3D Digital and Physical Reconstruction of a Collapsed Dome using SFM Techniques from Historical Images”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W11, pp. 217–224. DOI: 10.5194/isprs-archives-xlii-2-w11-217-2019.

Bevilacqua, M. G., G. Caroti, A. Piemonte and D. Ulivieri (2019). “Reconstruction of lost Architectural Volumes by Integration of Photogrammetry from Archive Imagery with 3-D Models of the Status Quo”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 119–125. DOI: 10.5194/isprs-archives-xlii-2-w9-119-2019.

Bitelli, G., M. Dellapasqua, V. A. Girelli, S. Sbaraglia and M. A. Tinia (2017). “Historical Photogrammetry and Terrestrial Laser Scanning for the 3d Virtual Reconstruction of Destroyed Structures: A Case Study in Italy”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5/W1, pp. 113–119. DOI: 10.5194/isprs-archives-xlii-5-w1-113-2017.

Bruschke, J., F. Niebling, F. Maiwald, K. Friedrichs, M. Wacker and M. E. Latoschik (2017). “Towards browsing repositories of spatially oriented historic photographic images in 3D web environments”. In: *Proceedings of the 22nd International Conference on 3D Web Technology - Web3D ’17*. ACM Press, pp. 1–6. DOI: 10.1145/3055624.3075947.

Bruschke, J., F. Niebling and M. Wacker (2018b). “Visualization of Orientations of Spatial Historical Photographs”. In: *Eurographics Workshop on Graphics and Cultural Heritage*, pp. 189–192. DOI: 10.2312/GCH.20181359.

Bruschke, J. and M. Wacker (2014). “Application of a Graph Database and Graphical User Interface for the CIDOC CRM”. In: *Access and Understanding–Networking in the Digital*

Era. Session J1. The 2014 annual conference of CIDOC, the International Committee for Documentation of ICOM, pp. 1–3.

Burdea, G. and P. Coiffet (2003). *Virtual reality technology.* John Wiley & Sons. ISBN: 9780471723752.

Callieri, M., P. Cignoni, M. Corsini and R. Scopigno (2008). “Masked photo blending: Mapping dense photographic data set on high-resolution sampled 3D models”. In: *Computers & Graphics* 32, 4, pp. 464–473. DOI: 10.1016/j.cag.2008.05.004.

Chum, O. and J. Matas (2005). “Matching with PROSAC — Progressive Sample Consensus”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05).* IEEE, pp. 220–226. DOI: 10.1109/cvpr.2005.221.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei (2009). “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

Efron, B. and R. J. Tibshiran (1994). *An introduction to the bootstrap.* CRC press. ISBN: 0412042312. DOI: https://doi.org/10.1111/1467-9639.00050.

Evens, T. and L. Hauttekeete (2011). “Challenges of digital preservation for cultural heritage institutions”. In: *Journal of Librarianship and Information Science* 43, 3, pp. 157–165. DOI: 10.1177/0961000611410585.

Fischler, M. A. and R. C. Bolles (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24, 6, pp. 381–395. DOI: 10.1145/358669.358692.

Fleming-May, R. A. and H. Green (2016). “Digital innovations in poetry: Practices of creative writing faculty in online literary publishing”. In: *Journal of the Association for Information Science and Technology* 67, 4, pp. 859–873. ISSN: 2330-1635. DOI: https://doi.org/10.1002/asi.23428.

Franken, T., M. Dellepiane, F. Ganovelli, P. Cignoni, C. Montani and R. Scopigno (2005). “Minimizing user intervention in registering 2D images to 3D models”. In: *The Visual Computer* 21, 8-10, pp. 619–628. ISSN: 0178-2789. DOI: https://doi.org/10.1007/s00371-005-0309-z.

Girardi, G., J. von Schwerin, H. Richards-Rissetto, F. Remondino and G. Agugiaro (2013). “The MayaArch3D project: A 3D WebGIS for analyzing ancient architecture and landscapes”. In: *Literary and Linguistic Computing* 28, 4, pp. 736–753. ISSN: 0268-1145. DOI: https://doi.org/10.1093/llc/fqt059.

Grussenmeyer, P. and O. Al Khalil (2017). “From Metric Image Archives to Point Cloud Reconstruction: Case Study of the Great Mosque of Aleppo in Syria”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W5, pp. 295–301. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-2-W5-295-2017.

Gutierrez, M., F. Vexo and D. Thalmann (2008). *Stepping into Virtual Reality.* Springer-Verlag GmbH. ISBN: 9781848001176. URL: https://www.ebook.de/de/product/19205394/mario_gutierrez_f_vexo_daniel_thalmann_stepping_into_virtual_reality.html.

Guttentag, D. A. (2010). “Virtual reality: Applications and implications for tourism”. In: *Tourism Management* 31, 5, pp. 637–651. ISSN: 0261-5177. DOI: https://doi.org/10.1016/j.tourman.2009.07.003.

Hartley, R. and A. Zisserman (2003). *Multiple view geometry in computer vision.* Cambridge: Cambridge university press. ISBN: 0521540518. DOI: https://doi.org/10.1017/CBO9780511811685.

Koutsoudis, A., F. Arnaoutoglou, A. Tsaouselis, G. Ioannakis and C. Chamzas (2015). “Creating 3D Replicas of Medium-to Large-Scale Monuments for Web-Based Dissemination Within the Framework of the 3D-Icons Project”. In: *CAA2015*, p. 971.

Li, J., Q. Hu and M. Ai (2018). “RIFT: Multi-modal Image Matching Based on Radiation-invariant Feature Transform”. In: *arXiv:1804.09493*, pp. 1–14.

Lowe, D. G. (2004). “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60, 2, pp. 91–110. ISSN: 0920-5691. DOI: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Maietti, F., R. Di Giulio, E. Piaia, M. Medici and F. Ferrari (2018). “Enhancing Heritage fruition through 3D semantic modelling and digital tools: the INCEPTION project”. In: *IOP Conference Series: Materials Science and Engineering.* Vol. 364. IOP Publishing, p. 012089. ISBN: 1757-899X. DOI: https://doi.org/10.1088/1757-899X/364/1/012089.

Maiwald, F., F. Henze, J. Bruschke and F. Niebling (2019a). “Geo-Information Technologies for a multimodal Access on Historical Photograps and Maps for Research and Communication in Urban History”. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W11, pp. 763–769. DOI: 10.5194/isprs-archives-xlii-2-w11-763-2019.

Maiwald, F. (2019). “Generation of a Benchmark Dataset Using Historical Photographs for an Automated Evaluation of Different Feature Matching Methods”. In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-2/W13, pp. 87–94. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-2-W13-87-2019.

Maiwald, F., D. Schneider, F. Henze, S. Münster and F. Niebling (2018). “Feature Matching of Historical Images Based on Geometry of Quadrilaterals”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2, pp. 643–650. ISSN: 2194-9034. DOI: https://doi.org/10.5194/isprs-archives-XLII-2-643-2018.

Maiwald, F., T. Vietze, D. Schneider, F. Henze, S. Münster and F. Niebling (2017). “Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities”. In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 447–452. ISSN: 1682-1750. DOI: https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017.

Matas, J., O. Chum, M. Urban and T. Pajdla (2004). “Robust wide-baseline stereo from maximally stable extremal regions”. In: *Image and vision computing* 22, 10, pp. 761–767. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2004.02.006.

Melero, F. J., J. Revelles and M. L. Bellido (2018). “Atalaya3D: Making Universities’ Cultural Heritage Accessible Through 3D Technologies”. In: *Eurographics Workshop on Graphics and Cultural Heritage*, pp. 31–35. DOI: 10.2312/GCH.20181338.

Milgram, P., H. Takemura, A. Utsumi and F. Kishino (1995). “Augmented reality: a class of displays on the reality-virtuality continuum”. In: *Telemanipulator and Telepresence Technologies*. SPIE, pp. 282–292. DOI: `10.1117/12.197321`.

Mishkin, D., J. Matas and M. Perdoch (2015a). “MODS: Fast and robust method for two-view matching”. In: *Computer Vision and Image Understanding* 141, pp. 81–93. ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2015.08.005`.

Moulon, P., P. Monasse and R. Marlet (2013). “Adaptive Structure from Motion with a Contrario Model Estimation”. In: *Computer Vision – ACCV 2012*. Springer Berlin Heidelberg, pp. 257–270. DOI: `10.1007/978-3-642-37447-0_20`.

Münster, S., C. Kamposiori, K. Friedrichs and C. Kröber (2018). “Image libraries and their scholarly use in the field of art and architectural history”. In: *International Journal on Digital Libraries* 19, 4, pp. 367–383. DOI: `10.1007/s00799-018-0250-1`.

Niebling, F., J. Bruschke and M. E. Latoschik (2018). “Browsing Spatial Photography for Dissemination of Cultural Heritage Research Results using Augmented Models”. In: *Eurographics Workshop on Graphics and Cultural Heritage*, pp. 185–188. ISSN: 2312-6124. DOI: `10.2312/GCH.20181358`.

Niebling, F., F. Maiwald, K. Barthel and M. E. Latoschik (2017). “4D Augmented City Models, Photogrammetric Creation and Dissemination”. In: *Digital Research and Education in Architectural Heritage*. Digital Research and Education in Architectural Heritage. Cham: Springer International Publishing, pp. 196–212. ISBN: 978-3-319-76992-9. DOI: `https://doi.org/10.1007/978-3-319-76992-9_12`.

Oliva, L. S., A. Mura, A. Betella, D. Pacheco, E. Martinez and P. Verschure (2015). “Recovering the history of Bergen Belsen using an interactive 3D reconstruction in a mixed reality space the role of pre-knowledge on memory recollection”. In: *2015 Digital Heritage*. IEEE, pp. 163–165. DOI: `10.1109/digitalheritage.2015.7413860`.

Paudel, D. P., A. Habed, C. Demonceaux and P. Vasseur (2015). “Robust and Optimal Sum-of-Squares-Based Point-to-Plane Registration of Image Sets and Structured Scenes”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 2048–2056. DOI: `10.1109/iccv.2015.237`.

Ross, S. and M. Hedstrom (2005). “Preservation research and sustainable digital libraries”. In: *International Journal on Digital Libraries* 5, 4, pp. 317–324. DOI: `10.1007/s00799-004-0099-3`.

Schindler, G. and F. Dellaert (2012). “4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections”. In: *Journal of Multimedia* 7, 2, pp. 124–131. ISSN: 1796-2048. DOI: `https://doi.org/10.4304/jmm.7.2.124-131`.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra (2017). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 618–626. DOI: `10.1109/iccv.2017.74`.

Simonyan, K. and A. Zisserman (2014). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv preprint arXiv:1409.1556*, pp. 1–14.

Slater, M. and M. V. Sanchez-Vives (2016). “Enhancing our lives with immersive virtual reality”. In: *Frontiers in Robotics and AI* 3, p. 74. ISSN: 2296-9144. DOI: `https://doi.org/10.3389/frobt.2016.00074`.

Styliani, S., L. Fotis, K. Kostas and P. Petros (2009). “Virtual museums, a survey and some issues for consideration”. In: *Journal of cultural Heritage* 10, 4, pp. 520–528. ISSN: 1296-2074. DOI: `https://doi.org/10.1016/j.culher.2009.03.003`.

Tschirschwitz, F., G. Büyüksalih, T. Kersten, T. Kan, G. Enc and P. Baskaraca (2019). “Virtualising an Ottoman Fortress - Laser Scanning and 3D Modelling for the Development of an Interactive, Immersive Virtual Reality Application”. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, 2/W9, pp. 723–729. DOI: `https://doi.org/10.5194/isprs-archives-XLII-2-W9-723-2019`.

Wu, C. (2013). “Towards linear-time incremental structure from motion”. In: *3D Vision-3DV 2013, 2013 International conference on.* IEEE, pp. 127–134. ISBN: 0769550673. DOI: `https://doi.org/10.1109/3DV.2013.25`.

Wu, Y., W. Ma, M. Gong, L. Su and L. Jiao (2015). “A Novel Point-Matching Algorithm Based on Fast Sample Consensus for Image Registration”. In: *IEEE Geosci. Remote Sensing Lett.* 12, 1, pp. 43–47. DOI: `https://doi.org/10.1109/LGRS.2014.2325970`.

Yoon, J. and E. Chung (2011). “Understanding image needs in daily life by analyzing questions in a social Q&A site”. In: *Journal of the American Society for Information Science and Technology* 62, 11, pp. 2201–2213. ISSN: 1532-2882. DOI: `https://doi.org/10.1002/asi.21637`.

4 An adapted Structure-from-Motion Workflow for the orientation of historical images

Chapter 4 published under CC-BY-NC-ND license in "The Photogrammetric Record" (eISSN: 1477-9730) as:

An Automatic Workflow For Orientation Of Historical Images With Large Radiometric And Geometric Differences

Authors: Ferdinand Maiwald[1], Hans-Gerd Maas[1]

[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany

Publication history: submitted January 2021, accepted March 2021, published June 2021

Full reference: Maiwald, F. and Maas, H.-G. (2021), An automatic workflow for orientation of historical images with large radiometric and geometric differences. Photogram Rec, 36: 77-103. `https://doi.org/10.1111/phor.12363`.

Abstract: This contribution proposes a workflow for a completely automatic orientation of historical terrestrial urban images. Automatic Structure-from-Motion (SfM) software packages often fail when applied to historical image pairs due to large radiometric and geometric differences causing challenges with feature extraction and reliable matching. As an innovative initialising step, the proposed method uses the neural network D2-Net for feature extraction and Lowe's mutual nearest neighbour matcher. The principal distance for every camera is estimated using vanishing point detection. The results were compared to three state-of-the-art SfM workflows (Agisoft Metashape, Meshroom and COLMAP) with the proposed workflow outperforming the other SfM tools. The resulting camera orientation data are planned to be imported into a web and virtual/augmented reality (VR/AR) application for the purpose of knowledge transfer in cultural heritage.

Keywords: feature matching, historical images, image orientation, neural networks, structure from motion

4.1 Introduction

Spanning a bridge between traditional photogrammetry and state-of-the-art machine learning technologies, this paper presents an adapted automatic image-orientation workflow for historical terrestrial images using feature points derived by a neural network. The value of historical images lies in the fact that they are often the only remaining visual source of destroyed buildings and are thus essential for cultural heritage research. With an increasing amount of digitisation in libraries and repositories, growing quantities of image data are published and a demand for localisation and orientation parameters of historical images arises. While a single image only depicts a piece of two-dimensional information, an oriented and geo-registered image collection, combined with three-dimensional (3D) building models, can provide an increased level of scene understanding. Further, the image data can be used for improving (accuracy, texturing) and even generating 3D models as well as obtaining better knowledge on the position and motifs of the photographer (Fig. 4.1). Possible fields of application are Web3D environments or virtual/augmented reality (VR/AR).

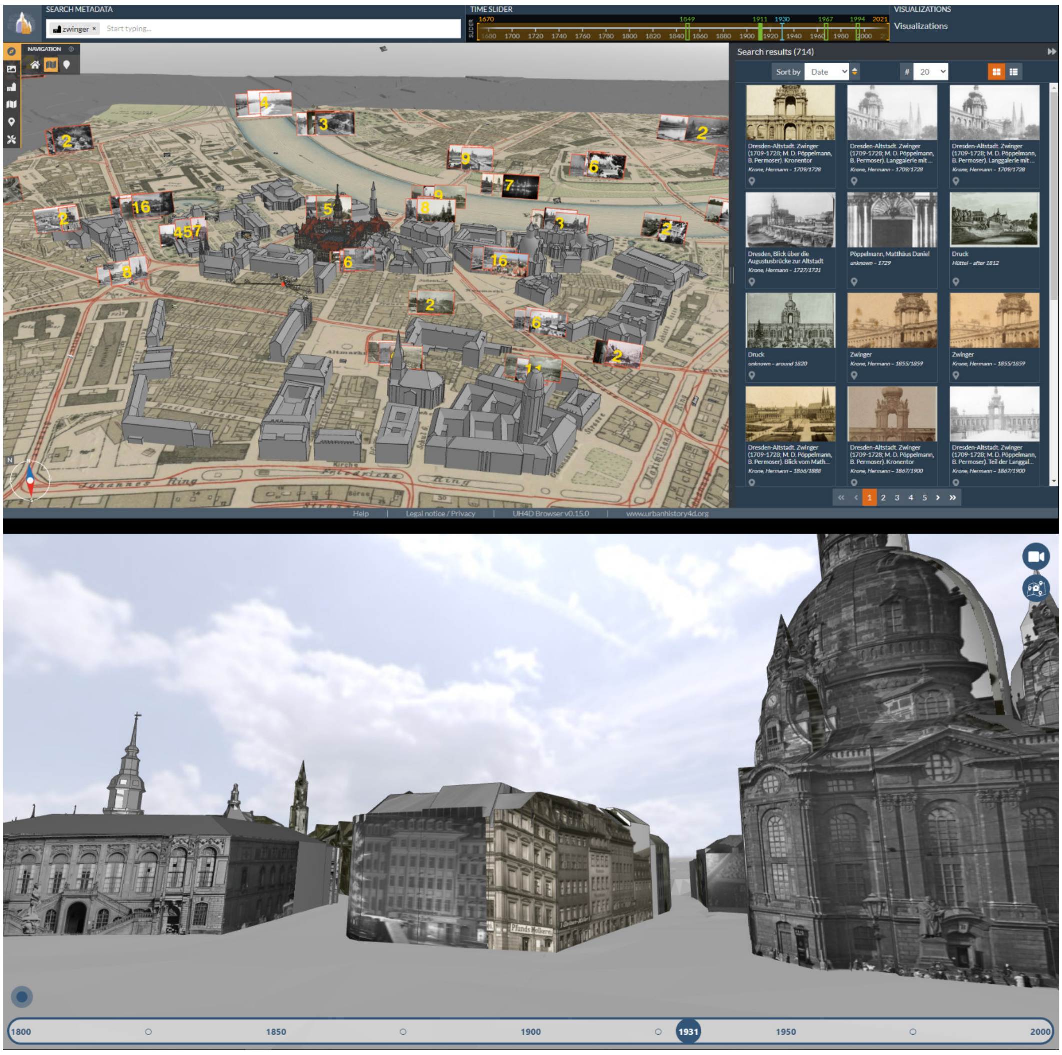


Figure (4.1): Two application scenarios for correctly oriented images. Top: 4D browser where the user can take up the position of the photographer (http://4dbrowser.urbanhistory4d.org). Bottom: Work in progress of a mobile 4D browser application where oriented images are projected on 3D models.

Photogrammetric methods are not easily applied to historical images due to the latter's characteristics. Most of the time, camera parameters are not available or are incorrect and thus have to be estimated. Especially when using digitised images, any information about the principal point's location and the principal distance are often lost. Additionally, the different images vary strongly in illumination and viewpoint of the depicted object, which makes a pairwise registration using feature matching methods more difficult. Digitisation artefacts caused by photographic film damage like mould or oxidation generate wrong or inaccurate feature matches. Most images are black and white, bleached out or sepia coloured (Fig. 4.2).

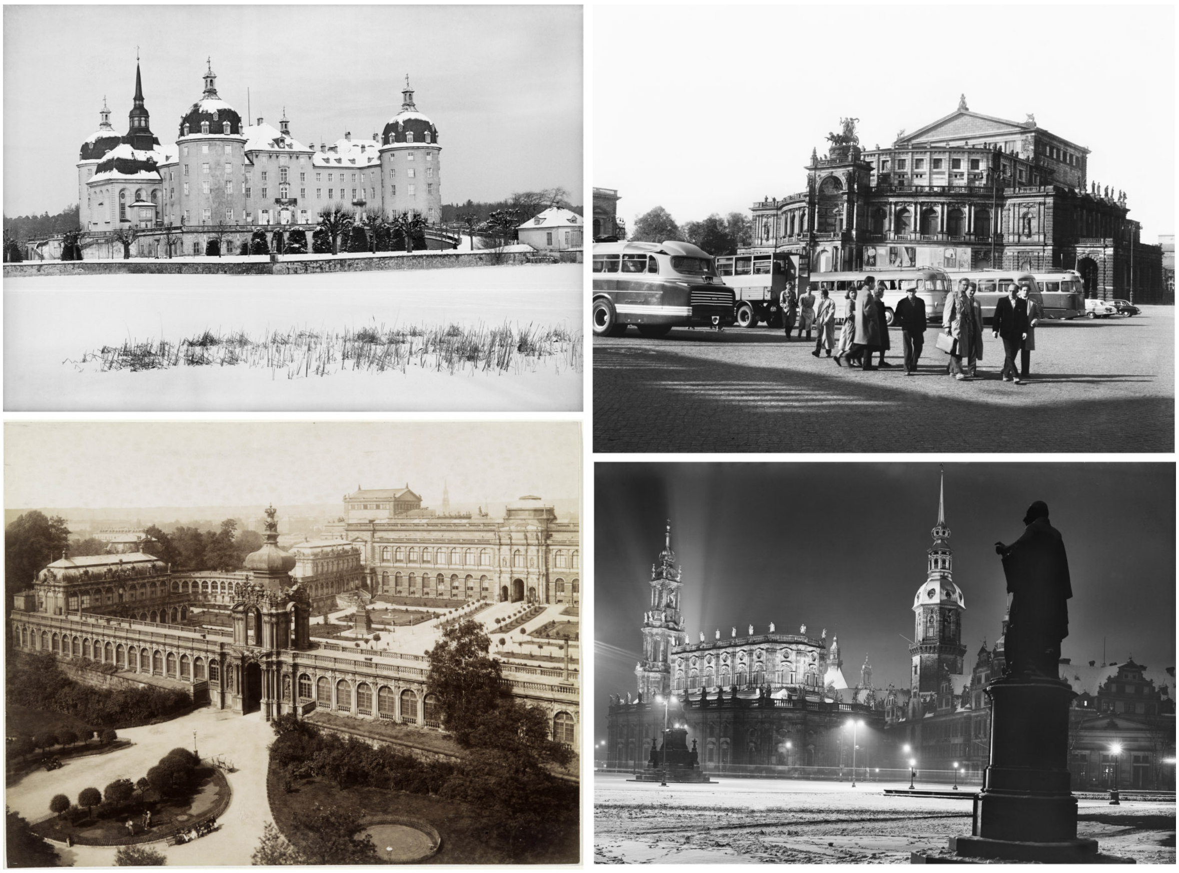


Figure (4.2): Four photographs showing the variety of historical images.

Consequently, most works in the field of historical image registration and 3D modelling are still done manually or semi-automatically: this contribution is intended to overcome this in providing a fully automatic workflow.

Conventional Structure-from-Motion (SfM) pipelines offer a workflow which consists of image selection, image orientation using feature extraction, description, pairwise matching and generation of sparse and/or dense 3D models. While these steps are optimised for recent images, the image orientation often fails for historical photographs. The critical step is the extraction of a large number of distinctive features and using an appropriate feature matching method including outlier removal. Considering future applications, this research mainly focuses on the correct image orientation, while the final step of generating a dense point cloud is left out. It has been shown that, due to changing or missing building elements and strongly varying point colours, the final dense 3D model is often not reasonable (Maiwald et al., 2017).

While it is difficult to integrate a researcher's own methods in proprietary software, this paper proposes the integration of D2-Net features (Dusmanu et al., 2019) with predefined parameters including the pre-calibration of the principal distance (camera constant) into the open-source SfM-routine COLMAP (Schönberger and Frahm, 2016). The features are matched between image pairs and additional images are added incrementally to the reconstructed scene. A sparse point cloud is created using bundle adjustment. The results on several historical image datasets are compared with state-of-the-art SfM pipelines of Agisoft Metashape, Meshroom (AliceVision, 2018) and the default version of COLMAP.

The paper is structured into four parts. The next section indicates related research in the field of cultural heritage as well as in computer vision, focusing on feature detection and matching, especially in architectural scenes. The following justifies the selection of D2-Net features for historical images and explains the newly introduced method in detail. Additionally, the workflows of three end-to-end software solutions are compared. This is followed by a section that presents the six different datasets used for the evaluation of the nine SfM pipelines. The results of the different methods are presented in the penultimate section and are compared in predefined metrics such as the number of oriented images and reprojection error. Finally, conclusions are drawn and an outlook on future work is given. The Appendix provides images showing examples of correctly and incorrectly matched photographs (Figs. 4.A1 and 4.A2).

4.2 Related Research

This section is structured into two parts. The first relates research in the field of cultural heritage and architecture, where historical images are used for reconstruction purposes. This part will be divided into interactive, semi-automatic workflows and fully automatic approaches. In particular, the main issues of automatic camera calibration, together with feature detection and matching in historical images, are considered. As a side note, research on historical aerial images and archival video material is presented.

The second part focuses on computer vision aspects, where different algorithmic (also called handcrafted) and deep-learning feature matching methods are compared. Particular research interest lies in methods appropriate for large illumination and viewpoint differences between image pairs. In cultural heritage, many different research investigations rely on the information that historical images are able to provide. Still, a lot of work in this field is undertaken manually.

4.2.1 Historical images for 3D reconstruction

As a pioneer of architectural photogrammetry, Albrecht Meydenbauer (1834–1921) was one of the first persons to meticulously document cultural heritage using metric cameras (Albertz, 2001). In the early 2000s, an increasing interest in research on using these historical images for 3D reconstruction purposes emerged, while dealing with similar problems (see the Introduction section) as today (Hemmleb, 1999; Wiedemann et al., 2000; Heuvel, 2002). Especially, the determination of camera parameters for single uncalibrated images was the interest in various work (Caprile and Torre, 1990; Cipolla et al., 1999; Wilczkowiak et al., 2001) and the pre-calibration of the principal distance using Vanishing Point Detection (VPD) is still used in different fields (Li et al., 2010; Tang et al., 2016; Zhou et al., 2017). With rising digitisation of historical images, the applications and automation potential have increased.

In interactive and semi-automatic workflows, Grün et al. (2004) were able to reconstruct the Buddha of Bamiyan in Afghanistan using least-squares matching, while Bitelli et al. (2007) calibrated three historical images using additional control points and geometric constraints in PhotoModeler. Recent examples of semi-automatic calibrations are the reconstruction of the civic tower of Sant' Alberto in Ravenna, Italy (Bitelli et al., 2017), the Fortezza Vecchia in Livorno, Italy (Bevilacqua et al., 2019) and the Borough at Biskupin in Poland (Zawieska and Markiewicz, 2016) using manually determined tie points (also referred to as feature points) due to image quality. Schindler and Dellaert (2012) compared a manual selection of tie points with an automated use of Scale Invariant Feature Transform (SIFT) features (Lowe, 2004) but still received better results for their four-dimensional (4D) city model of Atlanta using the interactive approach.

With an increasing number of SfM software tools – especially the open-source software VisualSFM (Wu, 2013) and the proprietary tool Agisoft PhotoScan (now Metashape) – the processing of historical image material became a research interest in different fields. Falkingham et al. (2014) reconstructed dinosaur tracks using historical images, and Mölg and Bolch (2017) utilised historical aerial images for analysing changes in glacier surface elevation in the western Swiss Alps. Redweik et al. (2016) recovered aerial photographs from 1934 and 1938 which could contribute to the evolution study of cliffs in Lisbon, Portugal as well as the 3D modelling of the port in Lisbon in 1938, while Pérez et al. (2014) were able to analyse territorial changes in Spain using historical aerial images from 1950 to 2014.

In architectural photogrammetry, recent research shows that Agisoft PhotoScan/Metashape is often capable of reconstructing a building, or part of a building, using historical images. Limitations to be considered are the automatic selection of appropriate images, retrieving the inner orientation of the camera (Grussenmeyer and Al Khalil, 2017) and the final 3D model quality (Maiwald et al., 2017; Rodríguez Miranda and Valle Melón, 2017; Beltrami et al., 2019).

As an additional perspective, there is also ongoing research regarding historical aerial images (Verhoeven and Vermeulen, 2016; Feurer and Vinatier, 2018; Giordano et al., 2018) as well as historical video material (Condorelli et al., 2020). However, the algorithms used in these works are not always transferable to terrestrial urban images.

4.2.2 Algorithmic Feature Detection and Matching

It has been identified that, in addition to camera calibration, feature extraction and description is considered to be the most difficult step in an SfM workflow when trying to retrieve the orientation of multiple historical images (Ali and Whitehead, 2014; Maiwald, 2019). While there exist many comprehensive surveys (Tuytelaars and Mikolajczyk, 2007; Fan et al., 2015; Csurka et al., 2018) and benchmarks (Mikolajczyk et al., 2005; Schönberger et al., 2017; Jin et al., 2020) of a large number of feature matching methods, there is very little research on evaluating feature matching on historical images (Maiwald, 2019). Consequently, the second part of this section on related research focuses on feature matching in image pairs with large illumination and/or viewpoint differences since these are quite comparable to pairs of historical photographs.

So-called handcrafted features (algorithmic methods) are manually described keypoints or regions around a keypoint through, for example, neighbourhoods or relations (Csurka et al., 2018). Its most common representative is SIFT, which finds local extrema in scale space and describes them by using gradient magnitude and orientation in an image region around these points (Lowe, 2004). In contrast, deep (learned) features are retrieved by deep-learning approaches (such as Convolutional Neural Networks (CNNs)) and learning to find keypoints and/or their descriptors by training on large image datasets. Both Csurka et al. (2018) as well as Jin et al. (2020) show that deep features are state of the art and are continuously improved. But both papers demonstrate that with proper settings, classical algorithmic methods may still be able to outperform the features derived by CNNs.

Nonetheless, for the presented datasets (see the Datasets section), it has been noticed that SIFT and its variants RootSIFT (Arandjelovic and Zisserman, 2012) and DSP-SIFT (Dong and Soatto, 2015) did not perform well and produced only few, or even no, matching image pairs. Further algorithmic methods that have been evaluated on historical images (Maiwald et al., 2019b; Maiwald, 2019) are Maximally Stable Extremal Regions (MSER) (Matas et al., 2004)), Radiation-Invariant Feature Transform (RIFT) (Li et al., 2020)), Matching On Demand with View Synthesis (MODS) (Mishkin et al., 2015a)) and MODS-WxBS (Mishkin et al., 2015b). Initially, the MSER method developed for wide baselines between image pairs produced decent results for historical image pairs, but was outperformed by a combination of RIFT and MODS/MODS-WxBS. While RIFT provides high robustness against illumination changes by using the phase information retrieved in the frequency domain of an image, MODS is particularly strong for wide baselines in image pairs. MODS starts with generating synthetic views between the image pairs and combines multiple feature detectors in its procedure. MODS-WxBS adds the possibility of using multiple descriptors for the different detectors and places them in separate multidimensional k-D trees, with the drawback that it is slower than the raw MODS variant.

4.2.3 Feature Detection and Matching using Convolutional Neural Networks

Avoiding an explicit description of keypoints, deep neural networks learn and evaluate features and/or descriptors using large image datasets. In 2008, one of the first successful CNNs for the creation of local image descriptors was created (Jahrer et al., 2008), and in the following years many different approaches were developed (Osendorfer et al., 2013; Simo-Serra et al., 2015; Yi et al., 2016). Since the presented work is dealing with historical images, this research focuses on networks that achieve good results on unordered image collections. Especially, the Oxford Buildings dataset (Philbin et al., 2007), the Photo Tourism dataset (Winder and Brown, 2007) and the Aachen Day–Night dataset (Sattler et al., 2012) are comparable to terrestrial urban historical images. In combination, these datasets show similar difficulties in finding appropriate features.

Multiple network architectures could achieve good results on these datasets (featuring a large number of registered images, a large number of features and correct feature matches and a small reprojection error) and the pretrained weights of these networks are often available as open-source code (Noh et al., 2017; DeTone et al., 2018; Ono et al., 2018; Dusmanu et al., 2019; Revaud et al., 2019; Sarlin et al., 2020). Aspects of computation time are mostly neglected as this research focuses on the quality aspects of the feature matching.

The current work uses D2-Net for the detection and matching of features because this network outperformed other algorithmic as well as deep learned approaches in the Image Matching Challenge of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (https://image-matching-workshop.github.io/challenge/). Additionally, it could be seen in the authors' tests that this approach performs well on real historical datasets (see the Feature

Matching section). Since every historical image is introduced as a new camera, the principal distance is pre-calibrated (if possible) using VPD and the complete scene reconstruction is refined using the integrated bundle adjustment of COLMAP.

4.3 Feature Matching

This section explains the authors' approach for testing feature matching methods on a historical dataset with predefined ground truth. Six different algorithms were compared to justify the selection of the D2-Net method used for the image-orientation workflow (see the Workflow section).

The unordered image collections such as the Oxford Buildings dataset (Philbin et al., 2007) or the Aachen Day–Night dataset (Sattler et al., 2012) are large ground-truth image datasets consisting of contemporary images. The authors experienced that historical image pairs are often much more difficult to match than recent images because of the properties already explained in the Introduction section. Since only sparse ground truth for historical terrestrial images is available, a small dataset consisting of 24 real-world example images (relative orientation of eight image triples) has been created in previous research (Maiwald, 2019). The ground-truth image orientation between image triples is available in the format of a Trifocal Tensor (`http://dx.doi.org/10.25532/OPARA-24`). The images are high-resolution digital copies of film negatives and show the variety of historical image material. Therefore, the photographs differ in viewpoint and illumination. Mainly, algorithmic methods like SIFT (Lowe, 2004) and MODS (Mishkin et al., 2015b) have already been tested and are now compared to the deep-learning approaches SuperPoint (DeTone et al., 2018) and D2-Net (Dusmanu et al., 2019). Most feature detection and matching methods focus on the computer vision problem of matching exactly two views of a scene and thus for a comparative evaluation the ground-truth Fundamental Matrix is extracted from the given Trifocal Tensor. For every possible image pair, feature points are detected, described and matched by the tested methods. In order to allow an equivalent comparison, all images are downsized to a maximum edge length of 1600 pixels using bilinear interpolation.

For SIFT and RIFT the already determined results (Maiwald et al., 2019b; Maiwald, 2019) are used for a comparison. For MODS both published variations are used (MODS and MODS-WxBS) to generate matches between the image pairs. The results of both processes are stored in a single file since the methods only produce a few matches due to strong filtering. Only one matching pair is kept if duplicate feature matches are found by MODS and also MODS-WxBS. For SuperPoint the pretrained weights of the Tensorflow implementation (https://github.com/rpautrat/SuperPoint) are used: the images are not further resized and the maximum number of feature points is set to 2048. The derived feature points are matched using Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981)) for outlier removal.

For D2-Net, both the single-scale and the multiscale approach are used and compared. That means for the multiscale approach multiple sets of feature maps are calculated for an image pyramid (half resolution, input resolution, double resolution) of the given image. The different feature maps are resized using bilinear interpolation to enable the determination of relevant features across all resolutions. The single-scale approach calculates the set of feature maps only for the input image. The underlying algorithm is explained in more detail in the following section (Workflow). The default approach uses the pretrained weights from the creator's implementation (https://github.com/mihaidusmanu/d2-net). As proposed by Dusmanu et al. (2019), Lowe's ratio test (Lowe, 2004) is used for outlier removal, only with mutual nearest neighbours with the recommended threshold of 0.95.

Feature matches are accepted as inliers if they fulfil the epipolar constraint using the provided and already optimised ground-truth fundamental matrix F:

$$x'^T F x = 0 \tag{4.1}$$

Therefore, the feature points of every method are converted into homogenous coordinates allowing multiplication with the 3×3 fundamental matrix. Since for real feature matches the epipolar constraint rarely results in exactly 0 (zero), a distance threshold has to be set to determine the inliers. This threshold (or algebraic distance) is proportional to the distance of a point from the epipolar line that is defined by the respective corresponding point (Fathy et al., 2011). When visualising the respective feature matches of the different methods, it has been experienced that an algebraic distance of 0.01 has been reasonable to determine the inliers:

$$x'^T F x <= 0.01 \tag{4.2}$$

All tested methods are compared considering the absolute number of inliers and the ratio of inliers to the absolute number of extracted matches by the respective method (matching score). Results are shown for every dataset in Table 4.1. The depicted data represent the results for the whole benchmark dataset which is additionally shown in more detail in the Appendix (Tab. 4.A1 and 4.A2). It has to be mentioned that the matching scores for the complete dataset are slightly lower because of difficult image pairs which could not be reliably matched by any of the methods.

Table (4.1): Evaluation of six different feature matching methods on a historical benchmark dataset consisting of 24 image pairs. Results are shown for eight different image pairs. The methods are evaluated by calculating the correct matches divided by the total number of found matches to produce the score. Additionally, the mean value of the correct number of matches is depicted.

Serial Number	SIFT+ RANSAC		RIFT		MODS MODS-WxBS		SuperPoint		D2-Net (single-scale)		D2-Net (multiscale)	
	score	inliers/ all_m	score	inliers/ all_m	score	inliers/ all_m	score	inliers/ all_m	score	inliers/ all_m	score	inliers/ all_m
1	0.26	(24/92)	0.67	(4/6)	1.00	(86/86)	0.20	(2/10)	0.96	(102/106)	0.98	(64/65)
4	0.24	(18/75)	0.00	(0/0)	1.00	(148/148)	0.33	(2/6)	0.91	(20/22)	1.00	(85/85)
7	0.14	(11/78)	0.78	(42/54)	0.00	(0/76)	1.00	(24/24)	0.87	(152/175)	0.88	(108/123)
10	0.52	(39/75)	0.00	(0/0)	1.00	(82/82)	0.14	(1/7)	0.97	(30/31)	0.94	(45/48)
13	1.00	(866/869)	0.43	(17/40)	0.97	(359/369)	0.88	(220/249)	0.86	(1465/1698)	0.75	(993/1317)
16	0.67	(78/117)	0.00	(0/0)	0.92	(24/26)	1.00	(5/5)	0.47	(8/17)	0.96	(27/28)
19	1.00	(707/707)	0.48	(10/21)	0.98	(290/295)	0.72	(63/88)	0.89	(681/763)	0.90	(841/931)
22	0.50	(42/84)	0.15	(4/26)	1.00	(30/30)	1.00	(13/13)	0.92	(178/194)	0.96	(149/155)
Mean	**0.54**	**223**	**0.31**	**10**	**0.86**	**127**	**0.66**	**41**	**0.86**	**329**	**0.92**	**289**

It can be seen that, for this particular dataset, D2-Net outperforms the other tested approaches. While SIFT and RIFT have the lowest average matching scores, D2-Net using the multiscale approaches generates an average of 92% correct matches. The combination of MODS and MODS-WxBS also reaches a value of 86%, similar to single-scale D2-Net, but shows overall a significantly lower number of correct matches. However, none of the tested methods achieves an overall acceptable matching score for every single image pair, which can also be seen in the Appendix (Tab. 4.A1). The most promising results are achieved by D2-Net and, consequently, this method is chosen for the following workflow.

4.4 Workflow

This section shows the main focus of the research and explains the workflow for successfully registering a large number of historical images automatically. While it is often not possible to include keypoints, feature matches and/or camera parameters into end-to-end software solutions, the proposed method integrates these steps into a single COLMAP SQLite database file. One database file is generated for each dataset (see the following Datasets section) and is accessed using the command-line interface. In the following paragraphs the presented approach is explained in three consecutive steps in more detail (Fig. 4.3).

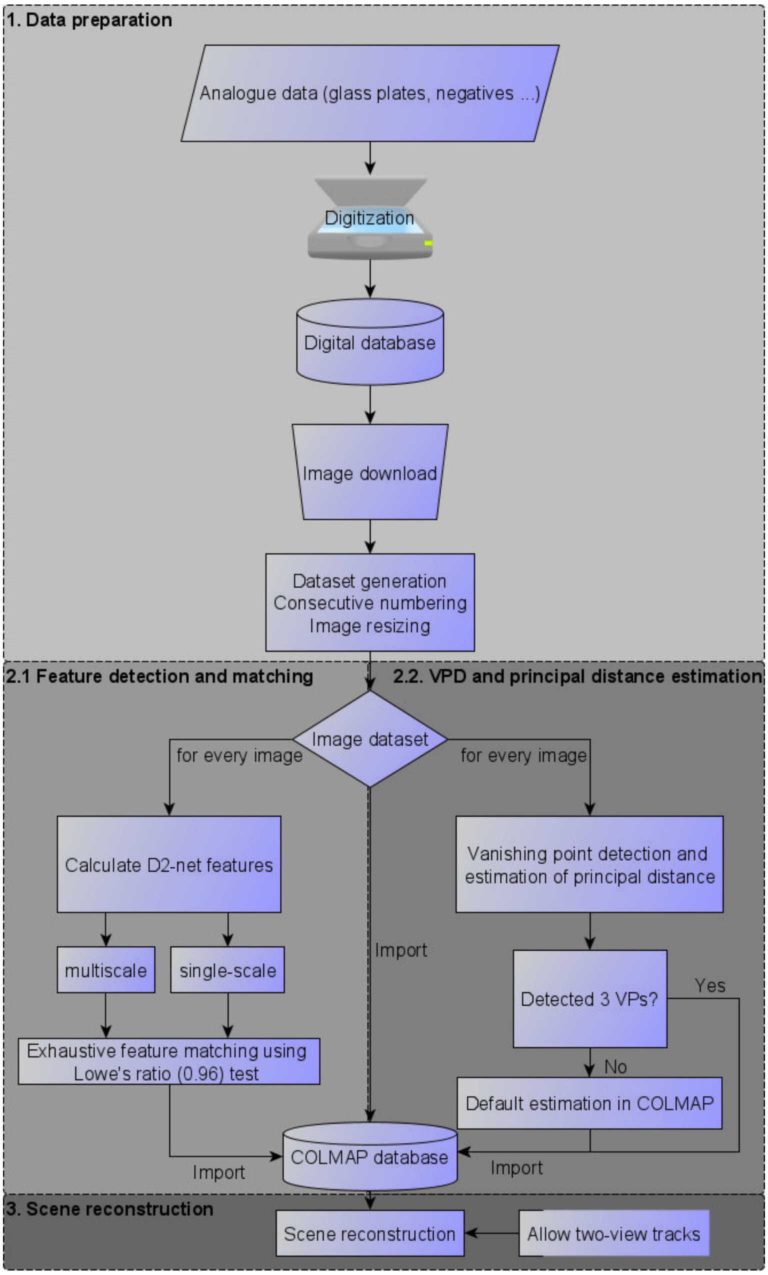


Figure (4.3): Workflow describing the three steps from retrieving historical datasets through scene reconstruction.

4.4.1 Step 1: Data preparation

Before working with historical images, the data have to be prepared (Fig. 4.3 – step 1). As a starting point, a digital copy has to be produced from the analogue original. Usually, a high-resolution scanner is used for this purpose and the result is uploaded into a digital database.

After that, to process the available datasets the images are downloaded in the exclusively provided JPEG file format and are numbered consecutively. For an equivalent comparison, all images are downsized to a maximum (long) edge length of 1600 pixels and a maximum sum of both image edges (long and short) of 2800 pixels using bilinear interpolation in OpenCV. If the images are already smaller than this they are not enlarged.

4.4.2 Step 2.1: Feature Detection and Matching

Importing precalculated features into COLMAP requires the preparation of a COLMAP database holding all images of one dataset (Fig. 4.3 – step 2.1). Then, for every single image, D2-Net features are calculated. D2-Net uses a “describe-and-detect” approach where a set of feature maps is computed for every image by a CNN. These feature maps are used for the detection of the strongest keypoints using a non-local maximum suppression followed by a non-maximum suppression. Simultaneously, the descriptors of these feature points are obtained by traversing through the set of feature maps. The presented approach uses the precomputed fine-tuned Caffe VGG16 (Simonyan and Zisserman, 2014) weights trained on the MegaDepth (Li and Snavely, 2018) dataset by Dusmanu et al. (2019). The features for the historical images are calculated using both the single-scale and multiscale approaches. That means that, for the multiscale approach, multiple sets of feature maps are calculated for an image pyramid (half resolution, input resolution, double resolution) of the given image. The different feature maps are resized using bilinear interpolation to enable the determination of relevant features across all resolutions. The single-scale approach calculates the set of feature maps only for the input image.

Next, the features of all images of a dataset are matched. Since it is expected that, for historical images, only a relatively small number of matches can be found between the image pairs, the proposed approach is to try feature matching between every possible combination of image pairs, a so-called exhaustive search. It has to be mentioned that this is not critical for smaller datasets but for the largest dataset 6 with $n = 467$ images (see the Datasets section) the exhaustive search calculates $(n^2 + n)/2 = 109278$ matching operations. On a i7-4790K quadcore CPU @ 4.00 GHz one matching operation takes around 5 seconds, leading to a total computation time of 151 hours. On a NVIDIA Quadro K2200 graphics processing unit (GPU) with 4 GB, the computation time for one operation drops below 1 second, still taking around 30 hours total time for the largest dataset. The feature matching pipeline has also been set up on a high-performance computing (HPC) system where the computation time is reduced to 5 hours working on one node and two Intel Haswell64 CPUs with a pre-allocated size of 5 GB memory per CPU. Possible optimisations are the parallelisation of the matching process or an automatic preselection of image pairs using a faster matching method in a first step. These solutions were not yet implemented for the tests shown in this paper.

As proposed by Dusmanu et al. (2019), Lowe’s ratio test (Lowe, 2004) is used for outlier removal only with mutual nearest neighbours; the resulting matches are imported into the COLMAP database. That means features are only kept as matches if the feature point P_m in image i maps only to the feature point P_n in image j and vice versa. Additionally, the distance to the second-closest neighbour has to be smaller than a certain ratio threshold. Dusmanu et al. (2019) recommend a threshold of 0.90 for the off-the-shelf descriptors and 0.95 for the fine-tuned ones. The presented work has shown that this ratio is a very critical point for the final result of the reconstruction. For a low threshold below 0.91 a reconstruction was not possible at all for any of the datasets, because almost all matches were rejected. In contrast to this, a threshold value above 0.97 often leads to incorrectly registered images and larger reprojection errors. This happens because almost none of the detected features are filtered out. As a measurement for the quality of the reconstruction a quotient considering the absolute number of registered images and the root mean square (RMS) reprojection error is used (eq. 4.3):

$$\text{quality quotient} = \frac{\text{RMS reprojection error}}{\text{number of oriented images}} \tag{4.3}$$

If this quality quotient is minimal, meaning a high number of registered images and a small RMS reprojection error, the most robust reconstruction for the dataset could be achieved. This quality quotient was calculated for all smaller datasets (see the Datasets section) in a range from 0.85 to 0.99 for both the single-scale and the multiscale approach (Fig. 4.4).

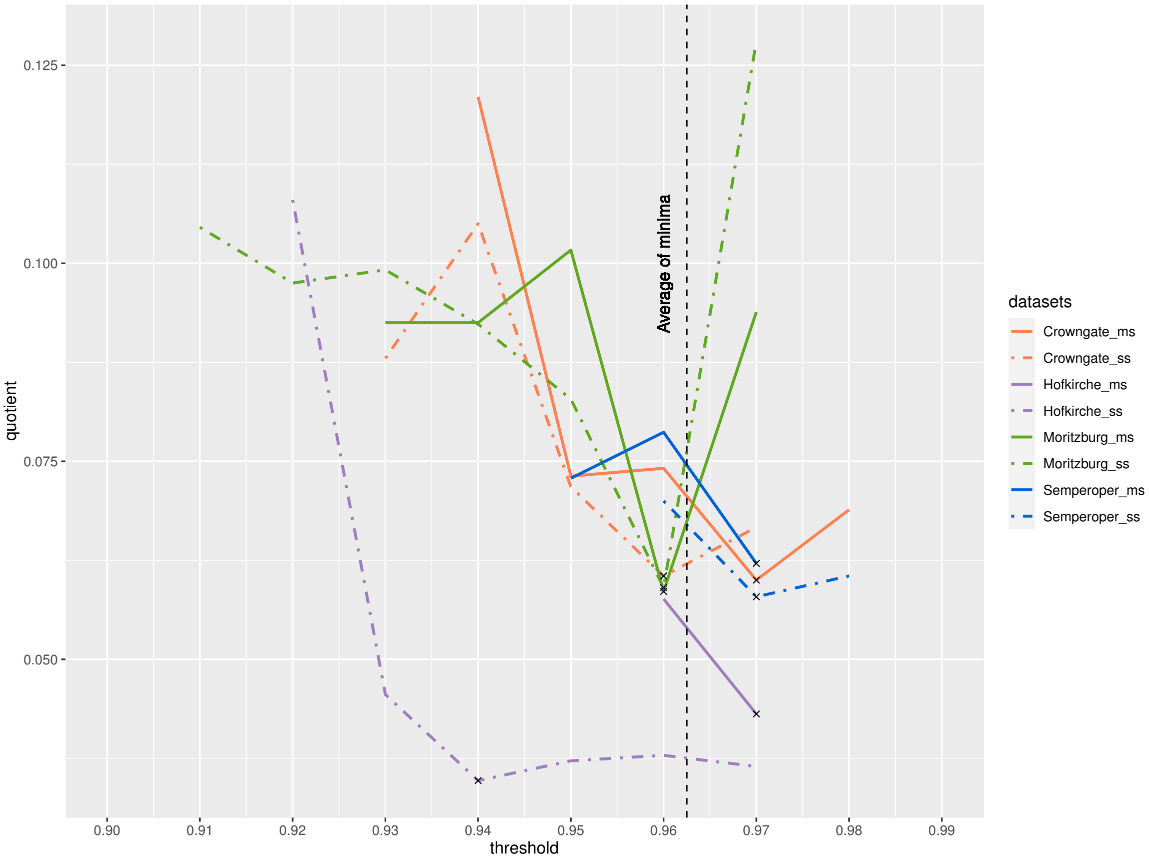


Figure (4.4): Determination of an appropriate ratio threshold for feature matching. A cross marks the minimum of every quality quotient of the eight datasets tested. The average of the minima is at 96.25.

It can be seen that for a threshold from 0.85 to 0.90 the relative orientations of the cameras could not be reconstructed because too few matches are accepted using the ratio test. A valid solution for the entirety of the datasets could only be calculated using the thresholds 0.96 and 0.97. Figure 4.4 shows that one minimum (the most robust reconstruction) lies at 0.94, three minima at 0.96 and four minima at 0.97. Because of the large deviations between 0.96 and 0.97, especially for the Moritzburg dataset in green, the authors recommend a fixed threshold of 0.96 using Lowe's ratio test for filtering historical image matches. It could also be considered using 0.97 for the multiscale approach of D2-Net since for three of the four datasets the quotient is lower than for 0.96. The resulting feature matches are imported into the COLMAP database. Correctly and also incorrectly matched features are shown in the examples in the Appendix (Figs. 4.A1 and 4.A2).

4.4.3 Step 2.2: Vanishing Point Detection and Principal Distance Estimation

Additional to the D2-Net features, an estimation of the principal distance is added to the database using VPD (Fig. 4.3 – step 2.2). Therefore, the proposed workflow uses the method of Li et al. (2010) since it performs more robustly and is faster than other approaches (Barnard, 1983; Košecká and Zhang, 2002). When detecting three vanishing points (VPs) in an image and with the assumption that the principal point lies in the image centre it is possible to estimate the principal distance (Barnard, 1983). The method of Li et al. (2010) detects lines in every image and calculates the intersections of every possible line pair. The intersection points are determined in a polar coordinate system where the origin is at the image centre. Accumulating the angle coordinates of that polar coordinate system in a discretised histogram with 600 bins mostly allows the distinctive determination of three VPs and consequently the principal distance.

However, the workflow of Li et al. (2010) always takes the three highest peaks in the accumulation diagram as three orthogonal VPs even if there are, for example, one, two or four clear peaks. Working with historical images, which sometimes do not have three orthogonal VPs, requires a filtering of the histogram. Thus, the proposed adaption searches for four local maxima in the accumulation histogram with the assumption that maxima are separated by at least 100 bins. A maximum is considered significant if its prominence is >50% of the maximum prominence of all peaks. This determines the absolute number of significant local maxima and, consequently, the question whether three VPs are visible. Only then is the principal distance calculated using the method of Li et al. (2010). Using this approach, if exactly three VPs could not be determined the default approximation of COLMAP ($f = 1.25 \cdot max(width_{px}, height_{px})$) is used and added to the existing database already holding the feature matches.

4.4.4 Step 3: Scene Reconstruction

The resulting database is imported into COLMAP, where the scene reconstruction is performed (Fig. 4.3 – step 3). A two-view reconstruction is used as a starting point, new images are registered incrementally including outlier removal and scene points are triangulated. After that, the reconstruction is refined using a bundle adjustment minimising the reprojection error. Further improvements could be achieved when allowing two-view feature tracks to be used for the reconstruction. That means that features are also used for the reconstruction if they only appear in two images. The default value is three but, considering the difficult matching between historical image pairs, the value has been lowered.

4.4.5 Comparison with Three Other State-of-the-Art SfM Workflows

The presented approach is mainly compared with three different SfM software tools. COLMAP 3.5 is used for the depicted workflow. The standard configuration using SiftGPU (Wu, 2013) and the default estimation of the focal length of the camera is compared with the proposed method using D2-Net features (single-scale and multiscale) and an integrated estimation of the principal distance using VPD. The final model size for the default configuration in COLMAP is set to a minimum of three images to allow a comparison to the other methods. For the proprietary software Agisoft Metashape 1.5.2, three different settings (low, high, very high) for automatic camera alignment and sparse point-cloud generation are tested. In AliceVision Meshroom 2019.2.0, a combination of different feature detectors and descriptor presets is tested: SIFT in combination with normal and ultra configurations, as well as SIFT and Accelerated KAZE (AKAZE) (Alcantarilla et al., 2013) with an ultra-descriptor preset. An overview of all the methods used is given in Table 4.2.

Table (4.2): Overview of the different workflow settings.

Software	Settings	
Agisoft Metashape		**Accuracy**
		Low
		High
		Very high
AliceVision Meshroom	**Descriptor**	**Descriptor preset**
	SIFT	Normal
	SIFT	Ultra
	AKAZE	Ultra
COLMAP	**Descriptor**	**Prior principal distance**
	SiftGPU	Default
	D2-Net (ss)	VPD/default
	D2-Net (ms)	VPD/default

4.5 Datasets

As already described, historical images show large variations in radiometric and geometric image quality. Therefore, this work tries to cover most of the issues using six different image datasets. Due to sparse image material, the fact that the images are from different moments in time is neglected, although this hampers the automatic reconstruction.

The first four smaller datasets consist of images which show famous landmarks near and in the city of Dresden in Germany and have already partly been described in Maiwald (2019). In particular, these are 35 images of the Hofkirche, 25 images of Moritzburg, 26 images of the Semperoper and 22 images of the Crowngate of the Dresden Zwinger. Since these are quite small datasets for an SfM workflow, two additional datasets were collected via a 4D browser application (Bruschke et al., 2018a). The images in the 4D browser were filtered using the metadata search for the term “Hofkirche”, resulting in 143 images, and for the term “Zwinger”, resulting in 467 images (Tab. 4.3). The datasets are consecutively numbered from 1 to 6. All images originate from the Saxon State and University Library Dresden (SLUB) and can be downloaded there or via `http://4dbrowser.urbanhistory4d.org` (Fig. 4.5).

Table (4.3): Overview of the six datasets used. The landmarks are all in the vicinity of Dresden, Germany and Fig. 4.5 provides an example image from each dataset.

Dataset	Size (no. of images)	Landmark	Time span
1	22	Crowngate	1880–1994
2	35	Hofkirche	1883–1992
3	25	Moritzburg	1884–1994
4	26	Semperoper	1869–1992
5	143	Hofkirche unordered	1883–2008
6	467	Zwinger unordered	1820–2009

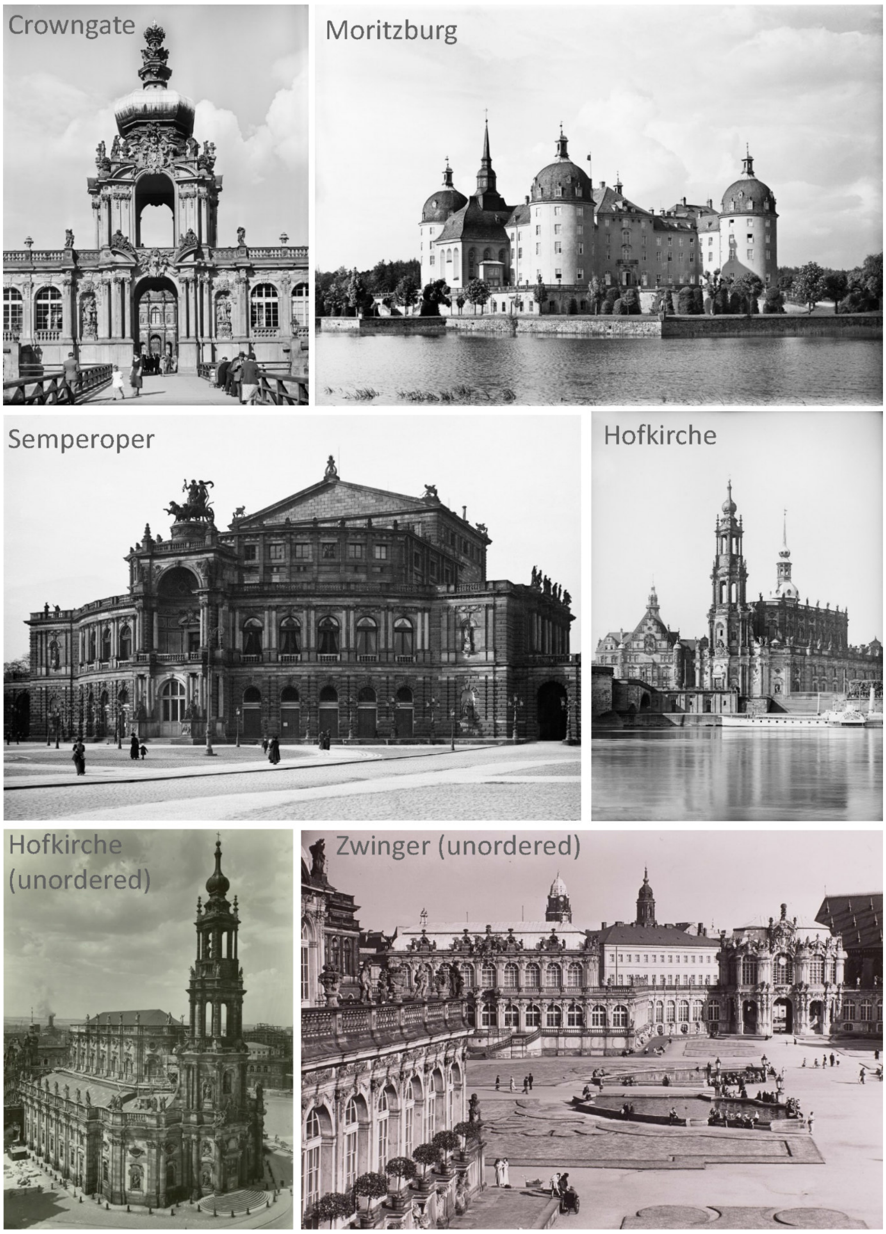


Figure (4.5): One example image for each of the six presented datasets showing the variety of historical photographs. All images originate from the Saxon State and University Library Dresden (SLUB).

Datasets 1 to 4 consist of preselected terrestrial images which, most of the time, show the respective landmark (or parts of it) in full and from different angles. However, the problem of, on the one hand, very small baselines and, on the other hand, extremely wide baselines exists. Additionally, some images are framed, vary in image colour (RGB, greyscale and sepia) and image artefacts occur. The images are from different instants in time as well as varying seasons: as such the depicted building changes in appearance. Sometimes objects (such as trams or cars) or people are in front of the landmark.

The more challenging unordered datasets (5 and 6) are directly exported from the 4D browser application (Bruschke et al., 2018a) and thus the images are not pre-filtered. In addition to the features of datasets 1 to 4 already outlined, these photographs have further properties hampering feature detection and matching. These images show not only the landmark in full but also building details over just part of the object. A small percentage (<5%) of these two datasets are oblique or nadir (vertical) aerial images. Furthermore, some photographs are taken at night-time, and a small number of these digital images are not photographs but plans or drawings. It has to be considered that the selection of unordered datasets may lead to wrong reconstruction results but it has been experienced that, in most approaches, a large number of images are already filtered out during feature matching. As an aside, the time span of dataset 6 starts in 1820, before the epoch of common photography, because there exist a few drawings of the building from this point in time. It is expected that a lower percentage of images of these datasets are able to be oriented by all given methods, since the image material is extremely diverse.

4.6 Results

Since there is no ground truth for the depicted datasets, the quality of the different approaches is measured using the number of oriented images, number of 3D points in the sparse cloud and the RMS of the reprojection error in pixels. A dense point cloud has not been created for any of the datasets since the results are often erroneous (see the Introduction section) and it is not the main focus of the presented research. Nevertheless, it has been observed that the sparse point cloud and the orientation of the cameras is more reasonable in this initial step. Calculating the dense cloud often leads to artefacts, inhomogeneous point colours or misaligned planes, especially for images of planar façades. Additionally, a visual control of the oriented images has been done to detect outliers and incorrectly aligned images. This occurs especially for very symmetrical and/or parallel parts of a building, or when setting Lowe's threshold to a value greater than 0.98 (see the Workflow section). All the results are shown in a single table comparing the nine different workflows, respectively, for the six datasets (Tab. 4.4). The best values are emphasised in bold letters. Special interest lies in results showing a large number of oriented images with an overall small reprojection error, since these results can be processed in further applications.

Table 4.4 shows that none of the methods is able to register all historical images of one dataset. Considering the pre-sorted datasets 1 to 4 (Tab. 4.4, rows 1 to 12), the results indicate that Meshroom in combination with the AKAZE descriptor and an ultra-descriptor preset produces a higher number of oriented images than the other default methods of COLMAP, Metashape and Meshroom.

Table (4.4): Results of nine different SfM workflows tested on six datasets. The table shows the number of oriented images, the RMS of the reprojection error and the number of 3D points of the sparse cloud. NV stands for "no valid reconstruction". The best results are highlighted in bold.

Dataset	SfM Workflow	Meshroom SIFT normal	Meshroom SIFT ultra	Meshroom AKAZE ultra	Metashape low	Metashape high	Metashape highest	COLMAP default	COLMAP D2 ss VPD (ours)	COLMAP D2 ss VPD (ours)
Crowngate (22 images)	Oriented images	5	5	**18**	4	3	4	5	**19**	**20**
	Reproj. error (px)	0.64	0.61	0.85	3.84	0.72	0.40	0.75	1.17	1.20
	3D points (sparse)	869	2100	14105	63	149	367	2456	3894	4472
Hofkirche (35 images)	Oriented images	2	7	10	2	8	2	7	10	**17**
	Reproj. error (px)	0.23	0.42	0.42	2.52	0.97	0.35	0.34	1.08	1.27
	3D points (sparse)	981	3274	9842	30	469	1048	3547	535	5633
Moritzburg (25 images)	Oriented images	2	2	4	2	3	2	2	**13**	**12**
	Reproj. error (px)	0.59	0.83	0.79	0.97	1.15	0.92	0.35	1.20	1.22
	3D points (sparse)	106	205	257	68	65	98	173	1075	1150
Semperoper (26 images)	Oriented images	3	2	13	3	2	3	3	**19**	**19**
	Reproj. error (px)	0.72	0.33	0.89	0.95	0.46	0.48	0.20	1.10	1.18
	3D points (sparse)	3182	929	3894	110	85	320	3682	2079	1975
Hofkirche unordered (143 imgs)	Oriented images	2	NV	NV	2	8	3	NV	**24**	**113**
	Reproj. error (px)	0.38	NV	NV	1.05	1.01	0.32	NV	1.25	1.34
	3D points (sparse)	207	NV	NV	80	1158	2469	NV	9772	13823
Zwinger unordered (467 imgs)	Oriented images	4	5	NV	6	13	**35**	25	27	**53**
	Reproj. error. (px)	0.54	0.51	NV	1.02	0.43	0.76	0.60	1.43	1.29
	3D points (sparse)	784	1564	NV	209	3305	12654	3599	12012	14742

However, the approach presented here outperforms the other methods by orienting around seven additional images per dataset (20% to 32%) in comparison to Meshroom AKAZE ultra. Looking at the absolute number of images, the proposed method is able to register 50% to 90% of the complete dataset. However, due to the higher number of oriented images, the overall RMS of the reprojection error increases up to 1.3 pixels. The other approaches stay mostly below 1.0 pixel.

Additionally, it is noticeable that the results with Meshroom cannot always be reproduced (not apparent in Table 4.4). The number of oriented images shows extreme variation, even when using exactly the same data and steering parameter set. Sometimes even a result for a reconstruction is not calculated by the software. A possible explanation for this might be in the random selection of the initial image pair which is essential for a reasonable scene reconstruction (Schönberger and Frahm, 2016). With COLMAP and Metashape it is almost always possible to reproduce the results.

The unordered historical datasets 5 and 6 are much more difficult to process with the standard routines of the three software packages. While scene reconstruction is possible for all workflows, the final output was often not valid. This means that a sparse model and a higher number of oriented images are generated but the resulting scene consists, for example, of multiple models of the same object, incorrectly merged models or extremely distorted camera models, especially with Meshroom. Examples of incorrectly matched images are depicted in the Appendix (Fig. 4.A2). Metashape turned out to be more reliable, since in the "high" and "very high" configurations it never produced incorrect results and for dataset 6 it provided a robust reconstruction of 35 images. Nevertheless, the number of oriented images is only between 2% and 7.5% of the total number of images. COLMAP in default configuration produced, for dataset 5, only non-valid models, while for dataset 6 multiple scenes were reconstructed. Four of these five scenes were valid and consisted, in sum, of around 50 images. The largest model (25 images) is depicted in Table 4.4. When using historical images this is advantageous because the single reconstructed scenes could be manually merged together in a subsequent step. Nevertheless, the authors' developed method using the D2-Net features also produced, overall, a higher number of registered images for the unordered datasets.

The number of 3D points in the sparse clouds varies between the different methods. Meshroom often produces more points for the same number of oriented images in comparison to COLMAP, while Metashape generates a reasonable number of points for the unordered datasets. When comparing the single-scale and multiscale feature detection of D2-Net on the respective datasets, it can be seen that – as expected – the multiscale approach performs slightly better than the single-scale approach. However, it has to be considered that the multiscale feature detection takes about three times more computation time.

4.7 Conclusions and Future Work

This contribution shows an approach to scene reconstruction exclusively using datasets of historical images. Three different software packages have been tested using varying configurations and were compared with the presented method on six challenging datasets. The depicted workflow, calculating D2-Net features and a prior principal distance estimation using VPD, outperforms state-of-the-art SfM software tools and is able to register the largest number of historical images with an acceptable RMS reprojection error. For the matching of D2-Net features, the use of Lowe's ratio test with mutual nearest neighbours and a fixed threshold of 0.96 is proposed. Furthermore, better reconstruction results can be achieved by allowing two-view tracks and disabling parallel bundle adjustment in COLMAP.

Since not all historical images could be added to the reconstruction in a first step, it is planned to add the missing images incrementally to include them in the already reconstructed scene of the respective landmark. An approach would be the use of a 6-degrees of freedom (6-DoF) hierarchical localisation pipeline. In a first step, a global search finds candidate images of the existing dataset that are potentially assignable with the query image. After that, the 6-DoF pose is estimated using extremely distinctive features (Sarlin et al., 2019). Considering that the datasets currently only consist of historical photographs showing landmarks, it is conceivable that, in a next step, further images from the side buildings or side streets will be added to the reconstruction and will then refine the calculated model. Due to the fact that more and more feature detection and feature matching methods using neural networks are being developed, it is possible that other descriptors are able to outperform D2-Net for the depicted datasets in the near future.

Additionally, it could be an option to register the images that could not be oriented by the proposed method using further recent images of the object. This approach requires that the building is still present today. It is also planned to automate the selection of appropriate images from the large database using single query images of the specific landmarks. The query images are given as input for a neural network and should provide images with similar perspectives of the building. An analysis of the respective image quality (blurred contents, artefacts) could then be realised using computer vision tools. The registered images will be imported into a browser application and can be compared with manually oriented photographs. A use in augmented/virtual reality (AR/VR) applications is also planned. The long-term goal is the automatic orientation of all urban terrestrial historical images of multiple databases retrieved by keyword search.

4.8 Acknowledgements

The research upon which this paper is based has received funding from the German Federal Ministry of Education and Research under grant agreement No. 01UG1630. Open access funding enabled and organised by ProjektDEAL.

4.A Appendix

This Appendix provides more detailed data and examples to supplement that in the main text. Table 4.A1 provides a matching score and Table 4.A2 the number of inliers, for each of the six feature matching methods for the four smaller datasets using 24 image pairs. Figure 4.A1 and 4.A2 provide, respectively, examples of correctly and incorrectly matched image pairs for datasets 1 to 4.

Table (4.A1): Evaluation of six different feature matching methods on a historical benchmark dataset consisting of 24 image pairs. The methods are evaluated calculating the ratio of correct matches divided by the total number of found matches (matching score). D2-Net shows the highest average matching score with 0.72.

	Serial	SIFT+		MODS		D2-Net	D2-Net
Dataset	Number	RANSAC	RIFT	MODS-WxBS	SuperPoint	(single-scale)	(multiscale)
Hofkirche_1_1-2	1	0.26	0.67	1.00	0.20	0.96	0.98
Hofkirche_1_1-3	2	0.35	0.25	0.00	0.20	0.73	0.81
Hofkirche_1_2-3	3	0.38	0.25	0.40	0.50	0.88	0.00
Hofkirche_2_1-2	4	0.24	0.00	1.00	0.33	0.91	1.00
Hofkirche_2_1-3	5	0.60	0.69	0.75	0.38	0.59	0.55
Hofkirche_2_2-3	6	0.96	0.84	0.89	1.00	0.62	0.64
Moritzburg_1_1-2	7	0.14	0.78	0.00	1.00	0.87	0.88
Moritzburg_1_1-3	8	0.13	0.83	0.00	1.00	0.78	0.69
Moritzburg_1_2-3	9	0.24	0.80	1.00	0.00	0.21	0.26
Moritzburg_2_1-2	10	0.52	0.00	1.00	0.14	0.97	0.94
Moritzburg_2_1-3	11	0.46	0.00	1.00	0.25	0.99	0.99
Moritzburg_2_2-3	12	0.52	0.73	0.33	1.00	0.38	0.37
Semperoper_1_1-2	13	1.00	0.43	0.97	0.88	0.86	0.75
Semperoper_1_1-3	14	0.63	0.00	0.69	0.90	0.78	0.81
Semperoper_1_2-3	15	0.74	0.00	1.00	1.00	1.00	1.00
Semperoper_2_1-2	16	0.67	0.00	0.92	1.00	0.47	0.96
Semperoper_2_1-3	17	0.53	0.09	0.00	0.02	0.11	0.11
Semperoper_2_2-3	18	0.78	0.25	0.97	0.60	0.00	0.94
Zwinger_1_1-2	19	1.00	0.48	0.98	0.72	0.89	0.90
Zwinger_1_1-3	20	1.00	0.00	0.99	1.00	0.87	0.94
Zwinger_1_2-3	21	0.88	0.50	0.95	0.69	0.92	0.89
Zwinger_2_1-2	22	0.50	0.15	1.00	1.00	0.92	0.96
Zwinger_2_1-3	23	0.63	0.47	0.63	0.32	0.68	0.68
Zwinger_2_2-3	24	0.66	0.63	0.39	0.06	0.29	0.21
Mean value		**0.58**	**0.37**	**0.70**	**0.59**	**0.70**	**0.72**

Table (4.A2): Evaluation of six different feature matching methods on a historical benchmark dataset consisting of 24 image pairs. This table shows the number of inliers compared to the number of all matches found by the respective methods. D2-Net in single-scale and multiscale shows the highest average number of inliers.

	Serial	SIFT+		MODS		D2-Net	D2-Net
Dataset	Number	RANSAC	RIFT	MODS-WxBS	SuperPoint	(single-scale)	(multiscale)
Hofkirche_1_1-2	1	(24/92)	(4/6)	(86/86)	(2/10)	(102/106)	(64/65)
Hofkirche_1_1-3	2	(17/48)	(2/8)	(0/0)	(1/5)	(19/26)	(13/16)
Hofkirche_1_2-3	3	(33/86)	(2/8)	(4/10)	(3/6)	(14/16)	(0/0)
Hofkirche_2_1-2	4	(18/75)	(0/0)	(148/148)	(2/6)	(20/22)	(85/85)
Hofkirche_2_1-3	5	(55/92)	(9/13)	(123/165)	(3/8)	(54/91)	(41/75)
Hofkirche_2_2-3	6	(89/93)	(16/19)	(156/176)	(13/13)	(56/90)	(52/81)
Moritzburg_1_1-2	7	(11/78)	(42/54)	(0/76)	(24/24)	(152/175)	(108/123)
Moritzburg_1_1-3	8	(10/75)	(35/42)	(0/44)	(41/41)	(295/376)	(145/210)
Moritzburg_1_2-3	9	(12/51)	(37/46)	(35/35)	(0/16)	(33/159)	(26/101)
Moritzburg_2_1-2	10	(39/75)	(0/0)	(82/82)	(1/7)	(30/31)	(45/48)
Moritzburg_2_1-3	11	(43/94)	(0/0)	(30/30)	(2/8)	(85/86)	(76/77)
Moritzburg_2_2-3	12	(47/90)	(8/11)	(45/135)	(7/7)	(47/125)	(62/168)
Semperoper_1_1-2	13	(866/869)	(17/40)	(359/369)	(220/249)	(1465/1698)	(993/1317)
Semperoper_1_1-3	14	(3285/5185)	(0/0)	(331/481)	(644/715)	(3884/4972)	(4285/5301)
Semperoper_1_2-3	15	(745/1007)	(0/0)	(189/189)	(421/421)	(3510/3514)	(3376/3377)
Semperoper_2_1-2	16	(78/117)	(0/0)	(24/26)	(5/5)	(8/17)	(27/28)
Semperoper_2_1-3	17	(333/633)	(3/34)	(0/96)	(1/65)	(85/776)	(75/705)
Semperoper_2_2-3	18	(83/107)	(1/4)	(31/32)	(3/5)	(0/0)	(16/17)
Zwinger_1_1-2	19	(707/707)	(10/21)	(290/295)	(63/88)	(681/763)	(841/931)
Zwinger_1_1-3	20	(411/411)	(0/0)	(108/109)	(20/20)	(168/193)	(342/362)
Zwinger_1_2-3	21	(1357/1534)	(46/92)	(231/243)	(82/118)	(1734/1887)	(1938/2169)
Zwinger_2_1-2	22	(42/84)	(4/26)	(30/30)	(13/13)	(178/194)	(149/155)
Zwinger_2_1-3	23	(59/93)	(23/49)	(31/49)	(15/47)	(216/316)	(202/299)
Zwinger_2_2-3	24	(113/172)	(20/32)	(26/66)	(2/31)	(138/475)	(91/438)
Mean value		**353**	**12**	**98**	**66**	**541**	**544**

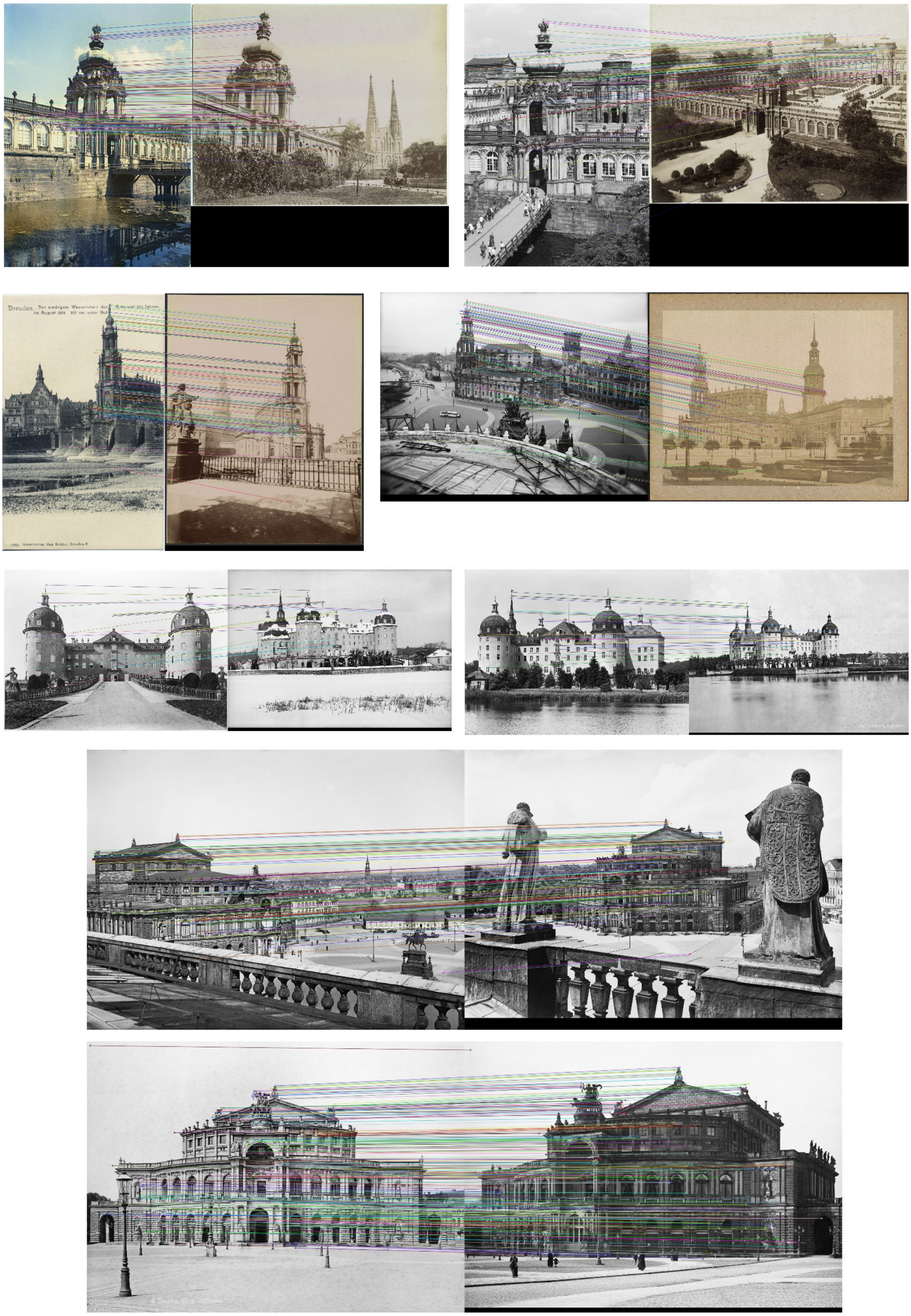

Dresden. Der niedrigste Wasserstand der Elbe seit 100 Jahren.
im August 1904. 227 cm unter Null

Figure (4.A1): Examples of difficult, but correctly matched, image pairs of datasets 1 to 4.

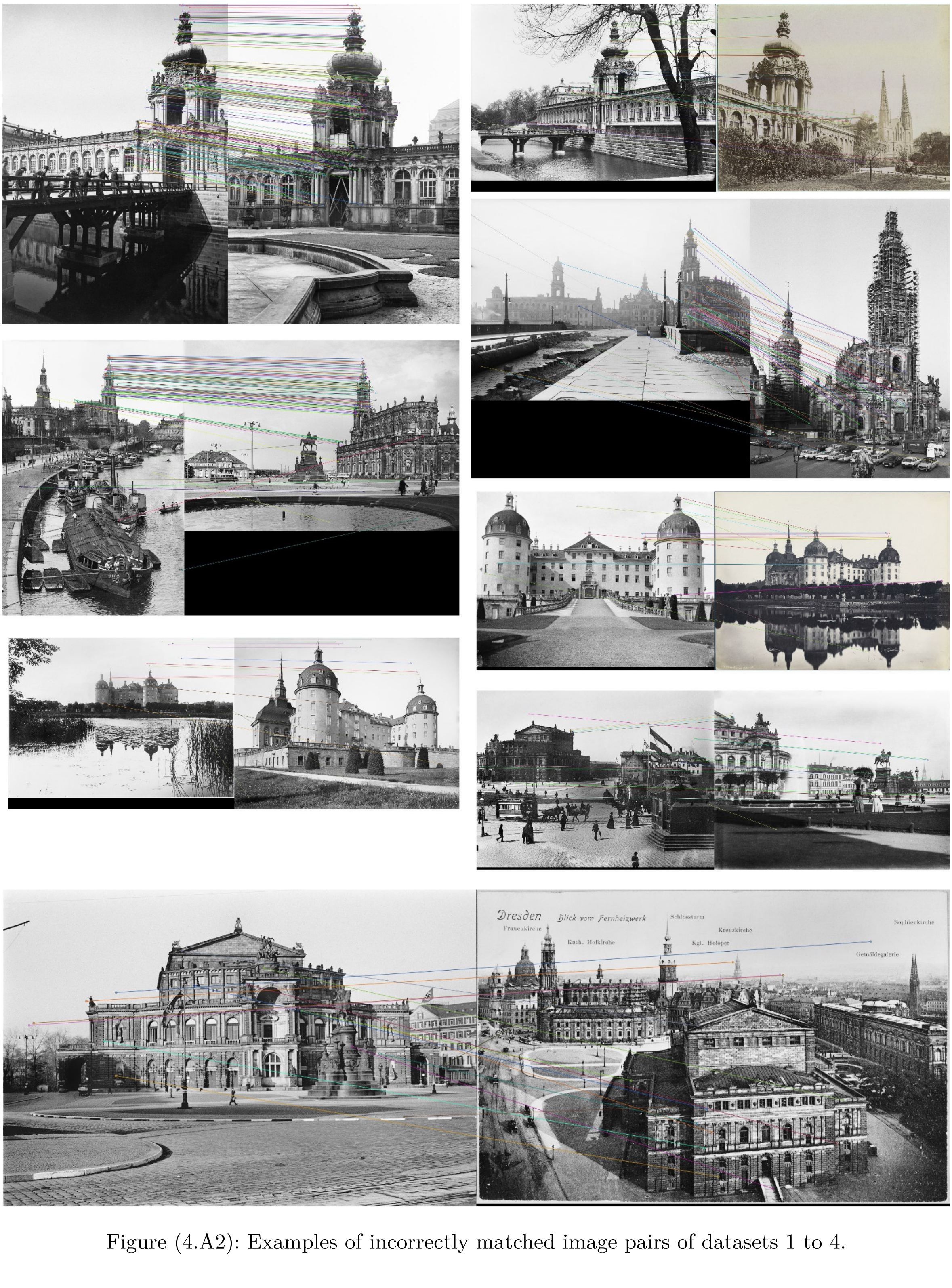


Figure (4.A2): Examples of incorrectly matched image pairs of datasets 1 to 4.

References

Albertz, J. (2001). "Albrecht Meydenbauer-Pioneer of photogrammetric documentation of the cultural heritage". In: *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34, 5/C7, pp. 19–25. ISSN: 1682-1750.

Alcantarilla, P., J. Nuevo and A. Bartoli (2013). "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces". In: *Procedings of the British Machine Vision Conference 2013*. British Machine Vision Association, pp. 1–11. DOI: 10.5244/c.27.13.

Ali, H. K. and A. Whitehead (2014). "Feature Matching for Aligning Historical and Modern Images". In: *Int. J. Comput. Appl.* 21, 3, pp. 188–201.

AliceVision (2018). *Meshroom: A 3D reconstruction software.* URL: https://github.com/alicevision/meshroom.

Arandjelovic, R. and A. Zisserman (2012). "Three things everyone should know to improve object retrieval". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2911–2918. DOI: 10.1109/cvpr.2012.6248018.

Barnard, S. T. (1983). "Interpreting perspective images". In: *Artificial Intelligence* 21, 4, pp. 435–462. DOI: 10.1016/s0004-3702(83)80021-6.

Beltrami, C., D. Cavezzali, F. Chiabrando, A. I. Idelson, G. Patrucco and F. Rinaudo (2019). "3D Digital and Physical Reconstruction of a Collapsed Dome using SFM Techniques from Historical Images". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W11, pp. 217–224. DOI: 10.5194/isprs-archives-xlii-2-w11-217-2019.

Bevilacqua, M. G., G. Caroti, A. Piemonte and D. Ulivieri (2019). "Reconstruction of lost Architectural Volumes by Integration of Photogrammetry from Archive Imagery with 3-D Models of the Status Quo". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 119–125. DOI: 10.5194/isprs-archives-xlii-2-w9-119-2019.

Bitelli, G., M. Dellapasqua, V. A. Girelli, S. Sbaraglia and M. A. Tinia (2017). "Historical Photogrammetry and Terrestrial Laser Scanning for the 3d Virtual Reconstruction of Destroyed Structures: A Case Study in Italy". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5/W1, pp. 113–119. DOI: 10.5194/isprs-archives-xlii-5-w1-113-2017.

Bitelli, G., V. A. Girelli, M. Marziali and A. Zanutta (2007). "Use of historical images for the documentation and the metrical study of cultural heritage by means of digital photogrammetric techniques". In: *AntiCIPAting the Future of the Cultural Past, Proceedings of the XXI International CIPA Symposium*, pp. 1–6.

Bruschke, J., F. Maiwald, S. Münster and F. Niebling (2018a). "Browsing and Experiencing Repositories of Spatially Oriented Historic Photographic Images". In: *Studies in Digital Heritage* 2, 2, pp. 138–149. DOI: 10.14434/sdh.v2i2.24460.

Caprile, B. and V. Torre (1990). "Using vanishing points for camera calibration". In: *International Journal of Computer Vision* 4, 2, pp. 127–139. DOI: 10.1007/bf00127813.

Cipolla, R., T. Drummond and D. Robertson (1999). “Camera Calibration from Vanishing Points in Image of Architectural Scenes”. In: *Procedings of the British Machine Vision Conference 1999*. British Machine Vision Association, pp. 382–391. DOI: `10.5244/c.13.38`.

Condorelli, F., F. Rinaudo, F. Salvadore and S. Tagliaventi (2020). “A Neural Networks Approach to Detecting Lost Heritage in Historical Video”. In: *ISPRS International Journal of Geo-Information* 9, 5, p. 297. DOI: `10.3390/ijgi9050297`.

Csurka, G., C. R. Dance and M. Humenberger (2018). “From handcrafted to deep local features”. In: *arXiv:1807.10254*, pp. 1–41.

DeTone, D., T. Malisiewicz and A. Rabinovich (2018). “SuperPoint: Self-Supervised Interest Point Detection and Description”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 337–33712. DOI: `10.1109/cvprw.2018.00060`.

Dong, J. and S. Soatto (2015). “Domain-size pooling in local descriptors: DSP-SIFT”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5097–5106. DOI: `10.1109/cvpr.2015.7299145`.

Dusmanu, M., I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii and T. Sattler (2019). “D2-Net: A Trainable CNN for Joint Description and Detection of Local Features”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8084–8093. DOI: `10.1109/cvpr.2019.00828`.

Falkingham, P. L., K. T. Bates and J. O. Farlow (2014). “Historical Photogrammetry: Bird’s Paluxy River Dinosaur Chase Sequence Digitally Reconstructed as It Was prior to Excavation 70 Years Ago”. In: *PLoS ONE* 9, 4, e93247. DOI: `10.1371/journal.pone.0093247`.

Fan, B., Z. Wang and F. Wu (2015). *Local Image Descriptor: Modern Approaches*. Springer Berlin Heidelberg. DOI: `10.1007/978-3-662-49173-7`.

Fathy, M. E., A. S. Hussein and M. F. Tolba (2011). “Fundamental matrix estimation: A study of error criteria”. In: *Pattern Recognition Letters* 32, 2, pp. 383–391. DOI: `10.1016/j.patrec.2010.09.019`.

Feurer, D. and F. Vinatier (2018). “Joining multi-epoch archival aerial images in a single SfM block allows 3-D change detection with almost exclusively image information”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 146, pp. 495–506. DOI: `10.1016/j.isprsjprs.2018.10.016`.

Fischler, M. A. and R. C. Bolles (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24, 6, pp. 381–395. DOI: `10.1145/358669.358692`.

Giordano, S., A. L. Bris and C. Mallet (2018). “Toward Automatic Georeferencing of Archival Aerial Photogrammetric Surveys”. In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2, pp. 105–112. DOI: `10.5194/isprs-annals-iv-2-105-2018`.

Grün, A., F. Remondino and L. Zhang (2004). “Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan”. In: *The Photogrammetric Record* 19, 107, pp. 177–199. DOI: `10.1111/j.0031-868x.2004.00278.x`.

Grussenmeyer, P. and O. Al Khalil (2017). “From Metric Image Archives to Point Cloud Reconstruction: Case Study of the Great Mosque of Aleppo in Syria”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W5, pp. 295–301. ISSN: 2194-9034. DOI: `10.5194/isprs-archives-XLII-2-W5-295-2017`.

Hemmleb, M. (1999). “Digital Rectification of historical images”. In: *CIPA International Symposium, Olinda, published on CD-ROM and in print*, pp. 1–6.

Henze, F., H. Lehmann and B. Bruschke (2009b). “Analysis of Historic Maps and Images for Research on Urban Development of Baalbek/Lebanon”. In: *Photogrammetrie - Fernerkundung - Geoinformation* 2009, 3, pp. 221–234. DOI: `10.1127/0935-1221/2009/0017`.

Heuvel, F. A. van den (2002). “Reconstruction from a single architectural image from the Meydenbauer archives”. In: *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34, 5/C7, pp. 699–706. ISSN: 1682-1750.

Jahrer, M., M. Grabner and H. Bischof (2008). “Learned local descriptors for recognition and matching”. In: *Computer Vision Winter Workshop*. Vol. 2, pp. 39–46.

Jin, Y., D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi and E. Trulls (2020). “Image Matching Across Wide Baselines: From Paper to Practice”. In: *International Journal of Computer Vision* 129, 2, pp. 517–547. DOI: `10.1007/s11263-020-01385-0`.

Košecká, J. and W. Zhang (2002). “Video Compass”. In: *Computer Vision — ECCV 2002*. Springer Berlin Heidelberg, pp. 476–490. DOI: `10.1007/3-540-47979-1_32`.

Li, B., K. Peng, X. Ying and H. Zha (2010). “Simultaneous Vanishing Point Detection and Camera Calibration from Single Images”. In: *Advances in Visual Computing*. Springer Berlin Heidelberg, pp. 151–160. DOI: `10.1007/978-3-642-17274-8_15`.

Li, J., Q. Hu and M. Ai (2020). “RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform”. In: *IEEE Transactions on Image Processing* 29, pp. 3296–3310. DOI: `10.1109/tip.2019.2959244`.

Li, Z. and N. Snavely (2018). “MegaDepth: Learning Single-View Depth Prediction from Internet Photos”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2041–2050. DOI: `10.1109/cvpr.2018.00218`.

Lowe, D. G. (2004). “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60, 2, pp. 91–110. ISSN: 0920-5691. DOI: `https://doi.org/10.1023/B:VISI.0000029664.99615.94`.

Maiwald, F. (2019). “Generation of a Benchmark Dataset Using Historical Photographs for an Automated Evaluation of Different Feature Matching Methods”. In: *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-2/W13, pp. 87–94. ISSN: 2194-9034. DOI: `10.5194/isprs-archives-XLII-2-W13-87-2019`.

Maiwald, F., J. Bruschke, C. Lehmann and F. Niebling (2019b). “A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR”. In: *Virtual Archaeology Review* 10, 21, pp. 1–13. DOI: `10.4995/var.2019.11867`.

Maiwald, F., T. Vietze, D. Schneider, F. Henze, S. Münster and F. Niebling (2017). “Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities". In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 447–452. ISSN: 1682-1750. DOI: `https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017`.

Matas, J., O. Chum, M. Urban and T. Pajdla (2004). "Robust wide-baseline stereo from maximally stable extremal regions". In: *Image and vision computing* 22, 10, pp. 761–767. ISSN: 0262-8856. DOI: `https://doi.org/10.1016/j.imavis.2004.02.006`.

Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. V. Gool (2005). "A Comparison of Affine Region Detectors". In: *International Journal of Computer Vision* 65, 1-2, pp. 43–72. DOI: `10.1007/s11263-005-3848-x`.

Mishkin, D., J. Matas and M. Perdoch (2015a). "MODS: Fast and robust method for two-view matching". In: *Computer Vision and Image Understanding* 141, pp. 81–93. ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2015.08.005`.

Mishkin, D., J. Matas, M. Perdoch and K. Lenc (2015b). "WxBS: Wide Baseline Stereo Generalizations". In: *Proceedings of the British Machine Vision Conference (BMVC).* BMVA Press, pp. 12.1–12.12. ISBN: 1-901725-53-7. DOI: `10.5244/C.29.12`.

Mölg, N. and T. Bolch (2017). "Structure-from-Motion Using Historical Aerial Images to Analyse Changes in Glacier Surface Elevation". In: *Remote Sensing* 9, 10, p. 1021. DOI: `10.3390/rs9101021`.

Noh, H., A. Araujo, J. Sim, T. Weyand and B. Han (2017). "Large-Scale Image Retrieval with Attentive Deep Local Features". In: *2017 IEEE International Conference on Computer Vision (ICCV).* IEEE, pp. 3476–3485. DOI: `10.1109/iccv.2017.374`.

Ono, Y., E. Trulls, P. Fua and K. M. Yi (2018). "LF-Net: Learning Local Features from Images". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 6237–6247. DOI: `https://arxiv.org/pdf/1805.09662.pdf`.

Osendorfer, C., J. Bayer, S. Urban and P. van der Smagt (2013). "Convolutional Neural Networks Learn Compact Local Image Descriptors". In: *Neural Information Processing.* Springer Berlin Heidelberg, pp. 624–630. DOI: `10.1007/978-3-642-42051-1_77`.

Pérez, J. A., F. M. Bascon and M. C. Charro (2014). "Photogrammetric Usage of 1956-57 Usaf Aerial Photography of Spain". In: *The Photogrammetric Record* 29, 145, pp. 108–124. DOI: `10.1111/phor.12048`.

Philbin, J., O. Chum, M. Isard, J. Sivic and A. Zisserman (2007). "Object retrieval with large vocabularies and fast spatial matching". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 1–8. DOI: `10.1109/cvpr.2007.383172`.

Redweik, P., V. Garzón and T. sá Pereira (2016). "Recovery of Stereo Aerial Coverage from 1934 and 1938 into the Digital Era". In: *The Photogrammetric Record* 31, 153, pp. 9–28. DOI: `10.1111/phor.12137`.

Revaud, J., P. Weinzaepfel, C. D. Souza, N. Pion, G. Csurka, Y. Cabon and M. Humenberger (2019). "R2D2: Repeatable and Reliable Detector and Descriptor". In: *arXiv:1906.06195*, pp. 1–11.

Rodríguez Miranda, A. and J. M. Valle Melón (2017). "Recovering Old Stereoscopic Negatives and Producing Digital 3d Models of Former Appearances of Historic Buildings". In: *ISPRS*

- *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W3, pp. 601–608. ISSN: 2194-9034. DOI: 10.5194/isprs-archives-XLII-2-W3-601-2017.

Sarlin, P.-E., C. Cadena, R. Siegwart and M. Dymczyk (2019). "From Coarse to Fine: Robust Hierarchical Localization at Large Scale". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 12708–12717. DOI: 10.1109/cvpr.2019.01300.

Sarlin, P.-E., D. DeTone, T. Malisiewicz and A. Rabinovich (2020). "SuperGlue: Learning Feature Matching with Graph Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947. DOI: arXiv:1911.11763[.

Sattler, T., T. Weyand, B. Leibe and L. Kobbelt (2012). "Image Retrieval for Image-Based Localization Revisited". In: *Procedings of the British Machine Vision Conference 2012*. Vol. 1. British Machine Vision Association, p. 4. DOI: 10.5244/c.26.76.

Schindler, G. and F. Dellaert (2012). "4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections". In: *Journal of Multimedia* 7, 2, pp. 124–131. ISSN: 1796-2048. DOI: https://doi.org/10.4304/jmm.7.2.124-131.

Schönberger, J. L. and J.-M. Frahm (2016). "Structure-from-Motion Revisited". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4104–4113. DOI: 10.1109/cvpr.2016.445.

Schönberger, J. L., H. Hardmeier, T. Sattler and M. Pollefeys (2017). "Comparative Evaluation of Hand-Crafted and Learned Local Features". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6959–6968. DOI: 10.1109/cvpr.2017.736.

Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer (2015). "Discriminative Learning of Deep Convolutional Feature Point Descriptors". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 118–126. DOI: 10.1109/iccv.2015.22.

Simonyan, K. and A. Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv preprint arXiv:1409.1556*, pp. 1–14.

Tang, Z., Y.-S. Lin, K.-H. Lee, J.-N. Hwang, J.-H. Chuang and Z. Fang (2016). "Camera self-calibration from tracking of moving persons". In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 265–270. DOI: 10.1109/icpr.2016.7899644.

Tuytelaars, T. and K. Mikolajczyk (2007). "Local Invariant Feature Detectors: A Survey". In: *Foundations and Trends® in Computer Graphics and Vision* 3, 3, pp. 177–280. DOI: 10.1561/0600000017.

Verhoeven, G. and F. Vermeulen (2016). "Engaging with the Canopy—Multi-Dimensional Vegetation Mark Visualisation Using Archived Aerial Images". In: *Remote Sensing* 8, 9, p. 752. DOI: 10.3390/rs8090752.

Wiedemann, A., M. Hemmleb and J. Albertz (2000). "Reconstruction of historical buildings based on images from the Meydenbauer archives". In: *International Archives of Photogrammetry and Remote Sensing* 33, B5/2; PART 5, pp. 887–893. ISSN: 0256-1840.

Wilczkowiak, M., E. Boyer and P. Sturm (2001). "Camera calibration and 3D reconstruction from single images using parallelepipeds". In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001.* IEEE Comput. Soc, pp. 142–148. DOI: `10.1109/iccv.2001.937510`.

Winder, S. A. J. and M. Brown (2007). "Learning Local Image Descriptors". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 1–8. DOI: `10.1109/cvpr.2007.382971`.

Wu, C. (2013). "Towards linear-time incremental structure from motion". In: *3D Vision-3DV 2013, 2013 International conference on.* IEEE, pp. 127–134. ISBN: 0769550673. DOI: `https://doi.org/10.1109/3DV.2013.25`.

Yi, K. M., E. Trulls, V. Lepetit and P. Fua (2016). "LIFT: Learned Invariant Feature Transform". In: *Computer Vision – ECCV 2016.* Springer International Publishing, pp. 467–483. DOI: `10.1007/978-3-319-46466-4_28`.

Zawieska, D. and J. Markiewicz (2016). "Development of Photogrammetric Documentation of the Borough at Biskupin Based on Archival Photographs - First Results". In: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection.* Springer International Publishing, pp. 3–9. DOI: `10.1007/978-3-319-48974-2_1`.

Zhou, Z., F. Farhat and J. Z. Wang (2017). "Detecting Dominant Vanishing Points in Natural Scenes with Application to Composition-Sensitive Image Retrieval". In: *IEEE Transactions on Multimedia* 19, 12, pp. 2651–2665. DOI: `10.1109/tmm.2017.2703954`.

5 Fully automated pose estimation of historical images

Chapter 5 published under CC-BY license in "ISPRS International Journal of Geo-Information" (eISSN: 2220-9964) as:

Fully Automated Pose Estimation of Historical Images in the Context of 4D Geographic Information Systems Utilizing Machine Learning Methods

Authors: Ferdinand Maiwald[1,2], Christoph Lehmann[3], Taras Lazariv[3]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]Chair for Digital Humanities, FSU Jena, Germany
[3]Centre for Information Services and High Performance Computing, TU Dresden, Germany

Publication history: submitted September 2021, accepted November 2021, published November 2021

Full reference: Maiwald, F.; Lehmann, C.; Lazariv, T. Fully Automated Pose Estimation of Historical Images in the Context of 4D Geographic Information Systems Utilizing Machine Learning Methods. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 748. `https://doi.org/10.3390/ijgi10110748`.

Abstract: The idea of virtual time machines in digital environments like hand-held virtual reality or four-dimensional (4D) geographic information systems requires an accurate positioning and orientation of urban historical images. The browsing of large repositories to retrieve historical images and their subsequent precise pose estimation is still a manual and time-consuming process in the field of Cultural Heritage. This contribution presents an end-to-end pipeline from finding relevant images with utilization of content-based image retrieval to photogrammetric pose estimation of large historical terrestrial image datasets. Image retrieval as well as pose estimation are challenging tasks and are subjects of current research. Thereby, research has a strong focus on contemporary images but the methods are not considered for a use on historical image material. The first part of the pipeline comprises the precise selection of many relevant historical images based on a few example images (so called query images) by using content-based image retrieval. Therefore, two different retrieval approaches based on Convolutional Neural Networks (CNNs) are tested, evaluated, and compared with conventional metadata search in repositories. Results show that image retrieval approaches outperform the metadata search and are a valuable strategy for finding images of interest. The second part of the pipeline uses techniques of photogrammetry to derive the camera position and orientation of the historical images identified by the image retrieval. Multiple feature matching methods are used on four different datasets, the scene is reconstructed in the Structure-from-Motion software COLMAP, and all experiments are evaluated on a newly generated historical benchmark dataset. A large number of oriented images, as well as low error measures for most of the datasets, show that the workflow can be successfully applied. Finally, the combination of a CNN-based image retrieval and the feature matching methods SuperGlue and DISK show very promising results to realize a fully automated workflow. Such an automated workflow of selection and pose estimation of historical terrestrial images enables the creation of large-scale 4D models.

Keywords: historical images; pose estimation; photogrammetry; 4D-GIS; cultural heritage; automation; feature matching; image retrieval; instance retrieval; convolutional neural network

5.1 Introduction

The idea of virtual time machines in digital environments like hand-held Virtual Reality (VR) or four-dimensional (4D) Geographic Information Systems (GIS) requires an accurate positioning and orientation (=pose) of historical images. This enables texturing three-dimensional (3D) models (see Figure 5.1) and offers new perspectives of spatial distributions of historical data (Niebling et al., 2020) as these photographs are sometimes the only visual remainder of buildings due to renovation or destruction of the respective object. The range of historical image data over time represents the fourth dimension as a temporal component, allowing the user to travel through space and time, and thus, makes cultural heritage tangible.

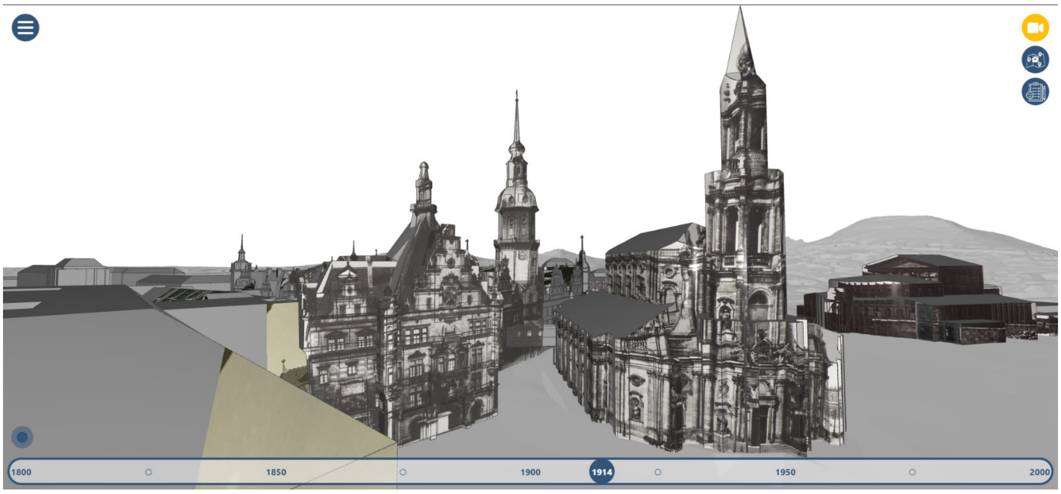


Figure (5.1): Example for projection of historical images on simple 3D models (under development) in application VR City `https://4dcity.org/`, accessed on 20 Sep 2021.

With the ongoing mass-digitization in archives and libraries, an increasing amount of image data became available to the public. Nowadays, the quality aspects of image data such as a higher resolution and lossless data formats, allow the use of automatic image analysis on historical image data. However, often the repositories show varying metadata and usability quality (Evens and Hauttekeete, 2011; Münster et al., 2018), and thus, the usage of historical data material for a photogrammetric reconstruction of buildings and structures relies heavily on archive browsing and manually selecting appropriate sources (Maiwald et al., 2017; Bevilacqua et al., 2019; Condorelli and Rinaudo, 2018; Kalinowski et al., 2021). This contribution tackles this problem focusing on a completely automated retrieval *and* pose estimation workflow for historical terrestrial images using deep learning methods.

The starting point is getting images of interest from an image database or repository. This step can be automated by using content-based image retrieval, whereby retrieval methods based on convolutional neural networks are used. Our approach provides the option to crawl an image database using multiple (in our experiments three) query images containing the object of interest. As metadata in databases and repositories is often inconsistent, misleading, or incorrect, it makes sense not to rely only on a metadata search and to apply image retrieval.

The retrieval result is directly used as input for an automatic 6-degrees of freedom (6-DoF) pose estimation workflow, i.e., calculation of the camera position and camera orientation up to scale (Fig. 5.2).

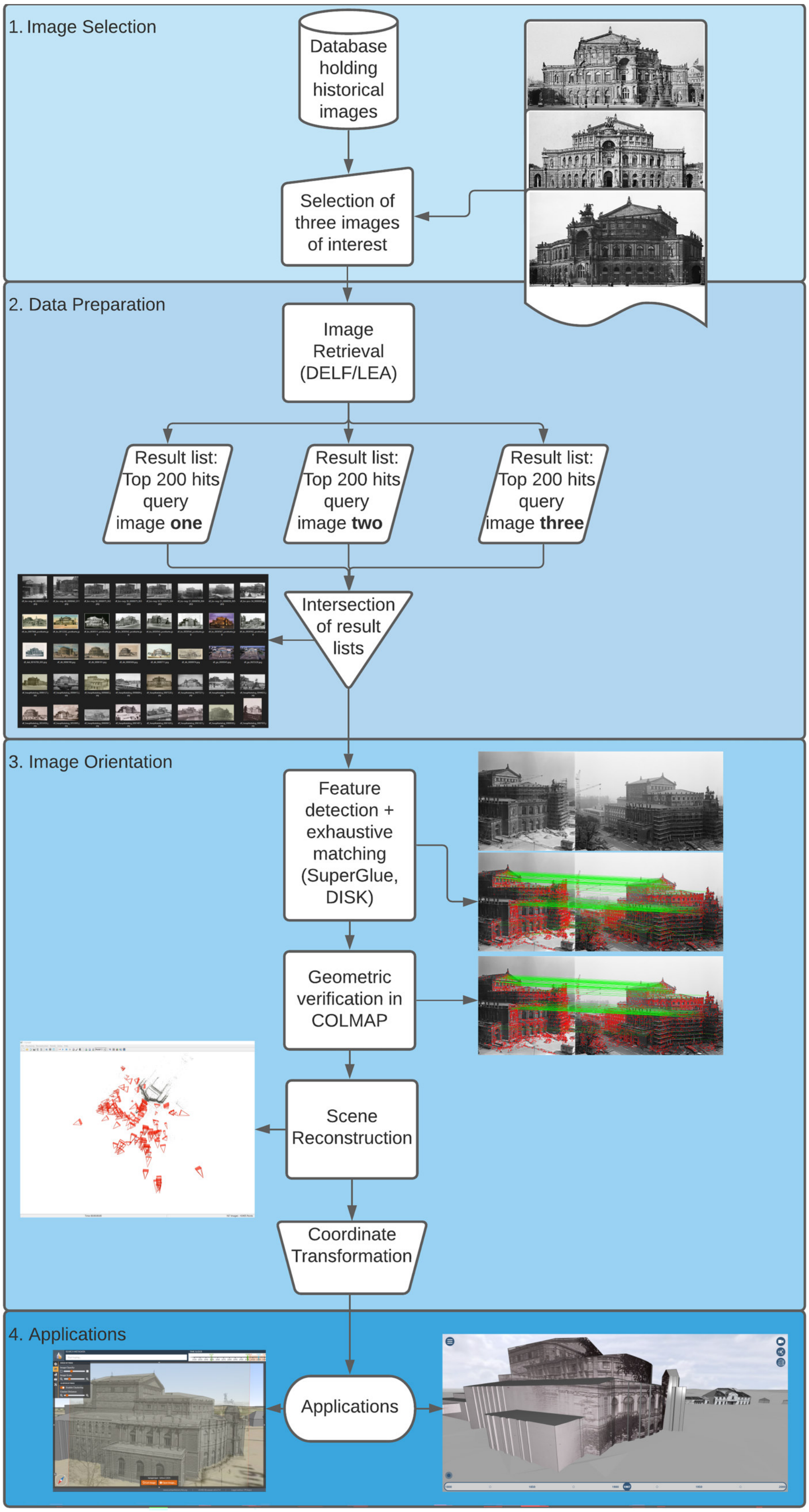

1. Image Selection
Database holding historical images
Selection of three images of interest
2. Data Preparation
Image Retrieval (DELF/LEA)
Result list: Top 200 hits query image **one**
Result list: Top 200 hits query image **two**
Result list: Top 200 hits query image **three**
Intersection of result lists
3. Image Orientation
Feature detection + exhaustive matching (SuperGlue, DISK)
Geometric verification in COLMAP
Scene Reconstruction
Coordinate Transformation
4. Applications
Applications

Figure (5.2): Workflow from an image database to scene reconstruction using image retrieval and photogrammetric methods.

Especially the automatic detection of distinctive features and their matching are the most difficult tasks when calculating the image pose (Maiwald and Maas, 2021). This is because the images were mostly made by different people at different times in no photogrammetric context, e.g., a convergent image block with significant overlap between the images. Additionally, large baselines, occlusions, environmental changes, and scale differences hamper the feature matching process. Thus, more advanced methods which are possibly able to deal with these difficulties are tested and evaluated.

However, for the evaluation of feature matching methods, no or only small datasets consisting of exclusively historical images are available. This contribution tries to overcome this issue, by creating and publishing four datasets with a mimimum of 20 images with manually determined feature points suitable for a Structure-from-Motion (SfM) workflow `http://dx.doi.org/10.25532/OPARA-146`, accessed on 03 Nov 2021. This benchmark data allows an evaluation of the performance of five different feature matching methods, all relying on deep neural networks.

In the second evaluation, all the methods are once again used to process the larger datasets derived by the preceding image retrieval. This workflow shown in Figure 5.2 and explained in detail in Section 5.4, enables the calculation of the orientation parameters of a large percentage of the historical images by only selecting three query images. The estimated poses can be transferred to different browser applications.

5.2 Related Work

This section intends to give a (nonexhaustive) overview of methods for image retrieval and feature matching in the photogrammetric context of calculating the camera position and orientation of historical images. For a summary of 3D building reconstruction approaches based on the works of Snavely et al. (2006) and Schindler and Dellaert (2012) using historical images refer to Maiwald and Maas (2021).

5.2.1 Image Retrieval

Content-based image retrieval (AKA content-based instance retrieval) is a classical task that computer vision is dealing with during the last decades already (e.g., cf. (Datta et al., 2008)). Thereby, it turned out to be a challenging task and was tackled with many different approaches. Image retrieval approaches based on deep neural networks show good or at least promising performance in comparison to former standard approaches as, e.g., VLAD, Bag-of-Words or improved Fisher vectors (IFV) (cf. ((Sharif Razavian et al., 2014), Section 4, Table 7), or ref. (Wan et al., 2014)). A good overview on the development of further more recent retrieval methods can be found in Zheng et al. (2018). The approach of (p. 255, Razavian et al. (2016)) that is based on a Convolutional Neural Network (CNN) even outperformed the common image representations as VLAD and IFV. Using a CNN for this task is plausible as such a network is able to catch the characteristics of an image, e.g., edges, shapes, lines within so-called feature maps (cf. (Zeiler and Fergus, 2014)). These feature maps can be used as descriptors of an image. By decomposing an image into several subimages, there are created multiple local descriptors as proposed in Razavian et al. (2016). In Maiwald et al. (2019b) and Münster et al. (2020) the retrieval approach of Razavian et al. (2016) was already successfully applied to heterogeneous historical images and it showed promising results to identify images that contain some given object of interest.

One approach to improve the usage of local descriptors is called attentive Deep Local Feature (DELF), cf. (Noh et al., 2017). Therein, the relevance of single local descriptors is learned by a second neural network. Furthermore, potential matches are geometrically verified with the Random Sample Consensus (RANSAC) algorithm, cf. (Fischler and Bolles, 1981). For image retrieval, the article at hand is intended to compare the approaches of Razavian et al. (2016) and Noh et al. (2017) and to investigate their utilization for finding relevant images for purposes of photogrammetry. To the best of our knowledge content-based image retrieval has not yet been applied exclusively to historical images within photogrammetry.

5.2.2 Feature Detection and Matching

While there is little research focusing on the evaluation of feature matching methods using exclusively historical images (Maiwald and Maas, 2021) a lot of methods are tested on difficult benchmark datasets, e.g. the Photo Tourism dataset (Winder and Brown, 2007) and the Aachen Day-Night dataset (Sattler et al., 2012). With an increasing use of neural networks for feature detection, matching, and scene reconstruction an overview of well-performing, recently published methods using diverse network architectures is given in this section.

Several studies show that the commonly used algorithmic ("handcrafted") method Scale Invariant Feature Transform (SIFT) (Lowe, 2004) in combination with advanced matching methods is still able to perform comparable to deep local invariant features (Balntas et al., 2017; Schönberger et al., 2017; Csurka et al., 2018; Jin et al., 2020). Thus, the training of a neural network for an end-to-end pipeline from feature detection to feature description up to feature matching seems like a promising approach. Other approaches even go further and include the final camera pose in the network architecture (Sarlin et al., 2021). The network should get an understanding what makes a good feature, how to describe it distinctively, how to combine multiple matches to generate a robust matching solution, and how to estimate a reasonable and accurate pose.

A lot of different approaches using deep neural networks were already developed in the past. Some of them learn optimized descriptors (Jahrer et al., 2008; Simo-Serra et al., 2015; Tian et al., 2017; DeTone et al., 2018), others combine the descriptor learning with a metric learning approach to include the feature matching into the network (Han et al., 2015; Zagoruyko and Komodakis, 2017).

As a further improvement, multiple recent methods focus on the joint detection *and* description of features in a single network architecture (Yi et al., 2016; Mishchuk et al., 2017; Ono et al., 2018). A focus lies especially on the methods D2-net (Dusmanu et al., 2019) and R2D2 (Revaud et al., 2019), which are used in the photogrammetric workflow and evaluation (see Section 5.4) of this contribution.

D2-Net is a winner of the Image Matching Challenge of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019) and already produced good results in recent research on historical images (Maiwald and Maas, 2021). The method jointly detects and describes stable feature matches using a CNN which extracts a set of feature maps for every single image. A local descriptor is obtained by traversing all feature maps at a spatial position in the image while the keypoint detections are obtained by performing a non-local-maximum suppression on a feature map followed by a non-maximum suppression across each descriptor (Dusmanu et al., 2019).

To train the network, D2-Net uses a loss function jointly optimizing the detection and description objectives by minimizing the distance of corresponding descriptors while maximizing the distance to other distracting descriptors.

R2D2 follows a similar approach to D2-net and jointly detects and describes keypoints by training a fully-convolutional network. The method achieved superior results compared to D2-net on benchmark datasets (Jin et al., 2020). R2D2 uses a slightly modified L2-Net (Tian et al., 2017) as a backbone and adds two different layer structures to obtain a respective map for repeatability and reliability (Revaud et al., 2019). This allows separate learning and optimizing of repeatability and reliability for the matching process. Still, both methods rely on a feature matching strategy using Lowe's ratio test.

As final and most recent approaches, end-to-end pipelines (including detection, description and matching) are compared, especially SuperGlue (Sarlin et al., 2020) and DISK (Tyszkiewicz et al., 2020). For the new promising approach Pixloc (Sarlin et al., 2021), which also integrates the 6-DoF-Pose into the neural network to learn the feature, the code was just recently published and could not be tested yet. In the strict sense, SuperGlue is only a feature matching method, however, the authors combine it in a complete pipeline using the SuperPoint feature detector (DeTone et al., 2018). The pipeline is a winner of the CVPR Image Matching Challenge 2020. SuperPoint uses a modified VGG16 network (Simonyan and Zisserman, 2014) as encoder and add two different decoders for keypoint detection as well as description.

Firstly, this feature detector is pretrained on synthetic data and is referred to as MagicPoint. Secondly, the geometric consistency of MagicPoint is improved using Homographic Adaption. This process is self-supervised and allows an iterative repetition. The resulting model is called SuperPoint and is used for the feature detection in the SuperGlue workflow.

SuperGlue introduces a novel method for the matching of the derived features using a graph neural network to solve a differentiable optimal transport problem where relevant information is directly learned from the data. SuperGlue uses two major components. An attentional graph neural network is used with a keypoint encoder to map feature positions and their descriptors into a single vector. Then, more powerful representations are generated in self- and cross attention layers. In the second component, an optimal matching layer finds the optimal partial assignment using the Sinkhorn algorithm (Sarlin et al., 2020).

DISK is a complete pipeline for extracting, describing and matching image features. The process of feature extraction is based on a modified U-net architecture (Ronneberger et al., 2015) while optimizing the absolute number of feature matches using Reinforcement Learning with positive and negative rewards for correct and incorrect matches, respectively. Therefore, gradients of expected rewards are estimated using Monte Carlo sampling and the gradient ascent is used for maximizing the quantity of obtained feature matches.

5.3 Data Preparation: Image Retrieval

This section of the article deals with the image retrieval process for getting images of interest that are processed later by methods of photogrammetry. Thereby, the idea is extended by looking in particular at experimental settings that correspond to the real situation a researcher in photogrammetry faces when searching for images that contain the Object of Interest (OoI). More specifically, it is investigated to which extent an image retrieval approach is suitable for refining search results from a metadata search, replacing the manual process of filtering hundreds or even thousands of images. Two CNN-based retrieval approaches are considered: (1) the approach according to Razavian et al. (2016), that will be called Layer Extraction Approach (LEA) in the following, (2) the approach according to Noh et al. (2017), which is the DELF approach. For every image retrieval there are two types of images: (a) the query image with the OoI, (b) the reference images that are to be compared.

Both approaches were implemented from scratch based on the corresponding publications. The implementations were realized in Python 3.8 with Pytorch 1.7. All experiments were performed on Nvidia A100 GPUs.

5.3.1 Experiment and Data

The experiments are intended to compare the performance between metadata (MD) search and image retrieval (IR). Thereby, the basic idea is to use the MD search as a kind of prefiltering and, in a second step, to improve the result quality by applying an IR to the MD search results. The concrete procedure of the experiment is as follows:

1. A fixed number of OoI in the vicinity of Dresden, Germany is defined: Frauenkirche, Hofkirche, Moritzburg, Semperoper, Sophienkirche, Stallhof, Crowngate

2. For each OoI, a MD search is performed returning a list of results. Note that each of these result lists contains a different number of images across the different OoI.

3. The result list from step 2. is sorted by two criteria: by name and recording date (ascending and descending each). As an outcome, there is a total of four result lists from the MD search for every single OoI.

4. To perform IR, for each OoI three query images are defined. Based on each query image, both IR approaches (LEA, DELF) are applied to the result list of every OoI from step 2. The outcome here is one result list per query image with the most similar images at the top.

5. All result lists from step 3. and step 4. are evaluated for the top-200 positions (more details on evaluation in Section 5.3.3).

The MD search was conducted via the Saxon State and University Library Dresden (SLUB) using the browser-based search interface in Deutsche Fotothek and resulted in a total of 4,737 images. These image data cover a time range from about 1850–2005 and contain historic photographs and a few drawings. Examples of the images are shown in Figure 5.7 in the Photogrammetry section. All found images were annotated manually whether they match the defined OoI or not. These chosen OoI are seven different sights in the vicinity of Dresden. As some buildings are located very close to each other in the inner city, several OoI may appear simultaneously on one image. Thus, some images are assigned to multiple OoI.

The data contain different file formats (jpg, tif, gif) and color spaces, as well as varying resolutions, with smallest images of size 257×400 and largest of size $3,547 \times 2,847$. During a preprocessing step all images were converted into jpg-format with RGB color channels without changing the resolution.

5.3.2 Methods

This section sketches the two IR approaches used within the experiments. Both are based on a CNN. More precisely, LEA is based on VGG-16 and DELF is based on Resnet50. The basic idea is similar for both: single network layers are used as a numerical representation of a single image for comparing and assessing the distance between many images. Using CNNs for that task seems plausible, as this type of neural networks is able to catch the main characteristics of an image, like edges, shapes, lines, etc.

For the sake of simplicity, the following method descriptions are mainly intended to convey the basic idea of an approach as well as the relevant parameters and their impact on the application. For further details we refer to the corresponding publications.

5.3.2.1 Layer Extraction Approach (LEA)

The IR was implemented according to Razavian et al. (2016). This approach is based on a pretrained CNN (here: VGG-16 from Simonyan and Zisserman (2014)), where one of the upper layers (here: the last max-pooling layer) is used as a vector-based image representation for each image. The output of the last max-pooling layer is 512 so-called feature maps, each with dimension 7×7. According to (p. 253, Razavian et al. (2016)), these feature maps are reduced in dimension by using spatial pooling (aka global max-pooling). This standard operation of max-pooling (cf. (Section 5.1.2, (Chollet, 2018))) can be performed using different parameters, mainly kernel size and stride. The resulting set of feature maps is flattened to a vector whose length depends on the parameters used. Within the experiments, two different parameter settings for max-pooling are used:

1. kernel size 7×7 with stride 0, leading to resulting feature maps of size 1×1 each (i.e., maximum reduction); flattened to vector of size $512 \times 1 \times 1 = 512$
2. kernel size 4×4 with stride 3, leading to resulting feature maps of size 2×2 each; flattened to vector of size $512 \times 2 \times 2 = 2,048$

The resulting vector representations within each setting have the same size for each image and are used to calculate a distance measure between the query image and each reference image. Here, the Euclidean distance is used. Within a last step, an ordered rank-list is created based on the distances calculated.

The result of a single IR based on one query image is a sorted list (rank-list) of images, containing the distance between the query image and each reference image, with best matches on top.

To improve the retrieval results, the query images and the reference images are divided into subpatches (part of image by cropping) according to (p. 254, Equations (1) and (2), (Razavian et al., 2016)). The level of division is characterized by the order parameter L. The number of resulting subpatches is determined by $\sum_{i=1}^{L} i^2$, e.g., an order of $L = 4$ leads to 30 subpatches. Within the experiments in Razavian et al. (2016) an order of $L = 3$ and $L = 4$ led to good results.

For the application at hand, we use a fixed query order of 3, i.e., 14 sub-patches for the query images. The number of subpatches of reference images, denoted as reference order, is varied for $L = 3$ and $L = 4$ within different experiment settings. Using subpatches of query and/or reference images, the above procedure of distance calculation is applied for every single pair of subpatches. Thereby, these distances need to be aggregated into one final distance number, which is done here following (p. 255, Equations (3) and (4), Razavian et al. (2016)). The chosen parameters (query order, reference order, max-pooling) for LEA are motivated from Razavian et al. (2016) and, especially for heterogeneous historical images, from Münster et al. (2020).

Moreover, in (p. 256, Razavian et al. (2016)) all images were resized to 600 × 600 pixels as preprocessing step and this size turned out to provide the best retrieval results based on the datasets used there. In contrast, we found that the retrieval results are quite robust against resizing and that there is no relevant change in the retrieval results. Furthermore, in (p. 255, Razavian et al. (2016)) a principal component analysis (PCA) and whitening is used after the max-pooling to further reduce the dimension of the vector representations. This kind of postprocessing turned out to be beneficial for the retrieval quality, cp. (p. 255, Table 1, Razavian et al. (2016)) or cf. Jégou and Chum (2012). Nevertheless, we could not observe any relevant change or even structural improvement of retrieval results while using such kind of postprocessing. One reason for this could be that, unlike Razavian et al. (2016) and Jégou and Chum (2012), we do not use fixed benchmark datasets, but real data that potentially varies widely across retrievals. Overall, for sake of simplicity we do not report the retrieval results from resizing and PCA.

5.3.2.2 Attentive Deep Local Features (DELF) Approach

The approach in this section is based on an attentive local feature descriptor and is referred to as DELF (Deep Local Feature) (Noh et al., 2017). The DELF IR pipeline consists of four steps: (1) dense localized feature extraction; (2) feature selection; (3) dimensionality reduction; (4) indexing and retrieval.

In the first step, the dense features are extracted by using the convolutional neural network ResNet50 (He et al., 2016), trained on ImageNet (Russakovsky et al., 2015) as a baseline. As in the original paper, the weights of the ResNet50 neural network are fine-tuned on the annotated dataset of landmark images (Google Landmark Challenge V2 (Weyand et al., 2020)), to achieve better discriminative power of the descriptors for the landmark retrieval problem. The output of the forth convolutional layer is used to obtain feature maps.

In the second step, to select a subset of relevant features from dense feature maps an attention-based selection technique is used. Therefore, a different neural network (attention network) is trained from scratch (again on Google Landmark Challenge V2), to learn which feature maps are relevant. This restriction to relevant features helps to improve accuracy and to increase computational efficiency. For more details the reader is referred to the original paper, ref. (Noh et al., 2017).

The selected feature maps with 1,000 features per image, each of size 1,024, are still too large to run efficient comparisons. Therefore, similarly as in the original paper, dimensionality reduction with PCA is applied. After normalization, the feature vectors of up to 500 images are collected and $k \in \{40, 120, 200\}$ principal components are calculated for the reference dataset.

In the last step, the actual matching of the query image to each image in the reference dataset is realized. The reference images are sorted according to the similarity to the query image. The Euclidian distance between the pairs of feature vectors is calculated and compared to a distance threshold $T \in \{1.0, 1.2, 1.4\}$. If the distance remains below the threshold, it is noted as a potential match and is discarded otherwise. Subsequently, potential matches are geometrically verified. If the images contain the same OoI, there exists an affine transformation, which transforms the positions of the features on the query image to those on the reference image. The RANSAC algorithm (Fischler and Bolles, 1981) is used for geometric verification. The number of verified matches is used as a measure of similarity, so-called score.

The distance threshold T has substantial impact on the retrieval results and should be chosen very carefully. Too large values of T lead to many false matches of feature maps within the considered image pair, while too small values provide no matches at all. To simplify the choice of the threshold, the distribution of scores can be evaluated by means of, e.g., a histogram. Ideally, the distribution should not be highly skewed to the left or right. For the application, here we found (using the histogram) $T = 1.2$ to deliver good results.

Another parameter, the number of principal components k, has less influence on the retrieval quality than the distance threshold. Here, we could find a good compromise between accuracy (large k) and computational performance of the retrieval (small k). The evaluation of retrieval results for different values of k are presented in Figure 5.5. Thus, we extend the investigation from Noh et al. (2017) who used the parameters for PCA $k = 40$ and distance threshold $T = 0.8$ and no other parameter settings were considered.

5.3.3 Results and Evaluation

The following results are based on the experimental setting described above in Section 5.3.1. More specifically, each metadata (MD) search result for the 7 objects of interest (OoI) considered can be sorted by recording date and location, in ascending and descending order, respectively. This leads to 28 result lists over all 7 OoI for the MD search. For each result list from the MD search an image retrieval (IR) is performed for improving the order of the list. These image retrievals are based on 3 different query images for every OoI, leading to 3 (potentially improved) result lists. Eventually, there are 7 OoI, 4 sorting criteria for MD search and 3 query images per OoI—leading to a total of $7 \times 4 \times 3 = 84$ result lists for each of the two approaches (LEA and DELF). As already described above, LEA and DELF can be configured by choosing different parameters like reference order, max-pooling parameters (for LEA) or distance threshold, PCA dimensions (for DELF). Within the experiments at hand, 3 different configurations for LEA and DELF were investigated. In total, this leads to $84 \times 3 \times 2 = 504$ result lists.

A single result list is a rank-list based on the top-200 images from the MD search. The sorting criteria are differing according to the applied method as follows:

- MD search: sorted by recording date and location, in ascending and descending order, respectively
- LEA, DELF: sorted by distance measure depending on the approach (top ranked positions have small distances to query image)

Based on the image annotation, every image can be categorized as “hit” or “miss” referring the OoI. A resulting ordered rank-list can be evaluated for each position on the list:

- relevant image *before* or at the considered position is a true positive (TP) or a hit
- irrelevant image that occurs *before* or at the considered position is a false positive (FP) or a miss

For evaluating these result lists, so-called precision-recall (PR) curves are used. Thereby, precision is defined as (cf. Ting (2016)) $precision = \frac{TP}{TP+FP}$. Recall is the fraction of collected hits w.r.t. to the total of possible hits. For every single position of the rank-list, precision and recall are calculated. Together they constitute a point on the PR curve (recall on x-axis; precision on y-axis). The perfect PR curve is a horizontal line at one, that describes a situation where all hits are directly one after another. More practically, a good PR curve tends to the upper-right corner. Note, that a result list is cut after the last possible hit, i.e., the last position always is a hit and all possible hits are collected therewith (i.e., recall equals one). Such a single PR curve is aggregated into one number by averaging the precision values along the curve, that leads to average precision (AP) ranging in the unit interval $[0, 1]$.

Figure 5.3 shows a selection of PR curves for both, MD search and IR (LEA and DELF) over all three queries for the seven different OoI. The selection of these curves is intended to provide a coarse overview of the range of quality of the search results.

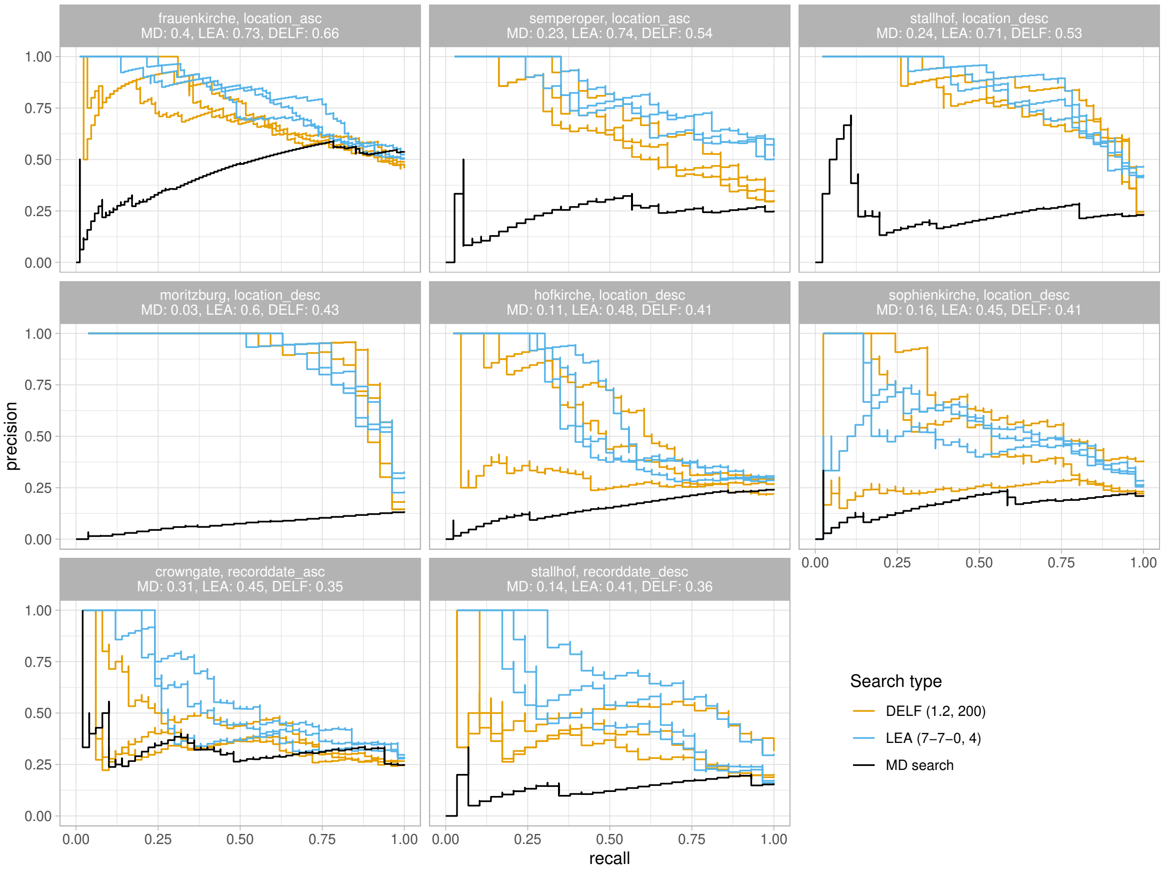


Figure (5.3): Selection of PR curves over all OoI for MD search and IR (LEA and DELF). Values in each subfigure refer to AP for MD search and mAP for LEA and DELF aggregating the corresponding PR curves.

The interpretation for OoI "Frauenkirche" (Fig. 5.3, top-left), for example, could be as follows. Finding roughly 50% (recall) of the OoI Frauenkirche means that approximately 70–90% (precision) of all images found to this point in the IR result list (blue and yellow curves) belong to OoI Frauenkirche. On the other hand, the precision of the MD search (black curve) for Frauenkirche is only 50%. From Figure 5.3, IR is leading to an improvement of MD search, i.e., the colored PR curves from IR (LEA, blue; DELF, orange) are substantially above the PR curve from the MD search in black. Furthermore, there is a slight difference in quality between the two IR approaches, LEA and DELF. This is indicated by the mean average precision (mAP). The mAP is an aggregation of multiple PR curves by calculating the mean over their single AP. The values in each subfigure refer to AP for MD search and mAP for LEA and DELF aggregating the corresponding PR curves. In Figure 5.3, the mAP is calculated separately for LEA and DELF, which shows slightly higher values for LEA.

To get the overall picture, Figures 5.4 and 5.5 contain the AP values of IR while considering different parameter settings (max-pooling and reference order for LEA; PCA dimensions for DELF. The data shown in Figures 5.4 and 5.5 refer to result lists that contain at least 20 hits for the OoI after MD search. This number of hits is regarded as a critical level due to practical considerations as this number of images is usually necessary to achieve enough redundancy for the SfM workflow.

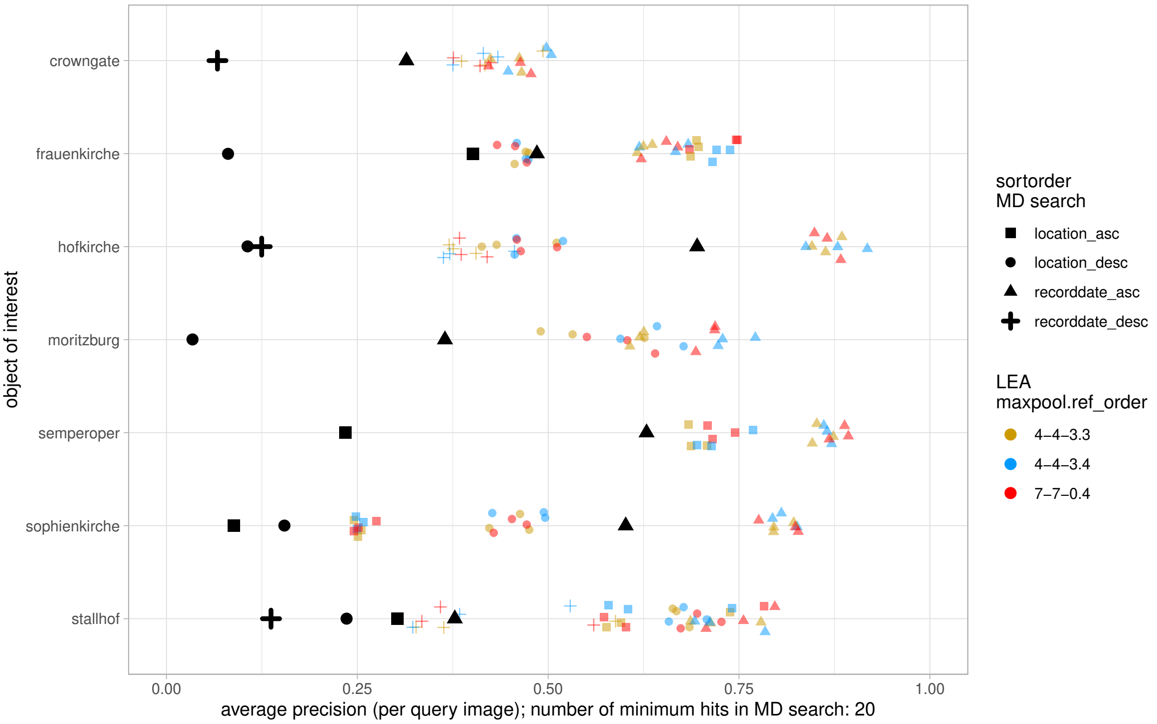


Figure (5.4): Average precision for MD search and LEA IR over all seven OoI.

To understand these figures, the following hints are essential:

- Basically, Figures 5.4 and 5.5 are to be read line by line for the different OoI.
- For each OoI, the AP of the MD search is marked with a large black symbol (square, circle, etc.) as a reference value.
- The 4 different symbol shapes (square, circle, triangle, cross) refer to the considered sorting order for the MD search.
- Within each OoI there might be less than 4 symbol shapes due to the requirement of at least 20 hits from the MD search.

- Besides the black symbols, the other colors refer to different parameter settings of the IR approach under consideration. Since the IR is based on 3 query images, there are 3 symbols of the same shape and color within each line for each IR setting.
- Finally, symbols of the same shape belong together within each line, but can be compared across the lines for comparing different OoI as well.

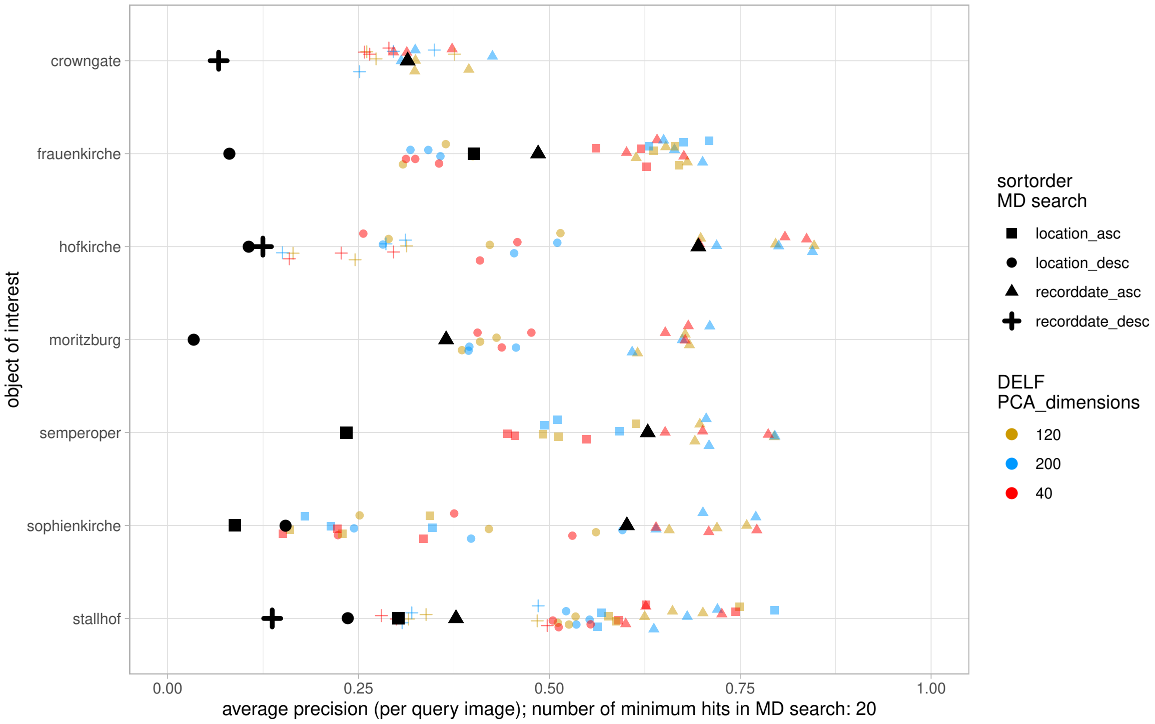


Figure (5.5): Average precision for MD search and DELF IR over all seven OoI.

In the following relevant observations that hold for both image retrieval approaches, LEA and DELF are listed.

1. Most sorting criteria for MD search provide poor result lists with AP below 30%. This magnitude quantifies the trouble of practitioners (see Section 5.1) while searching images of relevance in databases and repositories.
2. There are differences across the OoI referring to the IR quality and its deviation. Thereby, LEA performs slightly better and with less deviation than DELF.
3. IR improves the quality of every MD search result list. More precisely, the improvement factor ($AP_{\text{LEA}}/AP_{\text{MD}}$) for LEA is around 2.6 in 50% of the cases (factor 3.2, 75% of the cases). The improvement factor ($AP_{\text{DELF}}/AP_{\text{MD}}$) for the DELF approach is around 1.9 in 50% of the cases (factor 2.7, 75% of the cases).
4. Within the parameter settings illustrated there is no clear dependency on IR quality, i.e., no setting is preferred.

In summary, the results for AP are worse than those in Razavian et al. (2016) or Noh et al. (2017). Probably, this is caused by using different underlying data: the application at hand deals with a real dataset, that comes along with great heterogeneity in color, resolution, and image type, whereas benchmark datasets were used in Razavian et al. (2016) and Noh et al. (2017). In that sense, the data situation here is a realistic one and it provides insights about what is possible under real-world conditions.

Nevertheless, the absolute values for mAP in the experiments at hand are still in the middle lower range of values for the quality of different IR approaches for the Oxford5k dataset (approx. 50–80%) as reported in (p. 1236, Zheng et al. (2018)). Generally, from the practitioner's perspective an IR (no matter if LEA or DELF) substantially outperforms the MD search and can provide strong support for identifying images of interest.

5.4 Camera Pose Estimation of Historical Images Using Photogrammetric Methods

The main idea of this contribution is to use the output of the IR as a input for the pose estimation workflow. Usually, the simultaneous determination of 6-DoF pose and 3D object points for contemporary images is realized in conventional SfM software like e.g. Agisoft Metashape (proprietary) or Meshroom (open-source). However, these solutions often fail to orient historical images (Maiwald et al., 2019b) due to their heterogeneous radiometric and geometric properties.

Neural networks allow a more distinctive detection, description, and matching of features by training on thousands of ground truth image pairs. While these feature points and matches cannot yet be imported into the software solutions above, the presented method uses COLMAP (Schönberger and Frahm, 2016) to process the feature matches derived by several neural networks to reconstruct a scene including camera positions, orientations and a sparse point cloud. COLMAP is the standard tool for benchmark tests and evaluations of the Image Matching Challenge of the IEEE Conference on Computer Vision and Pattern Recognition.

Additionally, a benchmark dataset is generated and described to evaluate the results of the five different tested feature matching methods. All methods are also applied to larger datasets, which are directly derived by the image retrieval approach LEA (see Section 5.3.2.1) to bypass the process of manual image selection. The accuracy of the reconstruction and the total number of registered images can then again be verified by the benchmark dataset.

5.4.1 Data

For the evaluation of the automatic pose estimation workflow, two different lists (benchmark and retrieval) of respectively four datasets are created. The four datasets consist of historical terrestrial images in the vicinity of Dresden and all images originate from the SLUB. The analogue images are digitized by the the Deutsche Fotothek and can be downloaded in .jpg file format usually with a maximum edge length of 1,600 pixels via `http://www.deutschefotothek.de/`, accessed on 20 Sep 2021. To compare the results to previous research, four different landmarks (Crowngate, Hofkirche, Moritzburg, Semperoper) are chosen. An overview of all datasets and their characteristics is given in Table 5.1.

Table (5.1): Overview of eight datasets including total number of images, time span, and characteristics.

Dataset	Size	Landmark	Time Span	Characteristics
1	20	Crowngate	1880–1994	repetitive patterns, symmetry
2	33	Hofkirche	1883–1992	wide baselines, radiometric differences
3	24	Moritzburg	1884–1998	building symmetry
4	20	Semperoper	1869–1992	terrestrial and oblique aerial images
5	188	Crowngate	∼1860–2010	∼1,000 keyword hits
6	176	Hofkirche	∼1860–2010	∼3,000 keyword hits
7	200	Moritzburg	∼1884–2010	∼2,700 keyword hits
8	197	Semperoper	∼1869–2010	∼2,100 keyword hits

5.4.1.1 Benchmark Datasets

The four benchmark datasets with a small sample size consist of manually selected photographs mainly showing the corresponding landmark as a whole. The datasets' size varies between 20 and 33 images and all show variations in radiometric and geometric image quality with some particular challenging characteristics.

All images of the dataset Crowngate (of the Dresden Zwinger) show a high number of repetitive patterns, especially windows. Furthermore, the appearance of the landmark changed during time. From 1924 to 1936, a park in front of the monument was replaced by a water ditch (Marx, 1989).

For the Hofkirche dataset, it can be assumed that this will be the most difficult but also interesting dataset showing vast radiometric and geometric differences between the different images. It is the largest of the datasets.

The baroque palace Moritzburg castle has a very symmetric architecture with four almost identical looking round towers on an artificial island. Thus, it is difficult to distinguish between the different sides of the building.

The dataset of the Semperoper consists of terrestrial images with different viewpoints as well as oblique aerial images causing wide baselines and scale changes.

For all these images, there is (almost) no information available about camera type, camera model, and the digitization process. Nonetheless, this contribution tries to provide historical reference data to estimate the quality of the different feature matching methods (see Section 5.4.2).

To this end, for the benchmark datasets a minimum of 15 homologue points are interactively selected per single image of the respective dataset which allows a scene reconstruction and the calculation of reference poses up to scale for all images. It can be assumed that the image points can be determined to an accuracy of 1–2 pixels in the historical images while this process is very time-consuming. For selecting points, Agisoft Metashape is used and the tie points' coordinates (markers) are exported via XML. In the following, the coordinates of the selected homologue points are imported simultaneously with the images into a COLMAP database. Then, the scene is reconstructed up to scale using the bundle adjustment in COLMAP and setting the minimum number of inliers to 15 (see Section 5.4.2.2). This manual approach is evaluated using the reprojection error in pixels derived by COLMAP as a control measure as in equation (5.1),

$$\text{reprojection error} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_i)^2 + \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y}_i)^2}, \tag{5.1}$$

where (x_i, y_i) represent the image coordinates, i.e., the position of the matched 2D points, and $(\bar{x}_i, \bar{y}_i)$ are the reprojected values of the computed 3D coordinates within the bundle adjustment procedure (Remondino et al., 2017).

The reprojection error for all reconstructions is approximately 1 pixel (see Table 5.2) which verifies the assumed accuracy of tie point selection and the entirety of images could be oriented in the process. A visual control confirms that all images are pointing into the correct direction with respect to the sparse model and all projection centers are in a reasonable position (see Figure 5.6). As another value the 95% quantile of the reprojection error is calculated. This provides additional insight on the quality of the complete benchmark reconstruction as 95% of the single reprojection error of every 3D points and its corresponding image point are below that threshold.

Table (5.2): Determined features, reprojection error, and 95% quantile of the reprojection error of the benchmark datasets.

Dataset	Size	Landmark	Selected Features	Reprojection Error (px)	Reprojection Error$_{0.95}$ (px)
1	20	Crowngate	419	0.84	1.68
2	33	Hofkirche	1,108	1.23	2.10
3	23	Moritzburg	947	1.15	1.73
4	20	Semperoper	540	1.13	1.86

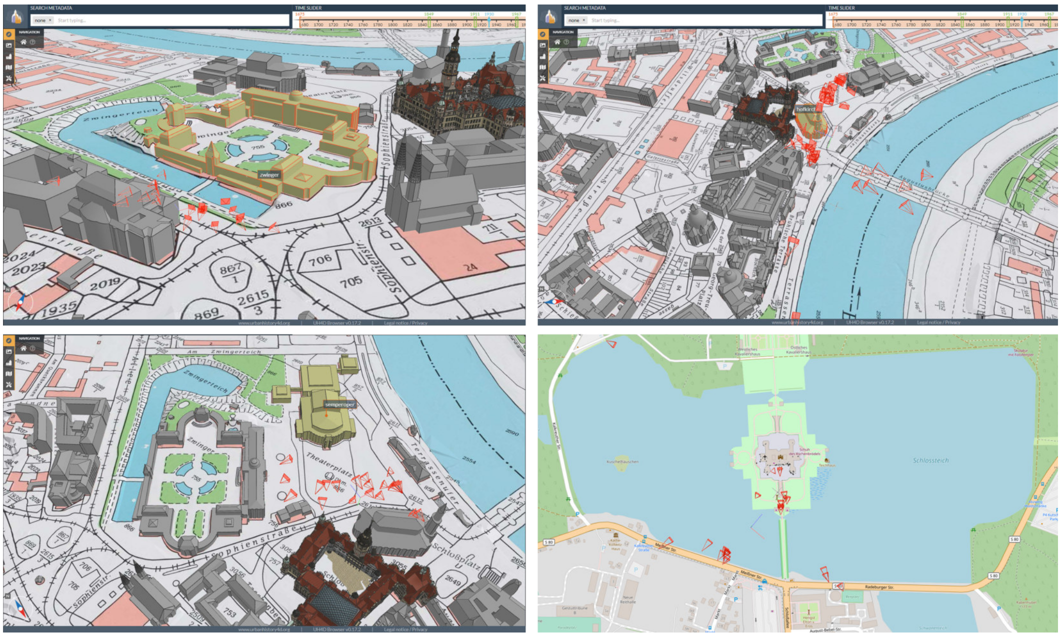


Figure (5.6): Reconstructed camera positions and orientations (small red frustums) derived by COLMAP manually blended into 3D/4D environment of 4D-browser (`https://4dbrowser.urbanhistory4d.org/`, accessed on 20 Sep 2021 for better scene understanding. Since there is no 3D model of Schloss Moritzburg yet, the OpenStreetMap (OpenStreetMap contributors, 2017) representation is used.

In the worst case, homologue image points do not necessarily belong to the similar object point in 3D because of building settlement, changes, destruction, or restorations during the vast time span. Coarse errors are filtered in the process of bundle adjustment but smaller errors could persist. Thus, we do refer to the oriented data as benchmark dataset (and not as ground truth). The benchmark data is especially useful to detect coarse outliers and to estimate the overall quality of the feature matching methods with the provided error measures (see Section 5.4.3). Due to copyright issues, it is not allowed to share the images directly, but all permalinks, specifications, license information, feature matches, and reference poses can be obtained via `http://dx.doi.org/10.25532/OPARA-146`, accessed on 03 Nov 2021.

5.4.1.2 Retrieval Datasets

Referring to the idea of a fully automated workflow (see Section 5.1) four datasets of the same landmarks are generated using exclusively the image retrieval process (see Section 5.3). Since the depicted approaches LEA and DELF show comparable performance, one fixed workflow is chosen to equally generate all retrieval datasets. For the experiments, LEA with a kernel size 4×4 with stride 3 and a reference order 4 is selected. To apply the method, the entirety of images of the resulting keyword search ("Kronentor", "Hofkirche", "Schloss Moritzburg", "Semperoper") are downloaded via the Deutsche Fotothek. Then, three different high quality images showing the landmark in full size (preferably from different perspectives) are manually chosen as query images (see Figure 5.7).

The image retrieval method yields a sorted list (rank-list) where all reference images are sorted according to their Euclidean distance to the query image. The first 200 images are taken for each query and the intersection of these three results forms the respective retrieval dataset. For three different query images the total 200 images shrink only from 0% up to 12% (see Table 5.1), and thus, it can be assumed that the most relevant and significant images of the landmark are kept in the process. Looking at the resulting data (Fig. 5.2, 2. Data Preparation), this method is capable of finding exclusively images of exterior views of the landmarks while also retrieving different perspectives around the building.

In contrast to the benchmark datasets, these images are not oriented interactively because of the extremely time-consuming procedure (clicking a minimum of 3,000 tie points for all datasets). However, images which coincidentally appear also in the benchmark datasets can be compared and used for transformation and accuracy estimation.

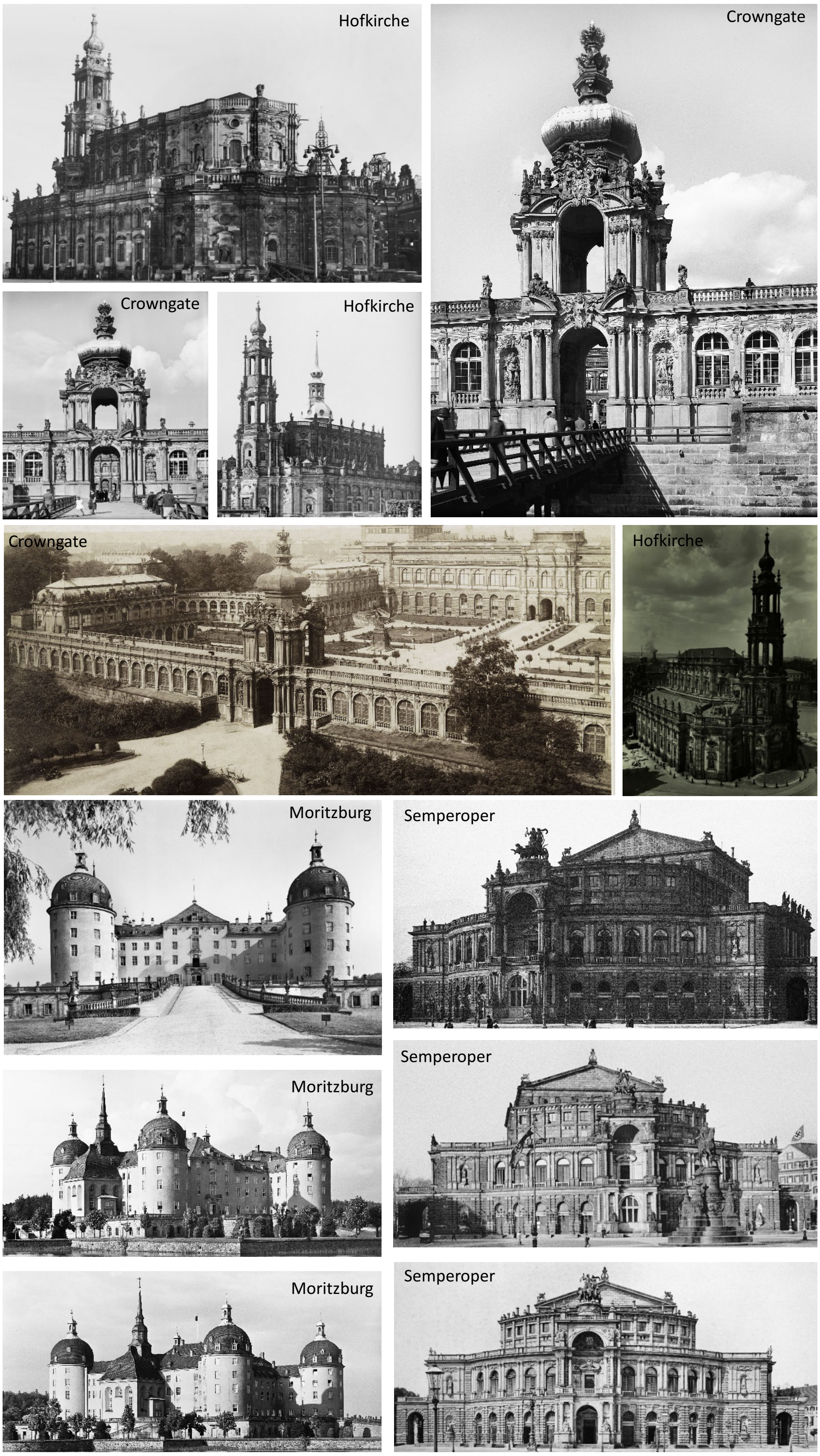


Figure (5.7): All query images used for automatic selection of retrieval datasets using LEA approach.

5.4.2 Methods

In the following, the full workflow beginning with feature detection up to calculation of the camera pose is explained in detail. The different feature matching methods are briefly explained and compared, and parameter settings are described. Several recent methods deriving features by using neural networks are chosen due to their performance on other challenging datasets. The goal is to retrieve the position of the projection center as well as the interior and exterior camera orientation parameters of a large number of historical images completely automatic up to scale. The images can then be transferred to a VR application or used for texture projection of 3D building models.

5.4.2.1 Feature Detection and Matching

Distinctive feature detection and matching in historical image pairs with large radiometric and geometric differences is a difficult task. Specialized feature detectors have to be used to enable successful feature matching (Maiwald et al., 2019b). This contribution compares five different feature matching methods on eight different datasets. The aim is, firstly, to maximize the quality of the feature detection and matching, and secondly, to make the different methods comparable.

From a user's perspective, we try to follow the descriptions and recommendations given in the corresponding publications and do not change elementary parts like the proposed matching method. In the following, a brief explanation of each method is given (see also Section 5.2) in the order of their publication and all modifications and parameter changes are explained. An example for detected keypoints by the different methods is given exemplarily for two historical images in the appendix (Fig. 5.A1). The process of feature detection and matching for the large datasets with approximately 200 images has a runtime of 5 min on a NVIDIA V100 GPU (DISK implementation). In contrast, the slowest approach is the D2-net multiscale implementation which runs only on one CPU and takes around 60 min for the same amount of images. All implementations are based on sequential processing of the images. Thus, there is plenty room for optimization (parallelization, GPU usage), which was not part of this investigation.

D2-Net (Single-Scale and Multiscale)

D2-Net jointly detects and describes features using a CNN (Dusmanu et al., 2019). In our experiments, we use the precomputed fine-tuned Caffe VGG16 (Simonyan and Zisserman, 2014) weights trained on the MegaDepth dataset (Li and Snavely, 2018) by Dusmanu et al. (2019) (d2_tf.pth). Features are calculated for every single image using the single-scale and the multiscale approach of D2-Net, respectively. The single-scale (-ss) option only processes the input image with a default maximum image width of 1600 pixels. The multiscale (-ms) approach calculates multiple sets of feature maps for an image pyramid consisting of half resolution, full resolution, and double resolution of the input image. The derived feature maps are resized according to the selected resolution using bilinear interpolation, enabling keypoint detection across the different resolutions.

For matching the detected features the proposed approach by Dusmanu et al. (2019) is used. Outliers are removed using a mutual nearest neighbor matching including Lowe's ratio test (Lowe, 2004). That means, features are only kept as matches if the feature point P_m in image i maps only to the feature point P_n in image j and vice versa. Additionally, the distance to the second-closest neighbor has to be smaller than a certain ratio threshold Dusmanu et al.

(2019) recommend a threshold of 0.90 for the off-the-shelf descriptors and 0.95 for the fine-tuned ones. The application on historical images has shown that this threshold is critical for retrieving robust results (Maiwald and Maas, 2021). Considering these findings, the threshold was slightly modified using 0.96 for the single-scale and 0.97 for the multiscale approach.

R2D2

R2D2 also jointly detects and describes features but uses two different layer structures to obtain a different map for reliability and repeatability of the features (Revaud et al., 2019). In our experiments, we use the precomputed model from Revaud et al. (2019). The model was trained with Web images (W), Aachen day-time images (A), Aachen day-night synthetic pairs (S), and Aachen optical flow pairs (F) and can be obtained via `https://github.com/naver/r2d2`, accessed on 20 Sep 2021 under the name r2d2_WASF_N16.pt. Features are then extracted using slight modifications. The maximum image width is set to 1,600 pixels and the maximum number of keypoints is fixed to 4,096 to allow a better comparison to the other approaches.

For matching the obtained keypoints a simple mutual nearest neighbor approach without a threshold is used, similarly done in the experiments by Revaud et al. (2019).

SuperGlue

SuperGlue (Sarlin et al., 2020) combines a feature matching method using a graph neural network with the feature detector SuperPoint (DeTone et al., 2018) in an end-to-end pipeline. In our experiments, we use the precomputed outdoor weights (superglue_outdoor.pth) trained on the MegaDepth dataset. Furthermore, we use the pipeline in `https://github.com/cvg/Hierarchical-Localization`, accessed on 20 Sep 2021 and allow a maximum image size of 1,600 instead of 1,024 for the superpoint_aachen configuration (cf. Sarlin et al. (2019)). The maximum number of detected keypoints is kept at the default value of 4,096 for a fair comparison.

DISK

Discrete Keypoints (DISK) is an end-to-end pipeline to learn local features relying on the policy gradient approach (Tyszkiewicz et al., 2020). The network is trained on the MegaDepth dataset (Li and Snavely, 2018) and we use the provided weights for our experiments. As the pipeline requires an input size of a multiple of 16 we use an image size of 1,600 × 1,600 pixels. Additionally, the maximum number of detected keypoints is set to 4,096. The recommended ratio threshold of 0.95 is used for matching the features.

5.4.2.2 Geometric Verification and Camera Pose Estimation

Almost all feature matching methods described, recommend the use of COLMAP for geometric verification of the features and calculation of the camera pose using the integrated bundle adjustment. As already shown, COLMAP is also used to reconstruct the camera positions from the benchmark datasets in a similar workflow (see Section 5.4.1.1).

Geometric verification implies, that for every image pair combination with feature matches derived by the prior methods, a Fundamental Matrix is calculated. If a minimum of 15 inliers is found via Locally-Optimized RANSAC (LO-RANSAC) (Chum et al., 2003) the image pair is considered valid and added to the scene graph (cf. (Schönberger and Frahm, 2016)). The scene graph holds the geometrically verified image pairs, their associated inlier correspondences and their geometric relation (cf. (Schönberger and Frahm, 2016)).

This structure is used to determine an initial image pair and incrementally register new images. The triangulated points and the derived camera poses are improved in a bundle adjustment with a maximum number of 100 iterations.

To improve the absolute number of registered images, two-view tracks are allowed in our workflow. That means, that features are used for the reconstruction even if they are only detected in two images (i.e., one image pair). For all historical images some assumptions have to be made. A simple radial camera model is chosen, modeling the principal distance (focal length) *f*, principal point coordinates c_x, c_y and only one radial distortion parameter *k*. As an initial approximation, f is set per default to $f = 1.25 \cdot \max(width_{px}, height_{px})$ and c_x, c_y lie in the image center.

These assumptions could be far off estimates considering that historical images have to be digitized and the geometric relation to the principal point's coordinates as well as to the principal distance is completely lost in the process. This may produce outliers which could not be filtered by the benchmark datasets because the same approximations are made for that approach.

As a result, the workflow provides one (or sometimes multiple) model(s) for each of the eight datasets. The model consists of the oriented images and their translation and rotation in the local coordinate system. Each image is introduced as a separate camera and the adjusted interior orientation can be exported. Additionally, a sparse point cloud is created and the number of 3D points, the reprojection error and further statistics are shown in COLMAP.

5.4.3 Results and Evaluation

All feature matching methods used, are evaluated by the number of oriented images, the pose error (i.e., euclidean distance between benchmark and estimated projection centers) and the angle error α in degree as in Equation (5.2),

$$2 \cdot \cos(\alpha) = \mathrm{trace}(R_{\mathrm{benchmark}}^{-1} R_{\mathrm{est}}) - 1, \tag{5.2}$$

where $R_{\mathrm{benchmark}}$ and R_{est} denote the Rotation matrices of the benchmark solution and the estimated solution, respectively. As in refs. Sattler et al. (2018), Jin et al. (2020) and Li and Ling (2021) the angle error is our main error metric to evaluate the results of the different reconstructions. Therefore, the reconstruction solution derived by the different methods has to be aligned with the benchmark poses. This requires a seven-parameter transformation (Helmert transformation) from the local coordinate frame of the automatic estimated solution into the local coordinate frame of the benchmark solution. This contribution uses the functionality of COLMAP to compute the alignment between the two reconstructions.

Images that are registered in both reconstructions are selected and an alignment is estimated using LO-RANSAC and the corresponding projection center coordinates. After that, the alignment is verified by reprojecting common 3D point observations. For LO-RANSAC a threshold has to be set to distinguish between inlier and outlier camera positions. This threshold is intentionally set to 0.05, which is slightly higher than necessary, to enable the transformation to the benchmark for almost all datasets. Then, it is possible to determine the quality and accuracy of the respective reconstruction using the mean and median values of the pose and angle errors.

Table (5.3): Results of different feature matching procedures tested on eight datasets in comparison to benchmark data. Table shows respectively number of oriented images, reprojection error in pixels, Euclidean distance (scale-less) and angle error in degree. Best results are highlighted in bold.

	Dataset and Parameters	D2-Net-ss	D2-Net-ms	R2D2	SuperGlue	DISK
Crowngate (20 images)	Oriented images	17	18	12	**19**	18
	Reproj. error	1.1	1.2	**0.9**	1.3	0.9
	Pose error (mean)	8,245.1	5.5	0.9	**0.8**	11.7
	Angle error (mean)	173.3	169.9	12.5	**4.3**	172.8
	Angle error (median)	174.1	173.3	10.7	**4.0**	172.6
Hofkirche (33 images)	Oriented images	27	28	NA	30	25
	Reproj. error	1.0	1.0	NA	1.3	0.8
	Pose error (mean)	4.8	4.6	NA	5.0	5.1
	Angle error (mean)	93.1	106.5	NA	88.5	92.2
	Angle error (median)	80.7	102.6	NA	108.6	153.3
Moritzburg (23 images)	Oriented images	20	13	14	**23**	20
	Reproj. error	1.1	1.1	0.8	1.2	**0.8**
	Pose error (mean)	3.4	6.2	2.9	1.7	**0.7**
	Angle error (mean)	33.0	159.3	43.7	**5.8**	7.2
	Angle error (median)	30.1	159.6	42.9	**0.8**	2.6
Semperoper (20 images)	Oriented images	19	19	14	**20**	19
	Reproj. error	1.0	1.1	0.8	1.2	**0.8**
	Pose error (mean)	16,042.5	0.9	2.2	**0.3**	**0.3**
	Angle error (mean)	7.8	7.3	7.1	3.0	**2.1**
	Angle error (median)	3.1	3.2	7.6	3.0	**2.1**
Crowngate (188 images, 12 in common)	Oriented images	175	178	178	184	**183**
	Reproj. error	1.3	1.4	1.1	1.4	**1.1**
	Pose error (mean)	3.6	1.1	1.2	1.2	**0.8**
	Angle error (mean)	164.0	65.6	176.9	169.4	**10.2**
	Angle error (median)	164.8	65.6	177.4	169.1	**10.7**
Hofkirche (176 images, 15 in common)	Oriented images	155	160	166	157	150
	Reproj. error	1.3	1.3	1.0	1.4	1.1
	Pose error (mean)	2.9	2.6	2.6	2.1	3.8
	Angle error (mean)	76.1	74.5	73.4	70.7	71.3
	Angle error (median)	7.4	6.0	4.0	5.7	2.6
Moritzburg (200 images, 15 in common)	Oriented images	129	151	NA	**176**	155
	Reproj. error	1.4	1.4	NA	1.3	**1.0**
	Pose error (mean)	0.7	2.2	NA	0.5	**0.3**
	Angle error (mean)	5.2	66.3	NA	4.2	**2.0**
	Angle error (median)	4.7	66.2	NA	4.1	**1.8**
Semperoper (197 images, 10 in common)	Oriented images	135	135	150	**167**	148
	Reproj. error	1.2	1.2	**0.9**	1.3	**0.9**
	Euclidean distance	1.9	0.8	0.4	**0.3**	**0.3**
	Angle error (mean)	111.2	6.1	**2.5**	2.6	2.8
	Angle error (median)	110.8	6.1	**2.3**	2.5	2.6

Lowering the LO-RANSAC threshold often leads to the result, that the two reconstructions cannot be aligned because there are too many outliers and no further statements about the accuracy could be made.

After the alignment, the mean pose and angle error are calculated for every corresponding image of both reconstructions. Additionally, the median is calculated over all angle errors. Here, the median provides a measure, whether the whole scene reconstruction is not well aligned or whether there are only few or even no incorrect camera positions. Since the pose error is dependent on the scale of the reconstruction it can be compared for all methods within a single dataset only. All results are shown in Table 5.3.

The analysis of the different models and their resulting error values allows to draw the following conclusions. First of all, it seems that the reprojection error is not a significant value to estimate the correct position and rotation of the historical images. For all different COLMAP models it varies between 0.8 and 1.4 but has no or only a small impact on the result, which is better described by the distance and error measures.

Especially, a combination of both error measures—in this particular case a mean pose error $<$ 1.0 and a mean angle error $<$ 5.0°—provides also visually good reconstruction results with no outlier camera positions and orientations. These values are in a range comparable to other state-of-the-art unordered datasets (Sattler et al., 2018; Jin et al., 2020; Li and Ling, 2021). Table 5.3 also shows that for all datasets, DISK and/or SuperGlue provide the highest number of oriented images as well as the most accurate results.

R2D2 performs very good on the Semperoper retrieval datasets but otherwise falls behind the two other methods. As already expected, the Crowngate and Hofkirche datasets are the most difficult to process for the methods. As the Crowngate shows a very symmetric structure all methods are mixing left- and right-side as well as exterior and interior views of the building also shown by angle errors close to 180°.

SuperGlue produces the only valid reconstruction for the Crowngate benchmark dataset while DISK is able to process the retrieval dataset still including multiple incorrectly registered images, which is reflected in the angle error of approximately 10°. For the Hofkirche dataset similar deviations from the real scene occur. The methods are not able to close the large viewpoint gap between the front of the building and the western side (see Figure 5.6).

This results in an erroneous pose estimation of images and the creation of multiple point clouds of the Hofkirche in one single reconstruction. For this particular building, it could be an option to close the gap using contemporary images to improve the image registration (Maiwald et al., 2017).

All methods are able to additionally filter the results of the image retrieval, e.g., for the Semperoper, 197 images are found by LEA. About 15 photographs show the old opera house before 1878, plans or drawings which are not included in the models because they are filtered during geometric verification. However, COLMAP is able to assign these 15 image to a separate reconstruction during bundle adjustment. In conclusion, the described method (LEA+SuperGlue/DISK) enables a completely automatic selection and registration of historical images if the landmark is photographed from multiple surrounding viewpoints.

5.5 Conclusions and Future Work

Basically, image retrieval applied to historical image works can provide assistance to the researcher for automation within a hand-crafted working process. For example, we demonstrated that LEA is able to generate large historical datasets, which would be an extremely time-consuming task using conventional metadata search. Therefore, the approach requires calculation on a GPU to be effectively implemented in practice, which refers especially to the feature extraction based on a CNN for instance retrieval. The usage of transfer learning (i.e., using a pretrained CNN) is very advantageous and drastically reduces the preparation time. All experiments carried out had a runtime of 5–30 min, depending on the concrete parameterization, i.e., retrieving a single query image within approximately 1,000 reference images takes 10–30 s.

Regarding the image retrieval, there are the following concluding remarks. LEA is less sophisticated, but produces better results than DELF. More precisely, the DELF approach with its optimized consideration of local features and a geometric verification did not lead to a substantial improvement of retrieval results. In contrast to LEA and Razavian et al. (2016), the original parameters for DELF from Noh et al. (2017) turned out not to be optimal and esp. DELF led to better results with different parameters. Both approaches lie behind the publications by Razavian et al. (2016) and Noh et al. (2017) in absolute numbers for mAP. The reason for these differences is most likely due to the underlying data: the pretrained networks used (ResNet50 and VGG-16) were trained on data that included historic buildings, but not historic photographs from a technical point of view. This aspect extends the statement from (p. 1240, Zheng et al. (2018)) by a more technical dimension, who tell that the used datasets for image retrieval usually refer to a particular type of instances such as landmarks or indoor objects, etc. The results are not based on a benchmark dataset, but on real, heterogeneous use-case data. Overall, there is still potential for improvement, but the approaches under consideration are already delivering practically relevant gains and making the impossible possible—fully automatic pose estimation of historical images.

It was shown that the image retrieval pipeline can be directly used to receive a large number of relevant images of one building by just selecting three query images. The tests have demonstrated that about 160 out of 190 relevant images can be localized up to scale using an adapted SfM workflow. In previous research, the method D2-net outperformed conventional feature matching methods. However, with the generation of a benchmark dataset and the comparison of several different error metrics, it could be shown that D2-net falls behind more recent methods. The methods SuperGlue and DISK in combination with COLMAP are not only able to match are large amount of images, but are also useful to reject images showing, e.g., different buildings or drawings not suitable for the reconstruction. Especially, for the retrieval datasets Moritzburg and Semperoper, the low mean angle error and the small reprojection error prove the impressing result considering the historical image material. These results could be obtained, though all neural networks were only trained on contemporary image data.

For the Hofkirche, the reconstruction is erroneous because certain perspective views are missing from the image retrieval (also due to their potential absence in the original database). In such a case, where not enough poses could be estimated correctly, we suggest the following strategy: Since the building is still present today, it could be an improvement to close the gap between the main historical views using contemporary images or even generate a contemporary SfM model for comparison.

The established benchmark dataset is to the knowledge of the authors the first dataset available which uses only historical images. The dataset can be easily extended by selecting more tie points and/or historical images and the obtainable data allows to reproduce the published results and also further metrics like homographies or depth maps of the historical image pairs. As the creation of such benchmark data is very time-consuming, this dataset may be of great benefit for the Cultural Heritage community.

For further development of the workflow the following aspects/ideas can be considered. An extension for optimizing the image retrieval results—when using multiple query images—could be an aggregation of the result lists. Note that different retrieval goals can be pursued by using appropriate query images. Thereby, query images can be quite similar in some sense but it also makes sense to use very distinct images. The whole end-to-end pipeline is currently being implemented in a scientific environment while a commercial use is not intended at the moment. Especially for the final web application, it is planned to directly integrate the obtained camera positions and orientations by applying a Helmert Transformation on the automatically retrieved poses. This is still under testing and would allow texture projection as well as immersive coloring of 3D models using historical images.

Author Contributions

Conceptualization, Ferdinand Maiwald and Christoph Lehmann; methodology, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; software, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; validation, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; formal analysis, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; investigation, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; resources, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; data curation, Ferdinand Maiwald; writing—original draft preparation, Ferdinand Maiwald and Christoph Lehmann; writing—review and editing, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; visualization, Ferdinand Maiwald, Christoph Lehmann and Taras Lazariv; funding acquisition, Ferdinand Maiwald and Christoph Lehmann. All authors have read and agreed to the published version of the manuscript.

Funding

The research for this paper was carried out in the project Denkmalschutz4D (Deutsche Bundesstiftung Umwelt, 35654) and in the competence center for Big Data and AI “ScaDS.AI Dresden/Leipzig” (German Federal Ministry of Education and Research BMBF, 01/S18026A-F). The authors received additional funding by the Chair of Photogrammetry, TU Dresden.

Data Availability Statement

The benchmark dataset can be found under `http://dx.doi.org/10.25532/OPARA-146`, accessed on 03 November 2021. The applications including the historical images can be found under `https://4dcity.org/`, accessed on 20 September 2021 and `https://4dbrowser.urbanhistory4d.org/`, accessed on 20 September 2021.

Acknowledgments

The authors gratefully acknowledge the GWK's support for this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden on HRSK-II. The authors would also like to thank Philipp Langen for his support in implementing the DELF approach.

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

4D	four-dimensional
CNN	convolutional neural networks
VR	Virtual Reality
GIS	geographic information system
3D	three-dimensional
6-DoF	six degrees-of-freedom
SfM	Structure-from-Motion
OoI	Object of Interest
LEA	layer extraction approach
MD	metadata
IR	image retrieval
SLUB	Saxon State and University Library Dresden
PCA	principal component analysis
RANSAC	Random Sample Consensus
LORANSAC	locally optimized RANSAC
TP	true positive
FP	false positive
PR curve	precision recall curve
AP	average precision
mAP	mean average precision
GPU	Graphics Processing Unit

5.A Appendix

This appendix contains examples for feature matches using different methods on various historical image pairs.

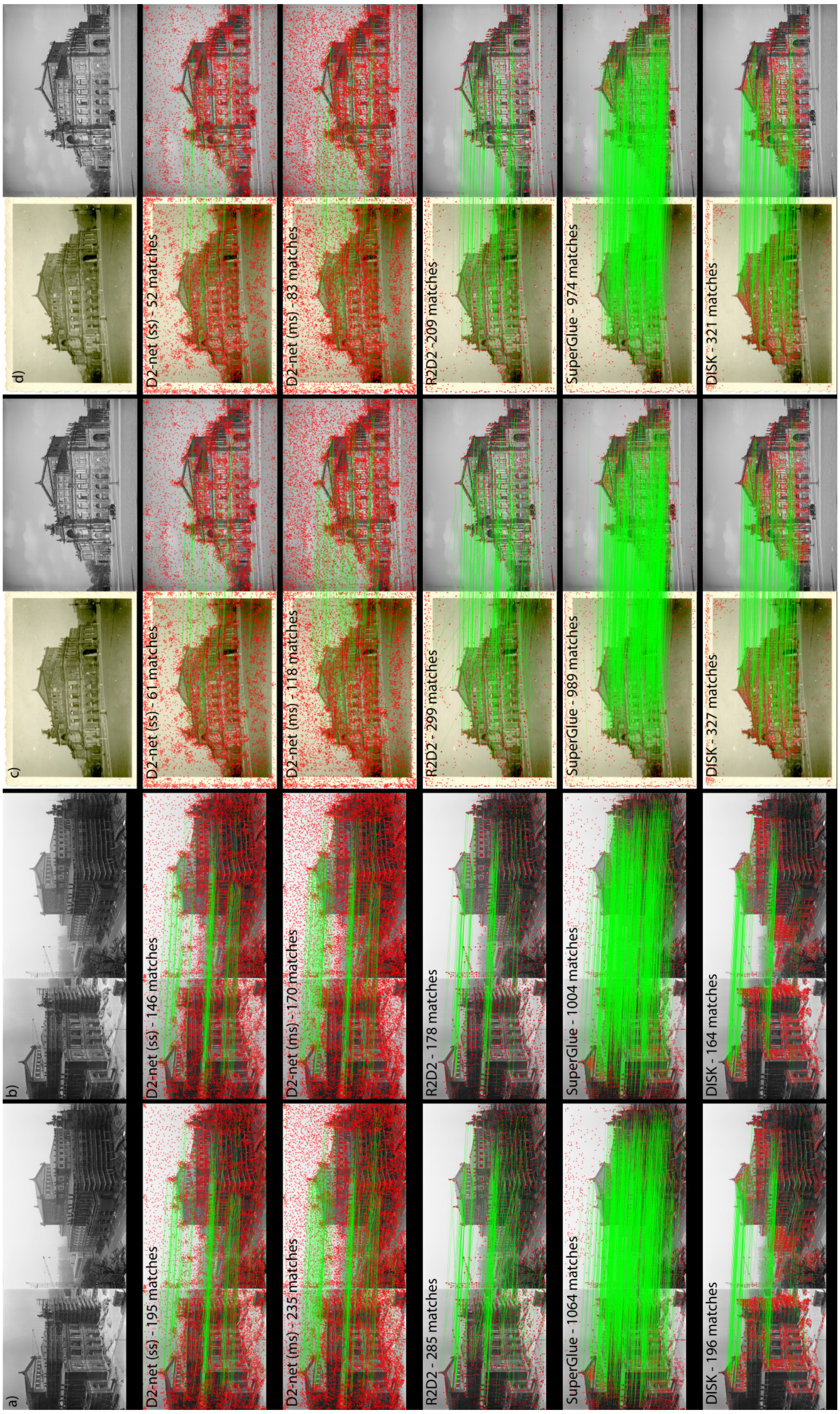


Figure (5.A1): Example for feature matches on two different historical image pairs (a–d). (a,c) show initial keypoints and feature matches derived by different methods D2-net, R2D2, SuperGlue and DISK. (b,d) show resulting remaining feature points after geometric verification in COLMAP.

References

Balntas, V., K. Lenc, A. Vedaldi and K. Mikolajczyk (2017). "HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3852–3861. DOI: 10.1109/cvpr.2017.410.

Bevilacqua, M. G., G. Caroti, A. Piemonte and D. Ulivieri (2019). "Reconstruction of lost Architectural Volumes by Integration of Photogrammetry from Archive Imagery with 3-D Models of the Status Quo". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 119–125. DOI: 10.5194/isprs-archives-xlii-2-w9-119-2019.

Chollet, F. (2018). *Deep learning with Python*. Shelter Island, NY: Manning Publications.

Chum, O., J. Matas and J. Kittler (2003). "Locally Optimized RANSAC". In: *Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 236–243. DOI: 10.1007/978-3-540-45243-0_31.

Condorelli, F. and F. Rinaudo (2018). "Cultural Heritage Reconstruction From Historical Photographs And Videos". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2, pp. 259–265. DOI: 10.5194/isprs-archives-xlii-2-259-2018.

Csurka, G., C. R. Dance and M. Humenberger (2018). "From handcrafted to deep local features". In: *arXiv:1807.10254*, pp. 1–41.

Datta, R., D. Joshi, J. Li and J. Z. Wang (2008). "Image Retrieval: Ideas, Influences, and Trends of the New Age". In: *ACM Computing Surveys* 40, 2, pp. 1–60. ISSN: 0360-0300. DOI: 10.1145/1348246.1348248.

DeTone, D., T. Malisiewicz and A. Rabinovich (2018). "SuperPoint: Self-Supervised Interest Point Detection and Description". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 337–33712. DOI: 10.1109/cvprw.2018.00060.

Dusmanu, M., I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii and T. Sattler (2019). "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8084–8093. DOI: 10.1109/cvpr.2019.00828.

Evens, T. and L. Hauttekeete (2011). "Challenges of digital preservation for cultural heritage institutions". In: *Journal of Librarianship and Information Science* 43, 3, pp. 157–165. DOI: 10.1177/0961000611410585.

Fischler, M. A. and R. C. Bolles (1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24, 6, pp. 381–395. DOI: 10.1145/358669.358692.

Han, X., T. Leung, Y. Jia, R. Sukthankar and A. C. Berg (2015). "MatchNet: Unifying feature and metric learning for patch-based matching". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3279–3286. DOI: 10.1109/cvpr.2015.7298948.

He, K., X. Zhang, S. Ren and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. DOI: `10.1109/CVPR.2016.90..`

Jahrer, M., M. Grabner and H. Bischof (2008). "Learned local descriptors for recognition and matching". In: *Computer Vision Winter Workshop.* Vol. 2, pp. 39–46.

Jégou, H. and O. Chum (2012). "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening". In: *European Conference on Computer Vision.* Springer, pp. 774–787. DOI: `10.1007/978-3-642-33709-3_55`.

Jin, Y., D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi and E. Trulls (2020). "Image Matching Across Wide Baselines: From Paper to Practice". In: *International Journal of Computer Vision* 129, 2, pp. 517–547. DOI: `10.1007/s11263-020-01385-0`.

Kalinowski, P., F. Both, T. Luhmann and U. Warnke (2021). "Data Fusion of Historical Photographs with Modern 3D Data for an Archaeological Excavation – Concept and First Results". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B2-2021, pp. 571–576. DOI: `10.5194/isprs-archives-xliii-b2-2021-571-2021`.

Li, X. and H. Ling (2021). "On the Robustness of Multi-View Rotation Averaging". In: *arXiv:2102.05454*, pp. 1–16. DOI: `https://doi.org/10.48550/arXiv.2102.05454`.

Li, Z. and N. Snavely (2018). "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 2041–2050. DOI: `10.1109/cvpr.2018.00218`.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* 60, 2, pp. 91–110. ISSN: 0920-5691. DOI: `https://doi.org/10.1023/B:VISI.0000029664.99615.94`.

Maiwald, F., J. Bruschke, C. Lehmann and F. Niebling (2019b). "A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR". In: *Virtual Archaeology Review* 10, 21, pp. 1–13. DOI: `10.4995/var.2019.11867`.

Maiwald, F. and H.-G. Maas (2021). "An automatic workflow for orientation of historical images with large radiometric and geometric differences". In: *The Photogrammetric Record* 36, 174, pp. 77–103. DOI: `10.1111/phor.12363`.

Maiwald, F., T. Vietze, D. Schneider, F. Henze, S. Münster and F. Niebling (2017). "Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities". In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 447–452. ISSN: 1682-1750. DOI: `https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017`.

Marx, H. (1989). *Matthäus Daniel Pöppelmann: der Architekt des Dresdner Zwingers.* Leipzig: VEB E.A. Seemann. ISBN: 3363004141.

Mishchuk, A., D. Mishkin, F. Radenović and J. Matas (2017). "Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4829–4840. ISBN: 9781510860964.

Münster, S., C. Kamposiori, K. Friedrichs and C. Kröber (2018). "Image libraries and their scholarly use in the field of art and architectural history". In: *International Journal on Digital Libraries* 19, 4, pp. 367–383. DOI: 10.1007/s00799-018-0250-1.

Münster, S., F. Maiwald, C. Lehmann, T. Lazariv, M. Hofmann and F. Niebling (2020). "An Automated Pipeline for a Browser-based, City-scale Mobile 4D VR Application based on Historical Images". In: *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*. ACM, pp. 33–40. DOI: 10.1145/3423323.3425748.

Niebling, F., J. Bruschke, H. Messemer, M. Wacker and S. von Mammen (2020). "Analyzing Spatial Distribution of Photographs in Cultural Heritage Applications". In: *Visual Computing for Cultural Heritage*. Springer International Publishing, pp. 391–408. DOI: 10.1007/978-3-030-37191-3_20.

Noh, H., A. Araujo, J. Sim, T. Weyand and B. Han (2017). "Large-Scale Image Retrieval with Attentive Deep Local Features". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 3476–3485. DOI: 10.1109/iccv.2017.374.

Ono, Y., E. Trulls, P. Fua and K. M. Yi (2018). "LF-Net: Learning Local Features from Images". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., pp. 6237–6247. DOI: https://arxiv.org/pdf/1805.09662.pdf.

OpenStreetMap contributors (2017). *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org.

Razavian, A. S., J. Sullivan, S. Carlsson and A. Maki (2016). "Visual instance retrieval with deep convolutional networks". In: *ITE Transactions on Media Technology and Applications* 4, 3, pp. 251–258. DOI: 10.3169/mta.4.251.

Remondino, F., E. Nocerino, I. Toschi and F. Menna (2017). "A Critical Review of Automated Photogrammetric Processing of Large Datasets". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W5, pp. 591–599. DOI: 10.5194/isprs-archives-xlii-2-w5-591-2017.

Revaud, J., P. Weinzaepfel, C. D. Souza, N. Pion, G. Csurka, Y. Cabon and M. Humenberger (2019). "R2D2: Repeatable and Reliable Detector and Descriptor". In: *arXiv:1906.06195*, pp. 1–11.

Ronneberger, O., P. Fischer and T. Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science*. Springer International Publishing, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Russakovsky, O., 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 (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115, 3, pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

Sarlin, P.-E., C. Cadena, R. Siegwart and M. Dymczyk (2019). "From Coarse to Fine: Robust Hierarchical Localization at Large Scale". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 12708–12717. DOI: 10.1109/cvpr.2019.01300.

Sarlin, P.-E., D. DeTone, T. Malisiewicz and A. Rabinovich (2020). "SuperGlue: Learning Feature Matching with Graph Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947. DOI: `arXiv:1911.11763[`.

Sarlin, P.-E., A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl and T. Sattler (2021). "Back to the Feature: Learning Robust Camera Localization from Pixels to Pose". In: *CVPR*. URL: `https://arxiv.org/abs/2103.09213`.

Sattler, T., W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl and T. Pajdla (2018). "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 8601–8610. DOI: `10.1109/cvpr.2018.00897`.

Sattler, T., T. Weyand, B. Leibe and L. Kobbelt (2012). "Image Retrieval for Image-Based Localization Revisited". In: *Procedings of the British Machine Vision Conference 2012*. Vol. 1. British Machine Vision Association, p. 4. DOI: `10.5244/c.26.76`.

Schindler, G. and F. Dellaert (2012). "4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections". In: *Journal of Multimedia* 7, 2, pp. 124–131. ISSN: 1796-2048. DOI: `https://doi.org/10.4304/jmm.7.2.124-131`.

Schönberger, J. L. and J.-M. Frahm (2016). "Structure-from-Motion Revisited". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4104–4113. DOI: `10.1109/cvpr.2016.445`.

Schönberger, J. L., H. Hardmeier, T. Sattler and M. Pollefeys (2017). "Comparative Evaluation of Hand-Crafted and Learned Local Features". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6959–6968. DOI: `10.1109/cvpr.2017.736`.

Sharif Razavian, A., H. Azizpour, J. Sullivan and S. Carlsson (2014). "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 512–519. DOI: `10.1109/CVPRW.2014.131.`.

Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer (2015). "Discriminative Learning of Deep Convolutional Feature Point Descriptors". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 118–126. DOI: `10.1109/iccv.2015.22`.

Simonyan, K. and A. Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv preprint arXiv:1409.1556*, pp. 1–14.

Snavely, N., S. M. Seitz and R. Szeliski (2006). "Photo tourism: exploring photo collections in 3D". In: *ACM Transactions on Graphics* 25, 3, pp. 835–846. DOI: `10.1145/1141911.1141964`.

Tian, Y., B. Fan and F. Wu (2017). "L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6128–6136. DOI: `10.1109/cvpr.2017.649`.

Ting, K. M. (2016). “Precision and Recall”. In: *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US, pp. 781–781. ISBN: 978-1-4899-7502-7. DOI: 10.1007/978-1-4899-7502-7_659-1.

Tyszkiewicz, M. J., P. Fua and E. Trulls (2020). “DISK: Learning local features with policy gradient”. In: *arXiv:2006.13566*, pp. 1–15.

Wan, J., D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li (2014). “Deep Learning for Content-Based Image Retrieval”. In: *Proceedings of the ACM International Conference on Multimedia - MM ’14*. ACM Press, pp. 157–166. DOI: 10.1145/2647868.2654948.

Weyand, T., A. Araujo, B. Cao and J. Sim (2020). “Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2575–2584.

Winder, S. A. J. and M. Brown (2007). “Learning Local Image Descriptors”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8. DOI: 10.1109/cvpr.2007.382971.

Yi, K. M., E. Trulls, V. Lepetit and P. Fua (2016). “LIFT: Learned Invariant Feature Transform”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, pp. 467–483. DOI: 10.1007/978-3-319-46466-4_28.

Zagoruyko, S. and N. Komodakis (2017). “Deep compare: A study on using convolutional neural networks to compare image patches”. In: *Computer Vision and Image Understanding* 164, pp. 38–55. DOI: 10.1016/j.cviu.2017.10.007.

Zeiler, M. D. and R. Fergus (2014). “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53.

Zheng, L., Y. Yang and Q. Tian (2018). “SIFT Meets CNN: A Decade Survey of Instance Retrieval”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5, pp. 1224–1244. DOI: 10.1109/TPAMI.2017.2709749.

6 Related publications

This chapter includes cover pages of five associated publications. A short description for every paper is given here to be able to rank the work in the context of this thesis.

The publication shown in Section 6.1 presents preliminary work mainly done interactively. Historical images are selected by hand and it is tested whether state-of-the-art proprietary SfM software is able to estimate the interior and exterior orientation of all images. This is only possible for a small number of images (13 out of 800) which is why first attempts are made to integrate contemporary images.

The publication shown in Section 6.2 presents a self-developed descriptor for feature matching in architectural scenes and especially façades with windows. Features which are extracted from the historical images are general quadrilaterals. The descriptor is able to describe the centroid of the quadrilateral distinctively using the position and orientation of detected neighboring quadrilaterals. While the method is robust for façades with windows, it is not possible to use it as a generally applicable feature descriptor.

The publication shown in Section 6.3 presents mainly a motivation for using Photogrammetry in Cultural Heritage applications while pointing out the still persisting issues in libraries. Especially, the link between image repositories and Web3D, VR and Augmented Reality (AR) is depicted and how photogrammetric methods are necessary to provide exact interior and exterior camera parameters for historical images. First results using specially adapted feature matching methods as Matching On Demand with View Synthesis (MODS) or Radiation-Invariant Feature Transform (RIFT) are outlined. The publication including its presentation won the Best Paper Award and was therefore fundamentally expanded and adapted in order to fit into the photogrammetric context of this thesis in Chapter 3.

The publication shown in Section 6.4 describes the idea of a complete pipeline from image repository to mobile VR application using preliminary results for Content-based image retrieval (CBIR) and the photogrammetric methods. For a subset of historical images it is now possible to determine the interior and exterior orientation using a combination of MODS and RIFT. The feature matches are imported into COLMAP and the bundle adjustment converges for this small dataset. The algorithmic methods briefly presented here cannot fully compete with the learned feature matchers as described in Chapter 4.

The publication shown in Section 6.5 describes the previously presented mobile VR application in more detail. Hereby, the focus lies on software and content design while the photogrammetric pipeline is a minor part. Still, the publication presents the initial combination of CBIR, feature matching with SuperGlue, and the integration in COLMAP. These preliminary results are visually demonstrated, however not tested on other datasets nor compared to other feature matching methods which is done in-depth in Chapter 5.

6.1 Photogrammetric analysis of historical image repositores for virtual reconstruction in the field of digital humanities

Paper 6.1 published in "International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences" (eISSN: 2194-90340).

Authors: Ferdinand Maiwald[1], Theresa Vietze[1], Danilo Schneider[1], Frank Henze[2], Sander Münster[3], Florian Niebling[4]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]Informationsverarbeitung im Bauwesen/Vermessungskunde, BTU Cottbus-Senftenberg, Germany
[3]Media Center, TU Dresden, Germany
[4]Human-Computer Interaction, JMU Würzburg, Germany

Publication history: submitted November 2016, accepted December 2016, published March 2017
Full reference: Maiwald, F., Vietze, T., Schneider, D., Henze, F., Münster, S., and Niebling, F.: PHOTOGRAMMETRIC ANALYSIS OF HISTORICAL IMAGE REPOSITORIES FOR VIRTUAL RECONSTRUCTION IN THE FIELD OF DIGITAL HUMANITIES, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W3, 447–452, `https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017`, 2017.

Abstract: Historical photographs contain high density of information and are of great importance as sources in humanities research. In addition to the semantic indexing of historical images based on metadata, it is also possible to reconstruct geometric information about the depicted objects or the camera position at the time of the recording by employing photogrammetric methods. The approach presented here is intended to investigate (semi-) automated photogrammetric reconstruction methods for heterogeneous collections of historical (city) photographs and photographic documentation for the use in the humanities, urban research and history sciences. From a photogrammetric point of view, these images are mostly digitized photographs. For a photogrammetric evaluation, therefore, the characteristics of scanned analog images with mostly unknown camera geometry, missing or minimal object information and low radiometric and geometric resolution have to be considered. In addition, these photographs have not been created specifically for documentation purposes and so the focus of these images is often not on the object to be evaluated. The image repositories must therefore be subjected to a preprocessing analysis of their photogrammetric usability. Investigations are carried out on the basis of a repository containing historical images of the Kronentor ("crown gate") of the Dresden Zwinger. The initial step was to assess the quality and condition of available images determining their appropriateness for generating three-dimensional point clouds from historical photos using a structure-from-motion evaluation (SfM). Then, the generated point clouds were assessed by comparing them with current measurement data of the same object.

Keywords: Historical images, structure-from-motion, image configuration, point cloud, virtual 3D reconstruction

6.2 Feature matching of historical images based on geometry of quadrilaterals

Paper 6.2 published in "International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences" (eISSN: 2194-90340).

Authors: Ferdinand Maiwald[1], Danilo Schneider[1], Frank Henze[1], Sander Münster[2], Florian Niebling[3]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]Media Center, TU Dresden, Germany
[3]Human-Computer Interaction, JMU Würzburg, Germany

Publication history: submitted January 2018, accepted February 2018, published May 2018
Full reference: Maiwald, F., Schneider, D., Henze, F., Münster, S., and Niebling, F.: FEATURE MATCHING OF HISTORICAL IMAGES BASED ON GEOMETRY OF QUADRILATERALS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2, 643–650, `https://doi.org/10.5194/isprs-archives-XLII-2-643-2018`, 2018.

Abstract: This contribution shows an approach to match historical images from the photo library of the Saxon State and University Library Dresden (SLUB) in the context of a historical three-dimensional city model of Dresden. In comparison to recent images, historical photography provides diverse factors which make an automatical image analysis (feature detection, feature matching and relative orientation of images) difficult. Due to e.g. film grain, dust particles or the digitalization process, historical images are often covered by noise interfering with the image signal needed for a robust feature matching. The presented approach uses quadrilaterals in image space as these are commonly available in man-made structures and façade images (windows, stones, claddings). It is explained how to generally detect quadrilaterals in images. Consequently, the properties of the quadrilaterals as well as the relationship to neighbouring quadrilaterals are used for the description and matching of feature points. The results show that most of the matches are robust and correct but still small in numbers.

Keywords: historical images, image orientation, feature matching, descriptor matching, quadrilaterals, geometry

6.3 Geo-information technologies for a multimodal access on historical photographs and maps for research and communication in urban history

Paper 6.3 published in "International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences" (eISSN: 2194-90340).

Authors: Ferdinand Maiwald[1], Frank Henze[1], Jonas Bruschke[2], Florian Niebling[2]
[1]Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany
[2]Human-Computer Interaction, JMU Würzburg, Germany

Publication history: submitted January 2019, accepted February 2019, published May 2019
Full reference: Maiwald, F., Henze, F., Bruschke, J., and Niebling, F.: GEO-INFORMATION TECHNOLOGIES FOR A MULTIMODAL ACCESS ON HISTORICAL PHOTOGRAPHS AND MAPS FOR RESEARCH AND COMMUNICATION IN URBAN HISTORY, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII-2/W11, 763–769, https://doi.org/10.5194/isprs-archives-XLII-2-W11-763-2019, 2019.

Abstract: This contribution shows ongoing interdisciplinary research of the project HistStadt4D, concerning the investigation and development of different multimodal access strategies on large image repositories. The first part of the presented research introduces different methods of access, where classical analogue access stands in contrast to digital access strategies such as online collections, Web3D, Augmented Reality (AR) and Virtual Reality (VR). We discuss the main persisting issues of libraries, advantages of digital methods, and different access tools. The second part shows technologies and workflows used to create various access possibilities. The photogrammetric and geo-informational work serves as a technical basis for a 3D WebGIS as well as multiple AR/VR applications, which require spatial oriented images, object coordinates, and further spatial data. We introduce a research environment that allows art historians spatial access to historical photography, integrating 3D/4D models with photographic documents of the respective architecture. For dissemination of research results in installations and museums, we present fully immersive VR as well as handheld AR applications allowing users a free exploration of historical photography in a spatial setting.

Keywords: Historical Photographs and Maps, Urban History, 3D Geo-Information System, Photogrammetry, Augmented Reality

6.4 An automated pipeline for a browser-based, city-scale mobile 4D VR application based on historical images

Paper 6.4 published in "SUMAC'20: Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents" (eISBN: 978-1-4503-8155-0).

Authors: Sander Münster[1], Ferdinand Maiwald[1], Christoph Lehmann[2], Taras Lazariv[2], Mathias Hofmann[3], Florian Niebling[4]
[1]Chair for Digital Humanities(Images/Objects), FSU Jena, Germany
[2]Centre for Information Services and High Performance Computing, TU Dresden, Germany
[3]Media Center, TU Dresden, Germany
[4]Human-Computer Interaction, JMU Würzburg, Germany

Publication history: submitted July 2020, accepted August 2020, published October 2020
Full reference: Sander Münster, Ferdinand Maiwald, Christoph Lehmann, Taras Lazariv, Mathias Hofmann, and Florian Niebling. 2020. An Automated Pipeline for a Browser-based, City-scale Mobile 4D VR Application based on Historical Images. In Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC'20). Association for Computing Machinery, New York, NY, USA, 33–40. DOI:https://doi.org/10.1145/3423323.3425748.

Abstract: The process for automatically creating 3D city models from contemporary photographs and visualizing them on mobile devices is now well established, but historical 4D city models are more challenging. The fourth dimension here is time. This article describes an automated VR pipeline based on historical photographs and resulting in an interactive browser-based device-rendered 4D visualization and information system for mobile devices. Since the pipeline shown is currently still under development, initial results for stages of the process will be shown and assessed for accuracy and usability.

6.5 Software and content design of a browser-based mobile 4D VR application to explore historical city architecture

Paper 6.5 published in "SUMAC'21: Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents" (eISBN: 978-1-4503-8155-0).

Authors: Sander Münster[1], Jonas Bruschke[1], Ferdinand Maiwald[1], Constantin Kleiner[1]
[1]Chair for Digital Humanities(Images/Objects), FSU Jena, Germany

Publication history: submitted July 2021, accepted August 2021, published October 2021
Full reference: Sander Muenster, Jonas Bruschke, Ferdinand Maiwald, and Constantin Kleiner. 2021. Software and Content Design of a Browser-based Mobile 4D VR Application to Explore Historical City Architecture. In Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC'21). Association for Computing Machinery, New York, NY, USA, 13–22. DOI:https://doi.org/10.1145/3475720.3484442.

Abstract: The Kulturerbe4D project aims at making the diversity and change processes of architectural monuments in the urban context virtually visible and experienceable, especially for children and young people, but also for residents and tourists. A virtual city tour providing cultural and historical information is to be combined with the transfer of knowledge about monuments, anthropogenic factors of influence, and protective measures. This article focuses on three main challenges in producing city-scale mobile 4D applications: (a) 4D content creation specifically for historical purposes is highly labour intensive, (b) web applications are better accepted by users but require more adoption to cope with technical limitations, (c) historically accurate 4D content is of disperse visual quality and visualization strategies are rarely empirically proven. Within this article we present our research and development work to overcome those issues.

7 Synthesis

The thesis aims to provide a Structure-from-Motion (SfM) workflow which is able to estimate the exterior camera parameters (camera pose) and interior camera parameters of diverse historical images in unordered photo collections. All scientific articles (Chapter 2-5) and their evaluations result in a final recommendation for such a pose estimation pipeline which is presented in this synthesis. A part of the final workflow is already presented in Chapter 5.

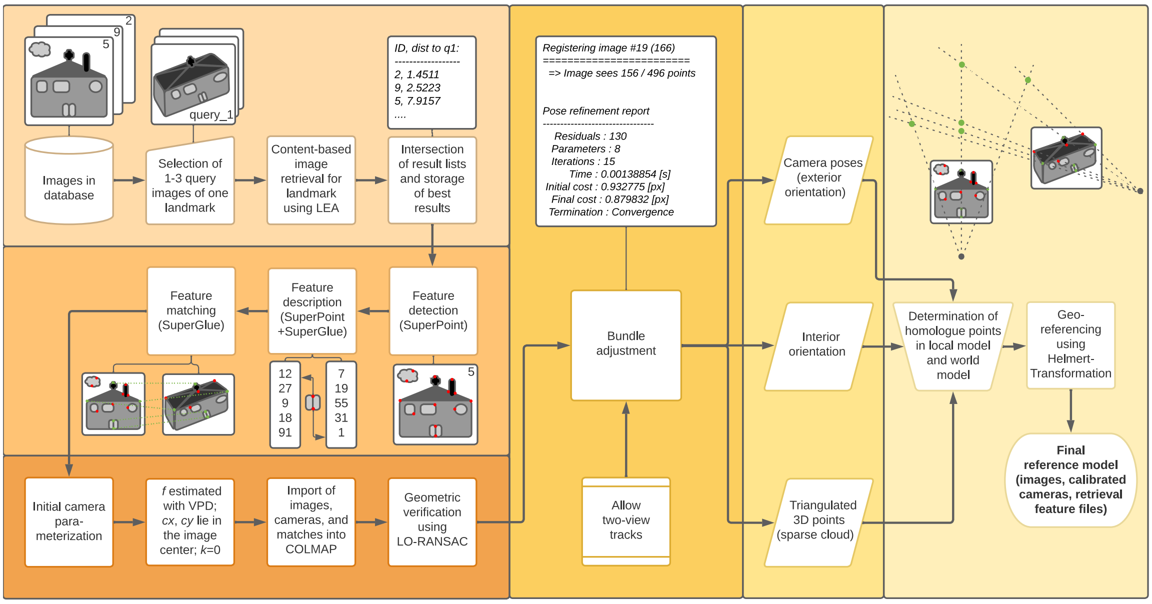
Further, this chapter presents an error assessment of the different steps of the workflow as well as a concise evaluation of accuracy using additional measurements. In order to show the efficacy of the workflow, the SfM pipeline is transferred to several further historical datasets and the results are discussed. An outlook on the development of the used methods is shown and how the different parts can be improved in the future. Finally, a glance on future trends and possible research opportunities is given.

7.1 Summary of the developed workflows

Regarding the evaluation of all steps for the pose estimation and determination of interior camera parameters of historical images, this section intends to condense the most important results into two final workflows.

The first one initializes a reconstruction, i.e. a completely new SfM model (=reference model) is generated for historical (or contemporary) images (Fig. 7.1). The workflow starts with a metadata search for a specific landmark in a repository. Then, three images showing the landmark in full size have to be manually selected. For all images a separate Content-based image retrieval (CBIR) is performed, using Layer Extraction Approach (LEA) with a kernel size 4×4 with stride 3 and a reference order 4. The three result lists are intersected, i.e., only those hits are kept which appear in all lists. Theoretically, it would be possible to use only the two most similar images for the reference reconstruction. In practice, the experiments show that when using only very few images, the latter bundle adjustment and reconstruction of camera poses fails. This is as already described, due to very small baselines or large radiometric differences between the image pairs. For the urban historical images that were used in this thesis showing one landmark in full size, an empirical value of 20 images could be determined.

However, the workflow can also be used for the local pose estimation of more than just 20 images as demonstrated in Chapter 5 and Section 7.1.3. For all images, the SuperGlue feature matching is performed, the matches, the images, and all cameras are initialized in COLMAP, and the feature matches are geometrically verified using Locally-Optimized RANSAC (LO-RANSAC). Then, the scene is reconstructed in a bundle adjustment procedure yielding interior and exterior camera orientation as well as a sparse point cloud. The


2
9
5
query_1
Images in database
Selection of 1-3 query images of one landmark
Content-based image retrieval for landmark using LEA
ID, dist to q1:

2, 1.4511
9, 2.5223
5, 7.9157
....
Intersection of result lists and storage of best results
Feature matching (SuperGlue)
Feature description (SuperPoint +SuperGlue)
Feature detection (SuperPoint)
12
27
9
18
91
7
19
55
31
1
5
Initial camera para-meterization
f estimated with VPD; cx, cy lie in the image center; k=0
Import of images, cameras, and matches into COLMAP
Geometric verification using LO-RANSAC
Bundle adjustment
Allow two-view tracks
Camera poses (exterior orientation)
Interior orientation
Triangulated 3D points (sparse cloud)
Determination of homologue points in local model and world model
Geo-referencing using Helmert-Transformation
Final reference model (images, calibrated cameras, retrieval feature files)

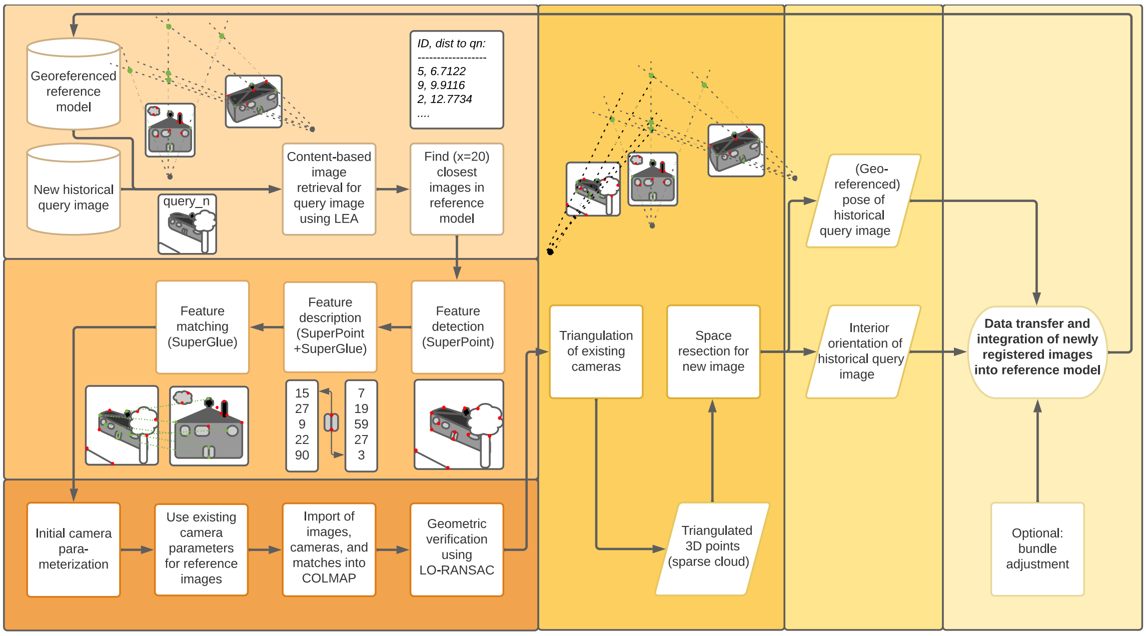
```
Registering image #19 (166)
==========================
  => Image sees 156 / 496 points


Pose refinement report
-------------------------------
    Residuals : 130
   Parameters : 8
   Iterations : 15
         Time : 0.00138854 [s]
 Initial cost : 0.932775 [px]
   Final cost : 0.879832 [px]
  Termination : Convergence
```

Figure (7.1): The different steps from unordered historical images in a database up to the generation of a georeferenced reference model.

following Helmert-Transformation requires point pairs between local and global coordinate system which have to be determined in a manual process. Finally, this allows the georeferencing of a 3D reference model and camera poses using exclusively historical images which can then be used for the estimation of interior and exterior camera parameters of further images.

The second workflow describes the process of adding historical images to the existing initial reference reconstruction (Fig. 7.2).


Georeferenced reference model
New historical query image
query_n
Content-based image retrieval for query image using LEA
ID, dist to qn:

5, 6.7122
9, 9.9116
2, 12.7734
....
Find (x=20) closest images in reference model
Feature matching (SuperGlue)
Feature description (SuperPoint +SuperGlue)
Feature detection (SuperPoint)
15
27
9
22
90
7
19
59
27
3
Initial camera para-meterization
Use existing camera parameters for reference images
Import of images, cameras, and matches into COLMAP
Geometric verification using LO-RANSAC
Triangulation of existing cameras
Space resection for new image
Triangulated 3D points (sparse cloud)
(Geo-referenced) pose of historical query image
Interior orientation of historical query image
Data transfer and integration of newly registered images into reference model
Optional: bundle adjustment

Figure (7.2): The different steps for pose estimation of an additional historical image using an a-priori calculated reference model.

Suppose a new image of the same landmark is uploaded to the database for which the pose is not estimated yet. Additionally, the reference reconstruction generated by the first workflow (Fig. 7.1) is imported. Then, LEA with a kernel size 4 × 4 with stride 3 and a reference order 4 is used again to find the 20 most similar images in the existing reconstruction. Therefore, the pre-calculated LEA features are stored in a database and can be imported to reduce computation time. For the new (query) image and all reference images, SuperGlue feature matching is performed and the matches are again geometrically verified.

All values are imported into COLMAP and the *existing* camera parameters are used for the triangulation of image feature points. The triangulated points are used to determine the exterior and interior orientation of the query image using space resection. All parameters are written into the reference model structure and can be used for iterative pose estimation of further query images. This workflow allows the integration of images which are not directly showing the landmark but side streets or only building parts. If images are integrated in edge areas of the reference model, a subsequent bundle adjustment should be performed in order to reduce errors in the model (Snavely et al., 2008). The second workflow does not involve user interaction anymore.

7.1.1 Error assessment

The pipelines depicted in Figure 7.1 and 7.2 propose the completely automatic estimation of interior and exterior camera parameters of historical images. In practice, the workflow will have its limitations depending on the data quality and heterogeneity. These limitations are to be elaborated here.

Considering the CBIR, it is shown that the user has to declare three query images of the landmark to retrieve relevant and meaningful results for a reference reconstruction. While this has mostly practical reasons as all images are gathered from the repository anyways, an automatic method for selecting query images would be preferable. An idea could be an initial clustering of photographs in order to select images that have a high correlation to other images. This method is called query expansion (Chum et al., 2007; Chum et al., 2011) and could be implemented in the future.

This would probably also improve the method's reliability on very heterogeneous datasets. It could be observed that the image retrieval sometimes yields non-landmark images due to so-called confusers (Chum et al., 2011), i.e. parts of the query image that do not show the landmark, such as the sky or the ground. Another effect especially relevant for the Deutsche Fotothek, are drawings that are a result of the CBIR but cannot be used directly in the SfM workflow as depicted in Figure 7.3.

These photographs which are usually a low percentage of all images can mostly be eliminated in the SuperGlue matching and geometric verification.

The CBIR works especially reliable if there is a high number of similar images available and if images in the metadata category can be separated well. Sometimes, the retrieval yields images of several sides of the depicted landmark. While this is not an error in the CBIR, this can lead to severe issues in the following step of feature matching.

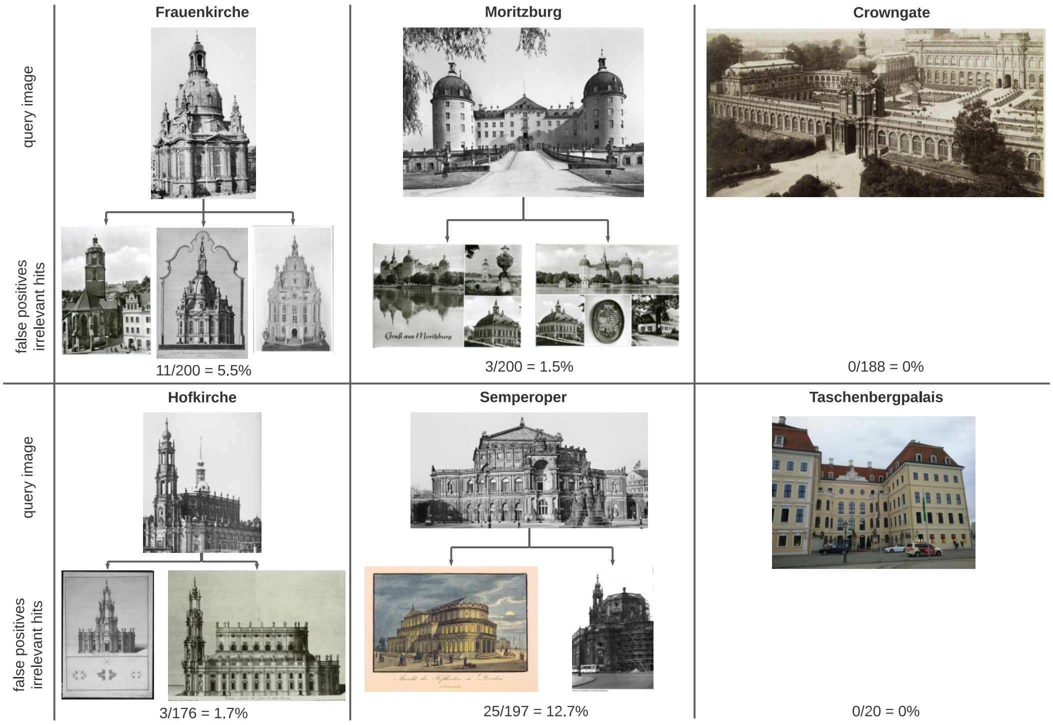
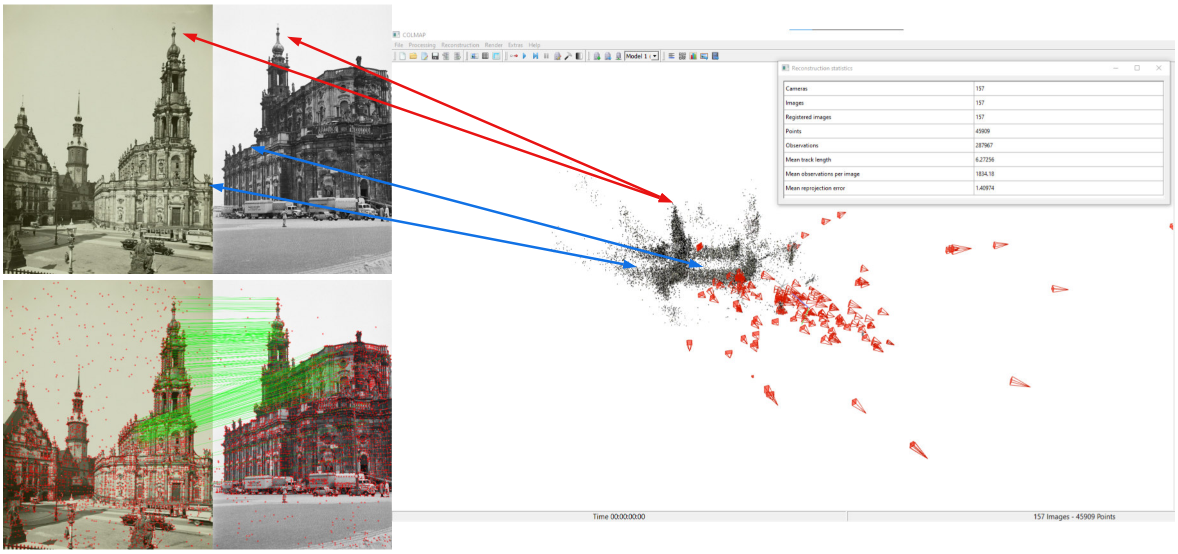


Figure (7.3): False positives or irrelevant hits for six different query images (excerpt) showing buildings in the vicinity of Dresden. For all datasets around 200 images are retrieved except for the Taschenbergpalais where only few images are available. For the Crowngate of the Dresden Zwinger and the Taschenbergpalais all results yielded by LEA show the specific object, while for e.g., the Frauenkirche, Hofkirche, Moritzburg or the Semperoper drawings and postcards are yielded to a small extent.



Reconstruction statistics

Cameras	157
Images	157
Registered images	157
Points	45909
Observations	287967
Mean track length	6.27256
Mean observations per image	1834.18
Mean reprojection error	1.40974

Time 00:00:00:00

157 Images - 45909 Points

Figure (7.4): An example for an erroneous SuperGlue matching. For the historical images of the Hofkirche, Dresden this happened between multiple image pairs resulting in gross errors in the final reconstruction. All reconstructed camera poses point from the same perspective to the building which is clearly incorrect. The tower (red arrow) and the balustrades (blue arrows) have been reconstructed at least twice but merged in a single model.

Symmetric building parts, feature-less regions, or repetitive structures still pose a problem for feature matching. Even for SuperGlue with following geometric verification gross errors occur rarely which falsify the final result of pose estimation to a large degree (Fig. 7.4).

This is partly compensated by COLMAP creating multiple models with fewer images of the same building but this can not always be guaranteed. As this happens especially when processing a large amount of images at once, this error can be probably bypassed by using the sequential workflows with fewer images depicted in Figure 7.1 and 7.2.

COLMAPs bundle adjustment is optimized for the reconstruction of architectural scenes and always converges for the urban photographs of Dresden. For other datasets (especially aerial images of Figure 7.9) the bundle adjustment does not converge and an initial reconstruction is not possible. This can be avoided by parameter tuning of the local bundle adjustment. The final reconstruction results are evaluated by absolute measures and visual errors in the following.

7.1.2 Accuracy estimation

In the photogrammetric context of this thesis, an estimation of the resulting accuracy is of importance. For the SfM pipeline, statements can be made about the quality and potential accuracy of the different steps.

In a first instance, the whole feature matching process is difficult to judge in terms of accuracy. While the Image Matching Challenge provides insights on the quality of feature matching methods on various image datasets, it could be observed that the quality drastically changes when using historical data. This raises the need for the creation of a historical dataset and requires the time-consuming interactive selection of homologue point pairs (point triples) in multiple images as shown in Chapter 2 and 5. With the provided Trifocal Tensor (or Fundamental Matrices) the matching score (or matching ratio) can be determined. The matching score is a common measure for evaluating feature matching methods (Mikolajczyk et al., 2005; Jin et al., 2020) and is calculated by $\frac{\text{correct matches (inliers)}}{\text{all matches}}$ (Appendix 8.1). It is desired to have a matching score of 100% and a preferably high number of correct matches. The results of a large number of different tested methods are already shown in the introduction in Figure 1.10.

The matches and initial interior orientation parameters are used in COLMAP's bundle adjustment. Unfortunately, COLMAP provides only few resulting values after bundle adjustment with its most significant parameter reprojection error which is described in Section 5 in detail. It corresponds to the reprojected object coordinates from matched 2D image correspondences after bundle adjustment.

Further local accuracy measures are the commonly used angle error and pose error (Sattler et al., 2018; Jin et al., 2020) also described in Section 5. Both can only be calculated if a reference model with superior accuracy exists. These error measures can provide an estimation about the local accuracy of the model and are also used to compare different feature matching methods. Though, the impact on the absolute values of the three-dimensional (3D) model's coordinates can only be quantified.

This fact leads to two different investigations of the accuracy and quality assessment of the workflows of Figure 7.1 and 7.2. The first approach aims for a visual control of the final result in a Web3D application which is of importance especially for the users.

The investigation is performed on the Taschenbergpalais, Dresden because of an already existing georeferenced level of detail (LOD)3 3D model which is probably more commonly available in municipalities than a SfM model. Homologue point pairs are selected in the local SfM reconstruction in COLMAP created by workflow 7.1 and in the georeferenced LOD3 model in COLMAP. This step can be error-prone as similar points have to be found, precisely selected and correctly matched by hand. Additionally, a LOD3 model still shows generalized structures and may be inaccurate on certain building parts. In order to minimize these errors, the Helmert transformation is performed using the least-squares approach of (Umeyama, 1991) including outlier detection with Random Sample Consensus (RANSAC) (Appendix 8.2). The final protocol of the transformation still shows errors in the range of 0.5 up to 3.5 meters ($\sigma_0 = 1.12$ m). These can also be seen in the depiction of the historical image as an overlay on the georeferenced model (Fig. 7.5).

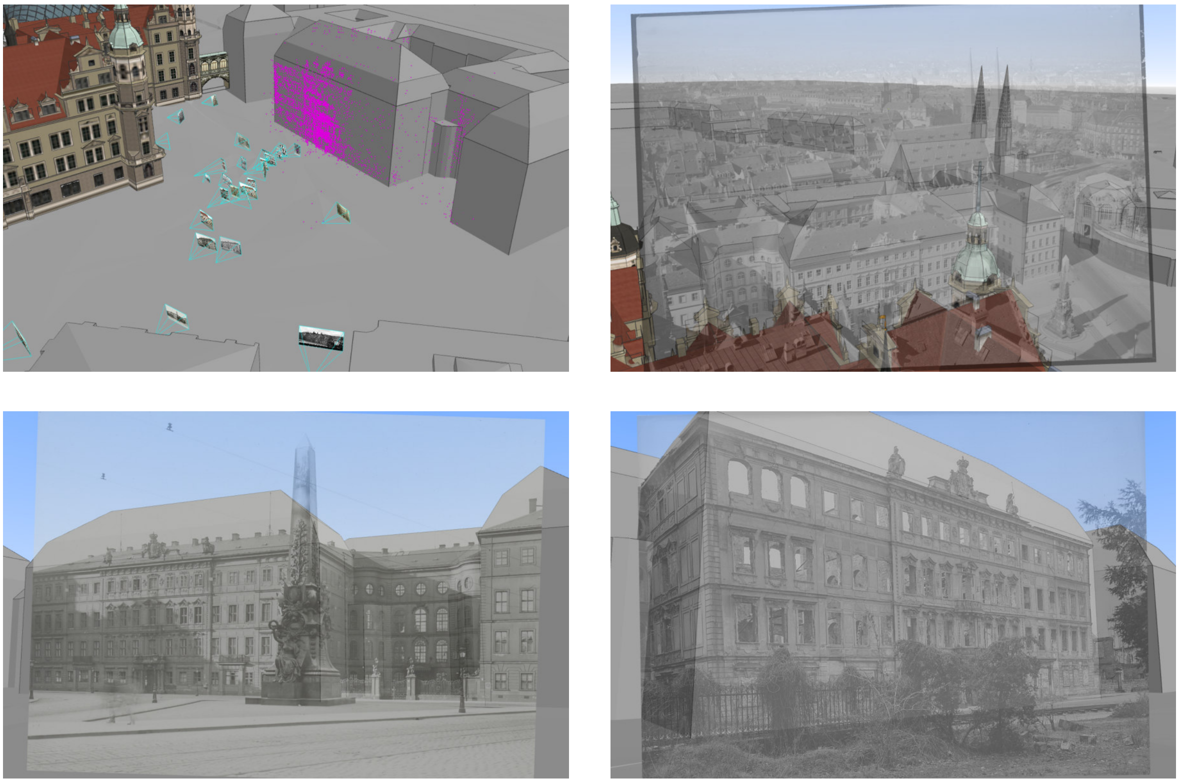


Figure (7.5): The top-left image shows a georeferenced SfM model using exclusively historical images of the Taschenbergpalais. The other semi-transparent images depict the perspective view looking from the camera center along the optical axis in direction of the 3D model.

Larger deviations can be seen particularly on the eaves of the building. This is possibly due to discrepancies between the actual building height and the height of the LOD3 model. Consequently, it is assumed that the manual selection of points in a LOD3 model may be the limiting factor for a more accurate georeferencing. To further evaluate this assumption, a model generated by terrestrial laserscanning could be used as reference in future research.

Nonetheless, this initial georeferenced model of the Taschenbergpalais is used for the evaluation of workflow 7.2. In this case that means, that the already calculated poses and interior orientation parameters of this initial model of the Taschenbergpalais are fixed. Then, feature matches are calculated between a new image and all existing images and the points are triangulated using the given poses. Subsequently, the pose and interior camera parameters of the new image is estimated via space resection. The final result of a completely automatic newly registered image propagates as expected similar inaccuracies (Fig. 7.6).

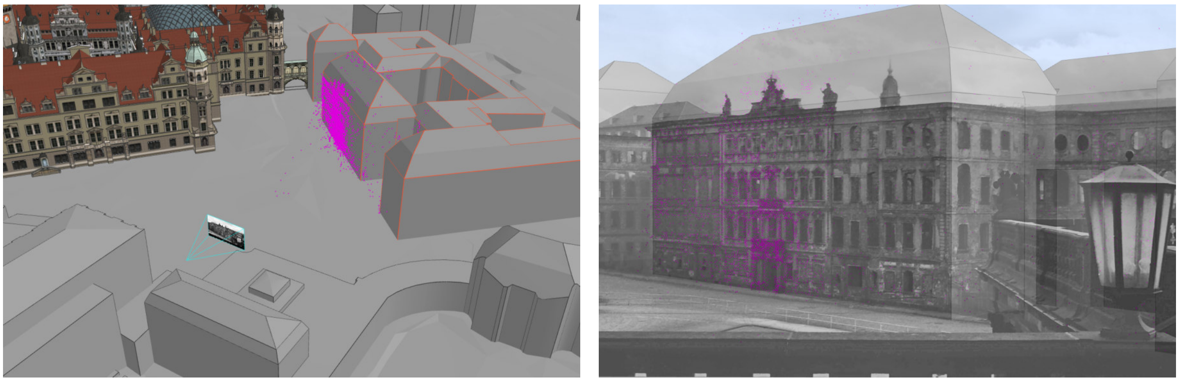


Figure (7.6): Newly georeferenced historical image using the formerly created reference model.

The second approach on accuracy estimation uses measurements on the Georg-Schumann-Bau, TU Dresden close to the Institute of Photogrammetry and Remote Sensing which made an a priori and a posteriori camera calibration possible. This allows the generation of a contemporary reference model created by the SfM pipeline. Historical images are mainly provided by the Münchner Platz Dresden Memorial while few are directly downloaded from the Deutsche Fotothek.

The contemporary model including the measurements with superior accuracy are processed simultaneously with the historical images generating the reference reconstruction with interior and exterior camera parameters. These are used for a direct comparison with the historical model generated by using workflow 7.1 out of exclusively historical images.

Initial results show that the absolute accuracy of the camera poses is still in the range of few meters matching with the visual check depicted in Figure 7.5 and 7.6. Especially the coordinate perpendicular to the building shows large deviations to its real value. This is mainly due to its direct correlation to the principal distance which can not always be estimated correctly.

These and further results of the accuracy analysis using additional measurements will be published in the Conference Proceedings of the ISPRS Congress 2022 in Nizza, France.

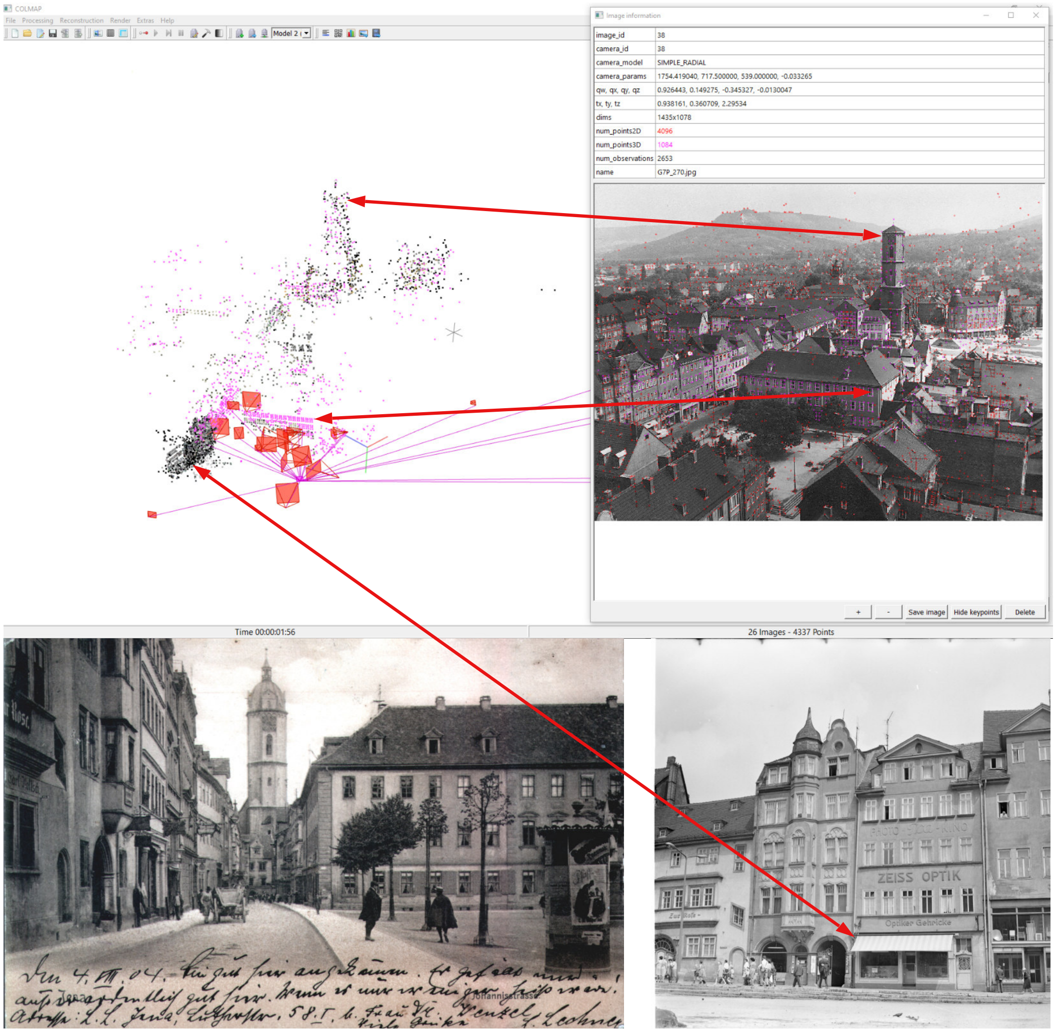
The accuracy analysis shows that there is still potential for improvements especially regarding the absolute accuracy of the camera poses. Nonetheless, the estimated poses are a valuable able tool for texturing 3D models and visualization of camera orientations and pose clusters in different points of time.

7.1.3 Transfer of the workflow

The SfM for historical images was mainly developed for the purpose of the reconstruction of urban architectural scenes in the vicinity of Dresden. The presented work up to this point shows successful reconstructions of Schloss Moritzburg, the Crowngate of the Zwinger, the Semperoper, parts of the Hofkirche, the Taschenbergpalais and the Georg-Schumann-Bau.

It is of high interest, if the approach is also functional for other cities with buildings from different epochs, for landscape scenes and for other data formats such as aerial images. In the following, multiple reconstruction results with several example images are briefly shown.

The Thüringer Universitäts- und Landesbibliothek Jena (ThULB) provided 82 historical images of Jena showing different locations in the city. Due to the sparse sample, the images are filtered manually resulting in 38 images showing the church St. Michael and the surrounding area with the Eichplatz. The dataset is extremely challenging containing few views in completely opposite directions including several postcards and image pairs show large radiometric differences. The pose of 26 out of 38 images could be reconstructed (Fig. 7.7). Images which could not be oriented show postcards and façade details.



image_id	38
camera_id	38
camera_model	SIMPLE_RADIAL
camera_params	1754.419040, 717.500000, 539.000000, -0.033265
qw, qx, qy, qz	0.926443, 0.149275, -0.345327, -0.0130047
tx, ty, tz	0.938161, 0.360709, 2.29534
dims	1435x1078
num_points2D	4096
num_points3D	1084
num_observations	2653
name	G7P_270.jpg

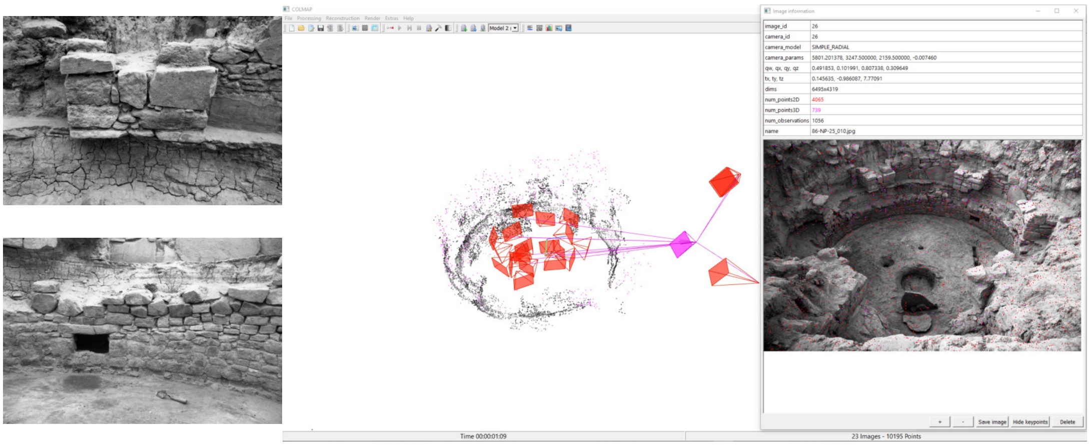
Time 00:00:01:56

26 Images - 4337 Points

Figure (7.7): Workflow 7.1 applied on 38 images of the historical center of Jena, Germany. It generates a successful and correct reconstruction of 26 camera poses.

Another 31 historical images from 1986 of are kindly provided by Robert Bischoff. These show the archaeological documentation of a kiva (ceremonial space of Ancestral Pueblo) from the Nancy Patterson Village site in southeastern Utah. The photographs are of high quality and taken with a single camera whose type is unknown. The images are most likely digitized from film. They show an overview from opposing sites and also detailed views of the kiva.

SuperGlue is able to deal with the extreme scale and perspective changes (Fig. 7.8). Hence, it is not possible to connect the images taken from opposing sites as SuperGlue is not able to deal with rotations larger than 45°between image pairs.



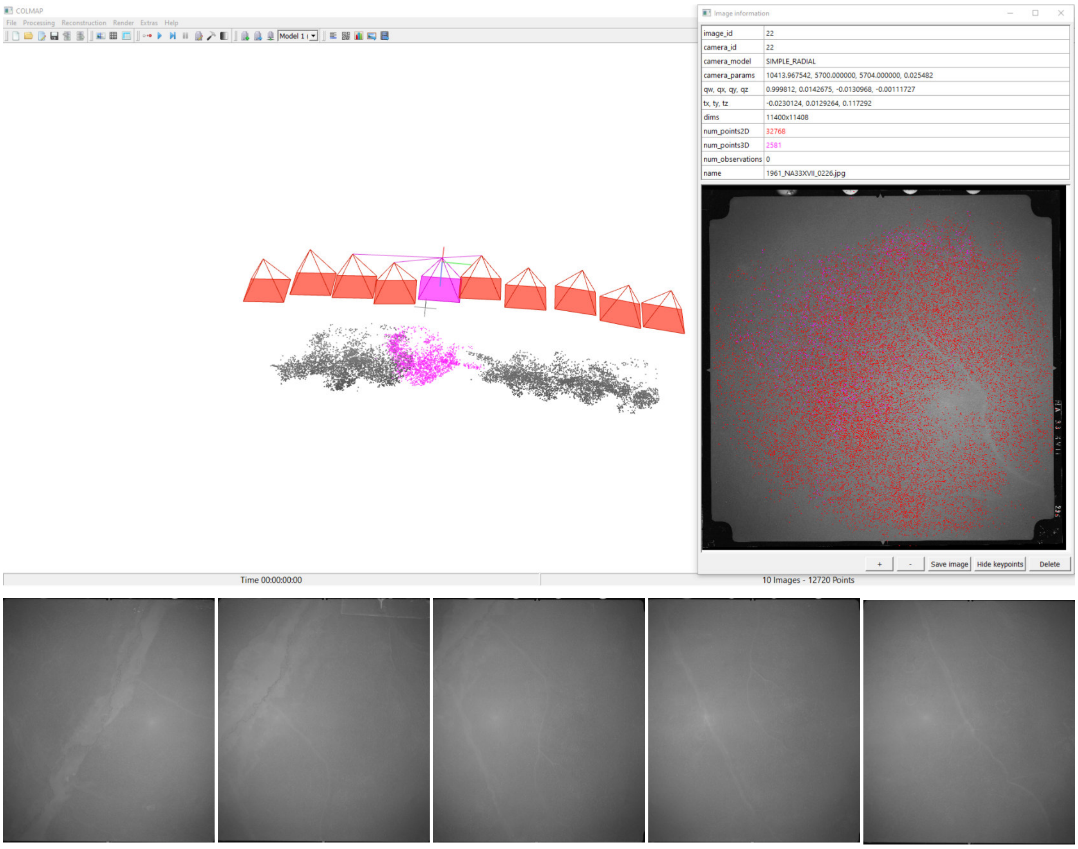
image_id	26
camera_id	26
camera_model	SIMPLE_RADIAL
camera_params	5801.201378, 3247.500000, 2159.500000, -0.007460
qw, qx, qy, qz	0.491853, 0.101991, 0.807338, 0.309649
tx, ty, tz	0.145635, -0.986087, 7.77091
dims	6495x4319
num_points2D	4065
num_points3D	739
num_observations	1056
name	86-NP-25_010.jpg

Time 00:00:01:09

23 Images - 10195 Points

Figure (7.8): Workflow 7.1 applied on 31 images of a kiva in Utah. It generates a successful and correct reconstruction of 23 camera poses.

The SfM workflow is also transferred to aerial images. Recently, the use of SuperGlue on historical aerial images has been published by Zhang et al. (2021).



image_id	22
camera_id	22
camera_model	SIMPLE_RADIAL
camera_params	10413.967542, 5700.000000, 5704.000000, 0.025482
qw, qx, qy, qz	0.999812, 0.0142675, -0.0130968, -0.00111727
tx, ty, tz	-0.0230124, 0.0129264, 0.117292
dims	11400x11408
num_points2D	32768
num_points3D	2581
num_observations	0
name	1961_NA33XVII_0226.jpg

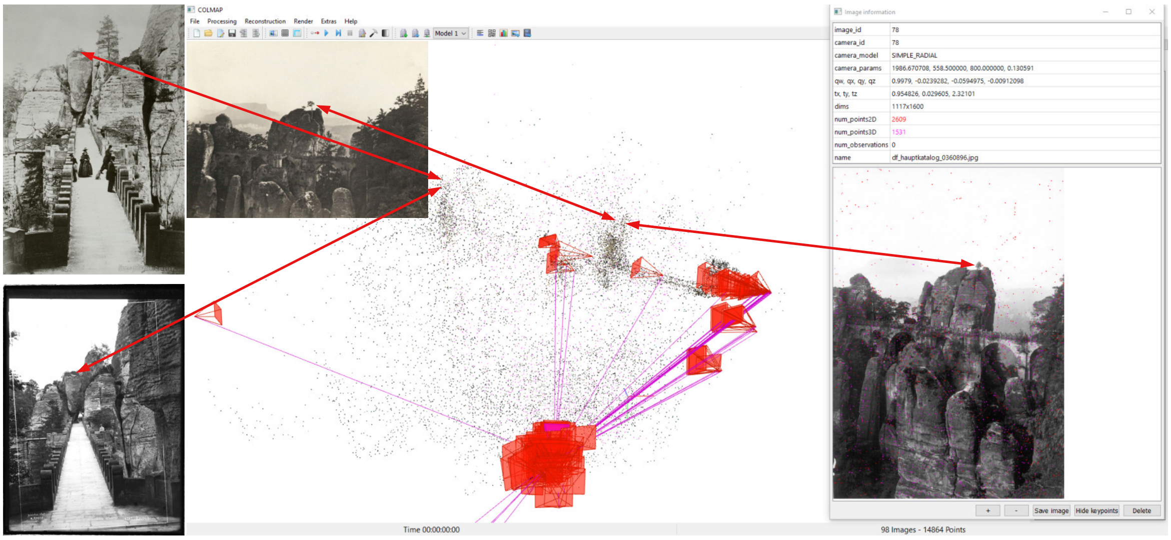
Time 00:00:00:00

10 Images - 12720 Points

Figure (7.9): Workflow 7.1 applied on 27 aerial images of the rain forest in Congo, Africa. All camera poses could be reconstructed in three different models respectively for every flight strip.

However in the depicted case, a more accurate model could be computed by using Discrete Keypoints (DISK) and changing several parameters of COLMAP's bundle adjustment. This includes e.g., lowering the initial minimum number of inlier matches and allowing two-view tracks (cf. Section 4.4.4). The historical images from 1961 showing the rain forest in Congo, Africa, test site of Lim et al. (2022) were provided by Denis Feurer and it is possible to generate feature matches for image pairs showing close to zero visible features even for a human. For each of the three flight strips a separate model is created. One flight strip is shown in Figure 7.9.

The last example shows landscape images that were acquired from the Deutsche Fotothek. For the Bastei in the Elbe Sandstone Mountains 169 photographs, postcards and maps are retrieved by using LEA on 1519 hits for the search query "bastei". The final reconstruction shows the camera pose of 98 of these images (Fig. 7.10). Images which could not be reconstructed are maps, postcards, drawing and stereo images.



image_id	78
camera_id	78
camera_model	SIMPLE_RADIAL
camera_params	1986.670708, 558.500000, 800.000000, 0.130591
qw, qx, qy, qz	0.9979, -0.0239282, -0.0594975, -0.00912098
tx, ty, tz	0.954826, 0.029605, 2.32101
dims	1117x1600
num_points2D	2609
num_points3D	1531
num_observations	0
name	df_hauptkatalog_0360896.jpg

Time 00:00:00:00

98 Images - 14864 Points

Figure (7.10): Workflow 7.1 applied on 169 images of the rock formation Bastei. 98 camera poses could be reconstructed in one model.

The experiments performed on different datasets show several findings. It is not yet possible to estimate the pose of all images yielded by LEA due to the reasons already explained above. Furthermore, this happens also because of the properties of SuperGlue.

SuperGlue is not completely rotation-invariant as already shown by Zhang et al. (2021) and in the above-mentioned examples. It is able to match image pairs up to an rotation of 45°which affects the reconstruction of Jena and the kiva. Additionally, the flight strips could not be connected at the points where the airplane made a turn. SuperGlue is trained on the MegaDepth dataset (Li and Snavely, 2018) and it is therefor recommend to use images with a maximum resolution of 2000x1500 pixels. Aerial images are often of higher resolution and must be divided and merged again (Zhang et al., 2021). In some cases, depending on the dataset, DISK may be a better choice for feature detection and matching. When the images are upright as in most archives and repositories, workflow 7.1 is able to deal with all kinds of images and may even be successful for interior views.

7.2 Developments and Outlook

At the starting point of this thesis, the generation of sparse point clouds using historical images was only possible for a very low percentage of all available images. The pose estimation for the cameras failed mainly due to finding correct correspondences between historical image pairs. Only through the enormous improvement by using neural networks and especially the Superpoint+SuperGlue process for finding homologue image points, this was made possible. Considering the recent advantages presented at the Image Matching Challenge 2021 of the CVPR, it is not conceivable that a different method will outperform SuperGlue in the near future like SuperGlue outperformed Scale Invariant Feature Transform (SIFT) for all kinds of images as shown in previous Section 7.1.3. Mainly, different methods are combined with each other (for instance SuperPoint+DISK) and produce slightly better results than the initial SuperPoint+SuperGlue approach.

Nonetheless, there is ongoing research on innovative neural network structures like Transformers which perform especially well on scenes with low texture (Sun et al., 2021). In first experiments done for this thesis, the above-mentioned method Local Feature Transformer (LoFTR) could not outperform SuperGlue. Others improve the final number of correct matching points by using advanced RANSAC methods for geometric verification as already mentioned in Section 1.3.3.4. All learned feature matching methods use contemporary datasets for training. Mainly, the MegaDepth dataset (Li and Snavely, 2018) is used which consists of 196 SfM models with their corresponding images and depth maps. As it is now possible to generate historical SfM models, it would be interesting to enrich the training process with historical models and see if this has an impact on the final results and quality of the feature matching.

The approaches presented in this thesis primarily deal with images of specific landmarks mainly due to data availability. Next steps could include historical photographs of outbuildings and side streets using an existing reference model with the workflow shown in Fig. 7.2.

As it has been shown that contemporary images can be used for estimation of interior and exterior camera parameters of historical photographs (Section 7.1.2), another approach could be the use of already existing camera poses for contemporary images. These could be provided e.g., by Google Street View `https://www.google.de/intl/de/streetview/` or Mapillary `https://www.mapillary.com/`. However, this has at the moment not been done for a larger amount of images.

Correctly estimating the camera pose of historical images, opens multiple new research topics and applications. An obvious application is the transfer of image content onto 3D city models also called texture projection and is already done in the Jena4D project (`https://4dcity.org/?scene=jena`) using the described methods. It is not solved yet how to merge co-existing textures on a 3D model taken in the same time but showing large radiometric differences.

As mainly contemporary 3D models exist at the moment, the texture projection may fail or show image artifacts. This raises a need for the creation of *historical* 3D models where the texture can be projected flawlessly. There is ongoing research on the automatic creation of 3D models by automatically vectorizing historical maps (which date back much further than historical photographs) available in digitized raster graphics. However, due to the strongly varying appearance of historical maps a general approach is not yet developed (Herold and Hecht, 2018; Heitzler and Hurni, 2020; Chen et al., 2021). Therefore, the ICDAR2021 Competition on Historical Map Segmentation was introduced in order to automate the process of

1. detecting building blocks
2. segmenting the map content area
3. locating graticule lines intersections

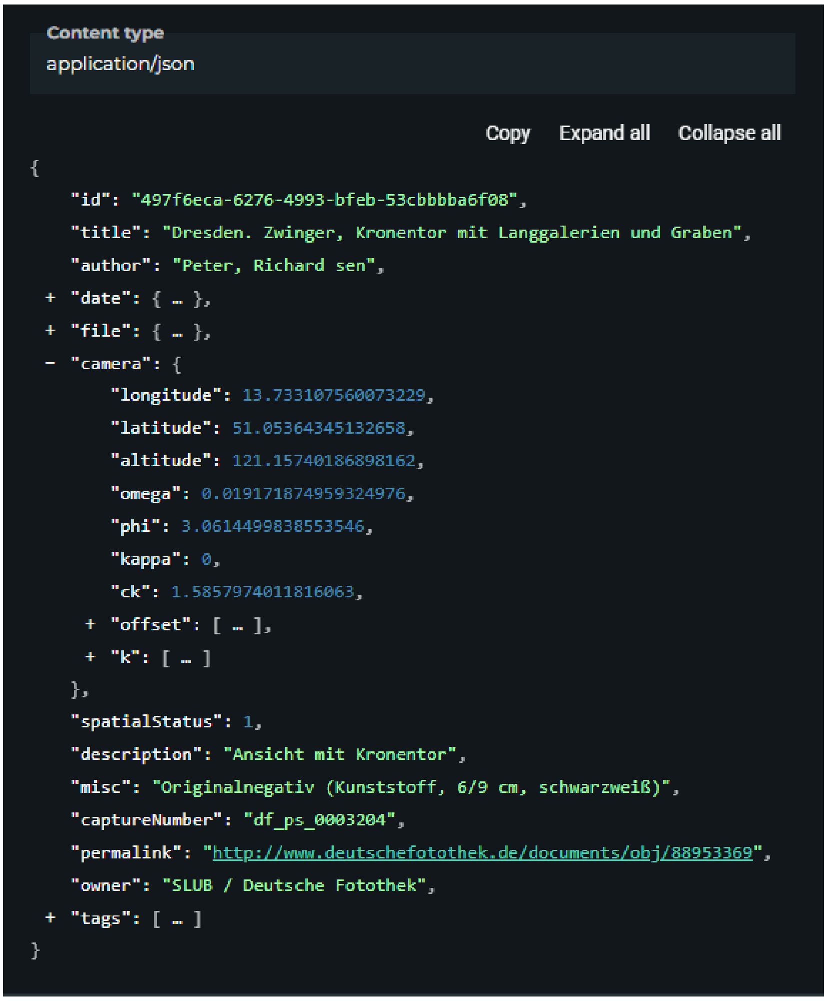
Especially, the automatic detection of building blocks that is important for extruding 3D models (wall elements), could not yet be solved generally (Chazalon et al., 2021).

The combination of georeferenced historical images and historical building models would allow the improvement of e.g. roof structures and the (automatic) assessment of former building heights (Farella et al., 2021). A further point could be the inclusion of additional material such as accurate perspective paintings, drawings, or plans. These are usually yielded by the CBIR, however are not able to be directly processed in the SfM workflow. At the moment, this is mainly done for certain objects by manual processing and regarding geometric constraints (Canciani et al., 2018; Kourniatis and Architect, 2019; Ansaldi, 2020).

Further, the experiments done with the CBIR approach LEA significantly improve the conventional metadata search which often yields not the expected results or neglected valuable images if the filters are set too narrow. The thesis uncovered in multiple experiments that the metadata quality is often not sufficient for filtering and retrieving all relevant images. While metadata and CBIR seem like contrary approaches because these are often opposed in the frame of this thesis, the best strategy seems to be a combination and alignment of both methods:

- The CBIR is able to control the assigned metadata for every object and maybe even improve it.
- The subsequent SfM pipeline can automatically tag images with spatial information (e.g. western view, top of the church, south of building XY).
- Metadata can be used to form clusters before CBIR significantly reducing computation time and reducing false positives as already demonstrated in Chapter 5.

At the moment, libraries and archives mainly work with metadata or sometimes geographical databases (`geonames.org`) in order to describe spatial information. It is conceivable that the depicted pose estimation workflow could be integrated into a library to directly store the interior and exterior orientation of the camera in combination with the metadata of the image. This is investigated in a first prototype (Fig. 7.11).



Content type
application/json

```
{
   "id": "497f6eca-6276-4993-bfeb-53cbbbba6f08",
   "title": "Dresden. Zwinger, Kronentor mit Langgalerien und Graben",
   "author": "Peter, Richard sen",
 + "date": { … },
 + "file": { … },
 − "camera": {
       "longitude": 13.733107560073229,
       "latitude": 51.05364345132658,
       "altitude": 121.15740186898162,
       "omega": 0.019171874959324976,
       "phi": 3.0614499838553546,
       "kappa": 0,
       "ck": 1.5857974011816063,
     + "offset": [ … ],
     + "k": [ … ]
   },
   "spatialStatus": 1,
   "description": "Ansicht mit Kronentor",
   "misc": "Originalnegativ (Kunststoff, 6/9 cm, schwarzweiß)",
   "captureNumber": "df_ps_0003204",
   "permalink": "http://www.deutschefotothek.de/documents/obj/88953369",
   "owner": "SLUB / Deutsche Fotothek",
 + "tags": [ … ]
}
```

Figure (7.11): Example for the JSON request for the historical image with captureNumber *df_ps_0003204*. For the image the camera pose is estimated (spatialStatus=1) and all exterior and interior camera parameters can be accessed. Image taken from `https://4dbrowser.urbanhistory4d.org/api/`.

Concluding, this thesis could prove the initial statement formulated by Ware (2021) and made in the introduction about the advantages of proper data visualization. Historical photographs directly linked with camera pose data enhance the raw 2D image by a spatial component valuable for e.g., architectural historians, tourists, or citizens. It enables a direct link to 3D applications drawing user attention and improving visual communication. The visualization of camera poses of historical images (Fig. 3.8) is able to give a new understanding of the distribution of the viewpoint of the photographers as well as the main object of interest (Bruschke et al., 2018b). CBIR and pose estimation of historical images can reveal errors in metadata while also being able to identify further unsolved issues as e.g., dealing with postcards and drawings. Due to these advantages, it is only a matter of time that the camera pose and interior camera parameters of (historical) images will be integrated in large repositories.

8 Appendix

This Appendix holds information about the final feature evaluation on the benchmark dataset created in Chapter 2. Additionally, the transformation from the local coordinate system in COLMAP using the OpenCV camera definition into the global four-dimensional (4D) browser coordinate system using the OpenGL camera definition is presented.

8.1 Setup for the feature matching evaluation

When the dataset was published in 2019, some of the methods used, were not published yet which is the main reason for a re-evaluation in the introduction of this thesis. The results of all methods are not further geometrically verified (Section 1.3.3.4) to allow a fair comparison.

As SuperPoint+SuperGlue is identified as the most robust variant in prior investigations the test parameters are oriented according to that method. That means, that all image triples derived from `http://dx.doi.org/10.25532/OPARA-24` are initially matched with SuperGlue in order to check the visual quality. For all images the resolution is halved and the images are converted to .jpg file format to allow the use of all methods. Then, the feature matches are visualized in OpenCV and SuperGlue is able to match all image pair combinations.

A qualitative analysis and comparison requires using a fixed setup for testing all methods. The given Trifocal Tensor for the image triples allows the extraction of the Fundamental Matrices relating all possible image pairs. Testing feature matches with given Fundamental Matrices can be done using the so-called epipolar threshold method. All feature matches are transformed into homogeneous coordinates x, x' which allows the direct solving of the Fundamental Matrix equation:

$$x'^T F x < \text{epipolar threshold} \tag{8.1}$$

If the equation yields a result lower than a fixed epipolar threshold the match is considered valid. For the experiments the epipolar threshold is set to 0.1 considering the low number of manually selected points of the ground truth dataset. Further the ground truth does only provide matches with pixel accuracy which could be extended in the future. Using this setup reveals that the accuracy and quality of the last image triple is not good enough for the comparison (i.e., all visual correct matches are invalid for the epipolar threshold method) so it is excluded for testing.

Nonetheless, this setup allows testing a lot of different methods in a qualitative comparison for 21 challenging image pairs. If possible, the number of detected feature points is limited to 4096. Apart from that, all methods are used in their standard implementation in OpenCV

(SIFT, MSER, ORB, AKAZE), in Matlab (RIFT) as Windows executable (MODS), and in Python (D2-Net-ss, D2-Net-ms, R2D2, SuperGlue, DISK). Exceptions are raising the number of detected points (minMatches parameter) in MODS after recommendation of the author and the matching threshold of D2-Net according to Chapter 4.

For all methods the absolute number of correct matches and the matching ratio $= \frac{\text{correct matches}}{\text{all matches}}$ is calculated over all possible image pairs, averaged, and depicted in Figure 1.10.

8.2 Transformation from COLMAP coordinate system to OpenGL

COLMAP exports the cameras' poses as projection from world to camera coordinate system as quaternions defined using Hamilton convention $R(q_w, q_x, q_y, q_z)$ plus the translation vector $T(X, Y, Z)$. To derive the projection center, the quaternions have to be normalized and multiplied with the negative translation vector 8.2.

$$\text{projection center}(X, Y, Z) = R(\hat{q_w}, \hat{q_x}, \hat{q_y}, \hat{q_z}) * -T(X, Y, Z) \tag{8.2}$$

Being able to transform the rotational component of the camera's pose, the rotation matrix $R(\omega, \phi, \kappa)$ has to be carefully derived from the original quaternions given in the Hamilton convention (equation 8.3).

$$R(\omega, \phi, \kappa) = \begin{bmatrix} 1 - 2q_y q_y - 2q_z q_z & 2q_x q_y - 2q_z q_w & 2q_x q_z + 2q_y q_w \\ 2q_x q_y + 2q_z q_w & 1 - 2q_x q_x - 2q_z q_z & 2q_y q_z - 2q_x q_w \\ 2q_x q_z - 2q_y q_w & 2q_y q_z + 2q_x q_w & 1 - 2q_x q_x \end{bmatrix} \tag{8.3}$$

The resulting rotation matrix has to be further processed. First, the second and third row of the matrix have to be negated to follow the definition of the coordinate system in OpenGL. Second, the resulting rotation matrix has to be transposed to matching OpenGLs order of elements in the rotation matrix (column-major in opposite to OpenCVs row-major definition) as shown in equation 8.4.

$$R(\omega, \phi, \kappa)_{OpenGL} = (R(\omega, \phi, \kappa)_{OpenCV} * \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix})^T \tag{8.4}$$

The local camera pose is now determined by the homogeneous synthesis of the rotation and translation component.

The transformation matrix can be derived from point correspondences in both coordinate systems. Least-squares estimation is used to find a transformation matrix between these correspondences (Umeyama, 1991).

To derive the global camera poses, the local pose is simply multiplied with the transformation matrix (equation 8.5).

$$\begin{bmatrix} R_{global} & & & T_{global} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} R_{transformation} & & & T_{transformation} \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} R_{local} & & & T_{local} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8.5}$$

References

Abdel-Hakim, A. and A. Farag (2006). “CSIFT: A SIFT Descriptor with Color Invariant Characteristics”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR’06).* IEEE, pp. 1978–1983. DOI: 10.1109/cvpr.2006.95.

Agapito, L. de, E. Hayman and I. Reid (1998). “Self-Calibration of a Rotating Camera with Varying Intrinsic Parameters”. In: *Procedings of the British Machine Vision Conference 1998.* British Machine Vision Association, pp. 105–114. DOI: 10.5244/c.12.11.

Agarwal, S., Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz and R. Szeliski (2011). “Building Rome in a day”. In: *Communications of the ACM* 54, 10, pp. 105–112. DOI: 10.1145/2001269.2001293.

Agarwal, S., K. Mierle and T. C. S. Team (2022). *Ceres Solver.* http://ceres-solver.org.

Albertz, J. (2009). “100 Jahre Deutsche Gesellschaft fur Photogrammetrie, Fernerkundung und Geoinformation e.V.” In: *Photogrammetrie - Fernerkundung - Geoinformation* 2009, 6, pp. 487–560. DOI: 10.1127/1432-8364/2009/0035.

Albl, C., A. Sugimoto and T. Pajdla (2016). “Degeneracies in Rolling Shutter SfM”. In: *Computer Vision – ECCV 2016.* Springer International Publishing, pp. 36–51. DOI: 10.1007/978-3-319-46454-1_3.

Alcantarilla, P., J. Nuevo and A. Bartoli (2013). “Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces”. In: *Procedings of the British Machine Vision Conference 2013.* British Machine Vision Association, pp. 1–11. DOI: 10.5244/c.27.13.

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

AliceVision (2018). *Meshroom: A 3D reconstruction software.* URL: https://github.com/alicevision/meshroom.

Ansaldi, B. (2020). “Touching and ‘Feeling’ the Feast of Herod by Benozzo Gozzoli: A Multisensory Communication Strategy Unveiling the Secrets of Painted Spaces to the Blind”. In: *Advances in Intelligent Systems and Computing.* Springer International Publishing, pp. 446–457. DOI: 10.1007/978-3-030-63403-2_40.

Arandjelovic, R. and A. Zisserman (2012). “Three things everyone should know to improve object retrieval”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 2911–2918. DOI: 10.1109/cvpr.2012.6248018.

Arandjelovic, R., P. Gronat, A. Torii, T. Pajdla and J. Sivic (2016). “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5297–5307. DOI: `10.1109/cvpr.2016.572`.

Balntas, V., K. Lenc, A. Vedaldi and K. Mikolajczyk (2017). “HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3852–3861. DOI: `10.1109/cvpr.2017.410`.

Barath, D., J. Matas and J. Noskova (2019). “MAGSAC: Marginalizing Sample Consensus”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10189–10197. DOI: `10.1109/cvpr.2019.01044`.

Barath, D., J. Noskova, M. Ivashechkin and J. Matas (2020). “MAGSAC++, a Fast, Reliable and Accurate Robust Estimator”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1301–1309. DOI: `10.1109/cvpr42600.2020.00138`.

Barnard, S. T. (1983). “Interpreting perspective images”. In: *Artificial Intelligence* 21, 4, pp. 435–462. DOI: `10.1016/s0004-3702(83)80021-6`.

Baumberg, A. (2000). “Reliable feature matching across widely separated views”. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*. IEEE Comput. Soc, pp. 774–781. DOI: `10.1109/cvpr.2000.855899`.

Bay, H., T. Tuytelaars and L. V. Gool (2006). “SURF: Speeded Up Robust Features”. In: *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, pp. 404–417. DOI: `10.1007/11744023_32`.

Bevilacqua, M. G., G. Caroti, A. Piemonte and D. Ulivieri (2019). “Reconstruction of lost Architectural Volumes by Integration of Photogrammetry from Archive Imagery with 3-D Models of the Status Quo”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W9, pp. 119–125. DOI: `10.5194/isprs-archives-xlii-2-w9-119-2019`.

Beyer, K., J. Goldstein, R. Ramakrishnan and U. Shaft (1999). “When Is “Nearest Neighbor” Meaningful?” In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 217–235. DOI: `10.1007/3-540-49257-7_15`.

Bitelli, G., M. Dellapasqua, V. A. Girelli, S. Sbaraglia and M. A. Tinia (2017). “Historical Photogrammetry and Terrestrial Laser Scanning for the 3d Virtual Reconstruction of Destroyed Structures: A Case Study in Italy”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-5/W1, pp. 113–119. DOI: `10.5194/isprs-archives-xlii-5-w1-113-2017`.

Bogdan, O., V. Eckstein, F. Rameau and J.-C. Bazin (2018). “DeepCalib”. In: *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production - CVMP ’18*. ACM Press, pp. 1–10. DOI: `10.1145/3278471.3278479`.

Bösemann, W. (2005). “Advances in photogrammetric measurement solutions”. In: *Computers in Industry* 56, 8-9, pp. 886–893. DOI: `10.1016/j.compind.2005.05.014`.

Bruschke, J., F. Niebling and M. Wacker (2018b). “Visualization of Orientations of Spatial Historical Photographs”. In: *Eurographics Workshop on Graphics and Cultural Heritage*, pp. 189–192. DOI: 10.2312/GCH.20181359.

Brutto, M. L. and P. Meli (2012). “Computer Vision Tools for 3D Modelling in Archaeology”. In: *International Journal of Heritage in the Digital Era* 1, 1_suppl, pp. 1–6. DOI: 10.1260/2047-4970.1.0.1.

Canciani, M., G. Spadafora, M. del Pilar Pastor Altaba, G. Formica, M. D’Angelico and C. Lebboroni (2018). “Geometric Constructive Traces in Drawings by Francesco Borromini”. In: *Advances in Intelligent Systems and Computing.* Springer International Publishing, pp. 208–218. DOI: 10.1007/978-3-319-95588-9_16.

Caprile, B. and V. Torre (1990). “Using vanishing points for camera calibration”. In: *International Journal of Computer Vision* 4, 2, pp. 127–139. DOI: 10.1007/bf00127813.

Cernea, D. (2020). “OpenMVS: Multi-View Stereo Reconstruction Library”. URL: https://cdcseacave.github.io/openMVS.

Chazalon, J., E. Carlinet, Y. Chen, J. Perret, B. Duménieu, C. Mallet, T. Géraud, V. Nguyen, N. Nguyen, J. Baloun, L. Lenc and P. Král (2021). “ICDAR 2021 Competition on Historical Map Segmentation”. In: *arXiv:2105.13265*, pp. 1–15.

Chen, Q., H. Wu and T. Wada (2004). “Camera Calibration with Two Arbitrary Coplanar Circles”. In: *Lecture Notes in Computer Science.* Springer Berlin Heidelberg, pp. 521–532. DOI: 10.1007/978-3-540-24672-5_41.

Chen, R., S. Han, J. Xu and H. Su (2019). “Point-Based Multi-View Stereo Network”. In: *arXiv:1908.04422*, pp. 1–13.

Chen, Y., E. Carlinet, J. Chazalon, C. Mallet, B. Duménieu and J. Perret (2021). “Vectorization of Historical Maps Using Deep Edge Filtering and Closed Shape Extraction”. In: *Document Analysis and Recognition – ICDAR 2021.* Springer International Publishing, pp. 510–525. DOI: 10.1007/978-3-030-86337-1_34.

Chum, O. and J. Matas (2005). “Matching with PROSAC — Progressive Sample Consensus”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05).* IEEE, pp. 220–226. DOI: 10.1109/cvpr.2005.221.

Chum, O., T. Werner and J. Matas (2005a). “Two-View Geometry Estimation Unaffected by a Dominant Plane”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05).* Vol. 1. IEEE, pp. 772–779. DOI: 10.1109/cvpr.2005.354.

Chum, O., A. Mikulik, M. Perdoch and J. Matas (2011). “Total recall II: Query expansion revisited”. In: *CVPR 2011.* IEEE. DOI: 10.1109/cvpr.2011.5995601.

Chum, O., J. Philbin, J. Sivic, M. Isard and A. Zisserman (2007). “Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval”. In: *2007 IEEE 11th International Conference on Computer Vision.* IEEE, pp. 1–8. DOI: 10.1109/iccv.2007.4408891.

Chum, O., J. Matas and J. Kittler (2003). “Locally Optimized RANSAC”. In: *Pattern Recognition.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 236–243. DOI: 10.1007/978-3-540-45243-0_31.

Chum, O., T. Pajdla and P. Sturm (2005b). “The geometric error for homographies”. In: *Computer Vision and Image Understanding* 97, 1, pp. 86–102. DOI: 10.1016/j.cviu.2004.03.004.

Cipolla, R., T. Drummond and D. Robertson (1999). “Camera Calibration from Vanishing Points in Image of Architectural Scenes”. In: *Procedings of the British Machine Vision Conference 1999*. British Machine Vision Association, pp. 382–391. DOI: 10.5244/c.13.38.

Cramer, M. and W. Kresse (2021). “Editorial PFG 5/2021”. In: *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 89, 5, pp. 369–369. DOI: 10.1007/s41064-021-00180-x.

Csurka, G., C. R. Dance and M. Humenberger (2018). “From handcrafted to deep local features”. In: *arXiv:1807.10254*, pp. 1–41.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei (2009). “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

DeTone, D., T. Malisiewicz and A. Rabinovich (2018). “SuperPoint: Self-Supervised Interest Point Detection and Description”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 337–33712. DOI: 10.1109/cvprw.2018.00060.

Dong, J. and S. Soatto (2015). “Domain-size pooling in local descriptors: DSP-SIFT”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5097–5106. DOI: 10.1109/cvpr.2015.7299145.

Dusmanu, M., I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii and T. Sattler (2019). “D2-Net: A Trainable CNN for Joint Description and Detection of Local Features”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 8084–8093. DOI: 10.1109/cvpr.2019.00828.

Eltner, A. and G. Sofia (2020). “Structure from motion photogrammetric technique”. In: *Developments in Earth Surface Processes*. Elsevier, pp. 1–24. DOI: 10.1016/b978-0-444-64177-9.00001-1.

Farella, E. M., E. Özdemir and F. Remondino (2021). “4D Building Reconstruction with Machine Learning and Historical Maps”. In: *Applied Sciences* 11, 4, p. 1445. DOI: 10.3390/app11041445.

Fischler, M. A. and R. C. Bolles (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24, 6, pp. 381–395. DOI: 10.1145/358669.358692.

Förstner, W. and E. Gülch (1987). “A fast operator for detection and precise location of distinct points, corners and centres of circular features”. In: *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*. Interlaken, pp. 281–305.

Friedman, J. H., J. L. Bentley and R. A. Finkel (1977). “An Algorithm for Finding Best Matches in Logarithmic Expected Time”. In: *ACM Transactions on Mathematical Software* 3, 3, pp. 209–226. DOI: 10.1145/355744.355745.

Grimm, A. (2021). “Albrecht Meydenbauer: Bauingenieur – Fotograf – Photogrammeter”. In: *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 89, 5, pp. 371–389. DOI: `10.1007/s41064-021-00183-8`.

Grün, A., F. Remondino and L. Zhang (2004). “Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan”. In: *The Photogrammetric Record* 19, 107, pp. 177–199. DOI: `10.1111/j.0031-868x.2004.00278.x`.

Grussenmeyer, P. and O. Al Khalil (2017). “From Metric Image Archives to Point Cloud Reconstruction: Case Study of the Great Mosque of Aleppo in Syria”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W5, pp. 295–301. ISSN: 2194-9034. DOI: `10.5194/isprs-archives-XLII-2-W5-295-2017`.

Han, X., T. Leung, Y. Jia, R. Sukthankar and A. C. Berg (2015). “MatchNet: Unifying feature and metric learning for patch-based matching”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3279–3286. DOI: `10.1109/cvpr.2015.7298948`.

Harris, C. and M. Stephens (1988). “A Combined Corner and Edge Detector”. In: *Procedings of the Alvey Vision Conference 1988*. Alvey Vision Club, pp. 1–6. DOI: `10.5244/c.2.23`.

Hartley, R. (1997a). “In defense of the eight-point algorithm”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 6, pp. 580–593. DOI: `10.1109/34.601246`.

Hartley, R. and A. Zisserman (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge university press. ISBN: 0521540518. DOI: `https://doi.org/10.1017/CBO9780511811685`.

Hartmann, W., M. Havlena and K. Schindler (2014). “Predicting Matchability”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 9–16. DOI: `10.1109/cvpr.2014.9`.

Hassner, T., V. Mayzels and L. Zelnik-Manor (2012). “On SIFTs and their scales”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1522–1528. DOI: `10.1109/cvpr.2012.6247842`.

Heinly, J., J. L. Schönberger, E. Dunn and J.-M. Frahm (2015). “Reconstructing the world* in six days”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3287–3295. DOI: `10.1109/cvpr.2015.7298949`.

Heitzler, M. and L. Hurni (2020). “Cartographic reconstruction of building footprints from historical maps: A study on the Swiss Siegfried map”. In: *Transactions in GIS* 24, 2, pp. 442–461. DOI: `10.1111/tgis.12610`.

Henze, F., H. Lehmann and B. Bruschke (2009b). “Analysis of Historic Maps and Images for Research on Urban Development of Baalbek/Lebanon”. In: *Photogrammetrie - Fernerkundung - Geoinformation* 2009, 3, pp. 221–234. DOI: `10.1127/0935-1221/2009/0017`.

Herold, H. and R. Hecht (2018). “3D Reconstruction of Urban History Based on Old Maps”. In: *Digital Research and Education in Architectural Heritage*. Springer International Publishing, pp. 63–79. DOI: `10.1007/978-3-319-76992-9_5`.

Hirsch, R. (2017). *Seizing the Light - A Social & Aesthetic History of Photography*. Routledge. DOI: `10.4324/9781315671994`.

Iglhaut, J., C. Cabo, S. Puliti, L. Piermattei, J. O'Connor and J. Rosette (2019). "Structure from Motion Photogrammetry in Forestry: a Review". In: *Current Forestry Reports* 5, 3, pp. 155–168. DOI: 10.1007/s40725-019-00094-3.

Jahrer, M., M. Grabner and H. Bischof (2008). "Learned local descriptors for recognition and matching". In: *Computer Vision Winter Workshop*. Vol. 2, pp. 39–46.

Jawahar, C., P. Jain, C. Jawahar and P. Jain (2006). "Homography Estimation from Planar Contours". In: *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pp. 877–884. DOI: 10.1109/3DPVT.2006.77.

Jiang, G. and L. Quan (2005). "Detection of concentric circles for camera calibration". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE, pp. 333–340. DOI: 10.1109/iccv.2005.73.

Jin, Y., D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi and E. Trulls (2020). "Image Matching Across Wide Baselines: From Paper to Practice". In: *International Journal of Computer Vision* 129, 2, pp. 517–547. DOI: 10.1007/s11263-020-01385-0.

Kalinowski, P., F. Both, T. Luhmann and U. Warnke (2021). "Data Fusion of Historical Photographs with Modern 3D Data for an Archaeological Excavation – Concept and First Results". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B2-2021, pp. 571–576. DOI: 10.5194/isprs-archives-xliii-b2-2021-571-2021.

Kazhdan, M. and H. Hoppe (2013). "Screened poisson surface reconstruction". In: *ACM Transactions on Graphics* 32, 3, pp. 1–13. DOI: 10.1145/2487228.2487237.

Kourniatis, N. and N. T. U. A. Architect (2019). "Leonardo da Vinci's The Last Supper: Reconstruction of the Room Using Reverse Geometric Perspective Processes". In: *Journal of Applied Mathematics and Physics* 07, 09, pp. 1941–1957. DOI: 10.4236/jamp.2019.79134.

Laguna, A. B., E. Riba, D. Ponsa and K. Mikolajczyk (2019). "Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 5835–5843. DOI: 10.1109/iccv.2019.00593.

Larsson, V., Z. Kukelova and Y. Zheng (2018). "Camera Pose Estimation with Unknown Principal Point". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2984–2992. DOI: 10.1109/cvpr.2018.00315.

Lepetit, V. and P. Fua (2006). "Keypoint recognition using randomized trees". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 9, pp. 1465–1479. DOI: 10.1109/tpami.2006.188.

Leutenegger, S., M. Chli and R. Y. Siegwart (2011). "BRISK: Binary Robust invariant scalable keypoints". In: *2011 International Conference on Computer Vision*. IEEE, pp. 2548–2555. DOI: 10.1109/iccv.2011.6126542.

Li, B., K. Peng, X. Ying and H. Zha (2010). "Simultaneous Vanishing Point Detection and Camera Calibration from Single Images". In: *Advances in Visual Computing*. Springer Berlin Heidelberg, pp. 151–160. DOI: 10.1007/978-3-642-17274-8_15.

Li, J., Q. Hu and M. Ai (2018). "RIFT: Multi-modal Image Matching Based on Radiation-invariant Feature Transform". In: *arXiv:1804.09493*, pp. 1–14.

Li, Z. and N. Snavely (2018). "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 2041–2050. DOI: 10.1109/cvpr.2018.00218.

Lim, F. K., N. Barbier, D. Feurer, S. Gourlet-Fleury, E. Paka, R. Pélissier, C. Pinet, G. V. Robin Pouteau, F. Vinatier and M. Réjou-Méchain (2022). "Using historical photographs and contemporary satellite images to explore 63 years of 3D changes in the forest structure of a Central African region". submitted to European Conference of Tropical Ecology, Montpellier, June 2022.

Lindeberg, T. (1998). "Feature Detection with Automatic Scale Selection". In: *International Journal of Computer Vision* 30, 2, pp. 79–116. DOI: 10.1023/a:1008045108935.

Longuet-Higgins, H. C. (1981). "A computer algorithm for reconstructing a scene from two projections". In: *Nature* 293, 5828, pp. 133–135. DOI: 10.1038/293133a0.

Lowe, D. (1999). "Object recognition from local scale-invariant features". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision.* Vol. 2. IEEE, pp. 1150–1157. DOI: 10.1109/iccv.1999.790410.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* 60, 2, pp. 91–110. ISSN: 0920-5691. DOI: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Lowe, D. G. (1987). "Three-dimensional object recognition from single two-dimensional images". In: *Artificial Intelligence* 31, 3, pp. 355–395. DOI: 10.1016/0004-3702(87)90070-1.

Luhmann, T. (2010). "Close range photogrammetry for industrial applications". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 65, 6, pp. 558–569. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2010.06.003.

Luhmann, T., S. Robson, S. Kyle and J. Boehm (2019). *Close-Range Photogrammetry and 3D Imaging.* De Gruyter. DOI: 10.1515/9783110607253.

Maddern, W., G. Pascoe, C. Linegar and P. Newman (2016). "1 year, 1000 km: The Oxford RobotCar dataset". In: *The International Journal of Robotics Research* 36, 1, pp. 3–15. DOI: 10.1177/0278364916679498.

Maiwald, F., J. Bruschke, C. Lehmann and F. Niebling (2019b). "A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR". In: *Virtual Archaeology Review* 10, 21, pp. 1–13. DOI: 10.4995/var.2019.11867.

Maiwald, F., T. Vietze, D. Schneider, F. Henze, S. Münster and F. Niebling (2017). "Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities". In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, pp. 447–452. ISSN: 1682-1750. DOI: https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017.

Martin-Brualla, R., N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy and D. Duckworth (2020). "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections". In: *arXiv:2008.02268*, pp. 1–15.

Matas, J., O. Chum, M. Urban and T. Pajdla (2002). “Robust Wide Baseline Stereo from Maximally Stable Extremal Regions”. In: *Procedings of the British Machine Vision Conference 2002*. British Machine Vision Association, pp. 1–10. DOI: `10.5244/c.16.36`.

Matas, J., O. Chum, M. Urban and T. Pajdla (2004). “Robust wide-baseline stereo from maximally stable extremal regions”. In: *Image and vision computing* 22, 10, pp. 761–767. ISSN: 0262-8856. DOI: `https://doi.org/10.1016/j.imavis.2004.02.006`.

Maybank, S. (1990). “The projective geometry of ambiguous surfaces”. In: *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* 332, 1623, pp. 1–47. DOI: `10.1098/rsta.1990.0099`.

McCarthy, J. (2014). “Multi-image photogrammetry as a practical tool for cultural heritage survey and community engagement”. In: *Journal of Archaeological Science* 43, pp. 175–185. DOI: `10.1016/j.jas.2014.01.010`.

Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. V. Gool (2005). “A Comparison of Affine Region Detectors”. In: *International Journal of Computer Vision* 65, 1-2, pp. 43–72. DOI: `10.1007/s11263-005-3848-x`.

Mikolajczyk, K. and C. Schmid (2004). “Scale & Affine Invariant Interest Point Detectors”. In: *International Journal of Computer Vision* 60, 1, pp. 63–86. DOI: `10.1023/b:visi.0000027790.02288.f2`.

Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi and R. Ng (2022). “NeRF”. In: *Communications of the ACM* 65, 1, pp. 99–106. DOI: `10.1145/3503250`.

Mishchuk, A., D. Mishkin, F. Radenović and J. Matas (2017). “Working Hard to Know Your Neighbor’s Margins: Local Descriptor Learning Loss”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 4829–4840. ISBN: 9781510860964.

Mishkin, D., J. Matas and M. Perdoch (2015a). “MODS: Fast and robust method for two-view matching”. In: *Computer Vision and Image Understanding* 141, pp. 81–93. ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2015.08.005`.

Moravec, H. (1977). “Towards Automatic Visual Obstacle Avoidance”. In: *Proceedings of 5th International Joint Conference on Artificial Intelligence (IJCAI ’77)*. Vol. 1, p. 584.

Morel, J.-M. and G. Yu (2009). “ASIFT: A New Framework for Fully Affine Invariant Image Comparison”. In: *SIAM Journal on Imaging Sciences* 2, 2, pp. 438–469. DOI: `10.1137/080732730`.

Muja, M. and D. G. Lowe (2009). “Flann, fast library for approximate nearest neighbors”. In: *International Conference on Computer Vision Theory and Applications (VISAPP’09)*. Vol. 3. INSTICC Press, pp. 1–21.

Noh, H., A. Araujo, J. Sim, T. Weyand and B. Han (2017). “Large-Scale Image Retrieval with Attentive Deep Local Features”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 3476–3485. DOI: `10.1109/iccv.2017.374`.

Ono, Y., E. Trulls, P. Fua and K. M. Yi (2018). “LF-Net: Learning Local Features from Images”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., pp. 6237–6247. DOI: `https://arxiv.org/pdf/1805.09662.pdf`.

Opdenbosch, D. V., M. Oelsch, A. Garcea, T. Aykut and E. Steinbach (2018). “Selection and Compression of Local Binary Features for Remote Visual SLAM”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 7270–7277. DOI: 10.1109/icra.2018.8463202.

Polic, M., S. Steidl, C. Albl, Z. Kukelova and T. Pajdla (2020). “Uncertainty Based Camera Model Selection”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5990–5999. DOI: 10.1109/cvpr42600.2020.00603.

Powell, M. J. D. (1964). “An efficient method for finding the minimum of a function of several variables without calculating derivatives”. In: *The Computer Journal* 7, 2, pp. 155–162. DOI: 10.1093/comjnl/7.2.155.

Pritchett, P. and A. Zisserman (1998). “Wide baseline stereo matching”. In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. Narosa Publishing House, pp. 754–760. DOI: 10.1109/iccv.1998.710802.

Rabin, J., J. Delon and Y. Gousseau (2009). “A Statistical Approach to the Matching of Local Features”. In: *SIAM Journal on Imaging Sciences* 2, 3, pp. 931–958. DOI: 10.1137/090751359.

Radenovic, F., A. Iscen, G. Tolias, Y. Avrithis and O. Chum (2018). “Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 5706–5715. DOI: 10.1109/cvpr.2018.00598.

Radenovic, F., J. L. Schonberger, D. Ji, J.-M. Frahm, O. Chum and J. Matas (2016). “From Dusk Till Dawn: Modeling in the Dark”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5488–5496. DOI: 10.1109/cvpr.2016.592.

Raguram, R., O. Chum, M. Pollefeys, J. Matas and J.-M. Frahm (2013). “USAC: A Universal Framework for Random Sample Consensus”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8, pp. 2022–2038. DOI: 10.1109/tpami.2012.257.

Razavian, A. S., J. Sullivan, S. Carlsson and A. Maki (2016). “Visual instance retrieval with deep convolutional networks”. In: *ITE Transactions on Media Technology and Applications* 4, 3, pp. 251–258. DOI: 10.3169/mta.4.251.

Remondino, F., E. Nocerino, I. Toschi and F. Menna (2017). “A Critical Review of Automated Photogrammetric Processing of Large Datasets”. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W5, pp. 591–599. DOI: 10.5194/isprs-archives-xlii-2-w5-591-2017.

Ressl, C. (2003). “Geometry, constraints and computation of the trifocal tensor”. Dissertation. TU Wien.

Revaud, J., P. Weinzaepfel, C. D. Souza, N. Pion, G. Csurka, Y. Cabon and M. Humenberger (2019). “R2D2: Repeatable and Reliable Detector and Descriptor”. In: *arXiv:1906.06195*, pp. 1–11.

Ronneberger, O., P. Fischer and T. Brox (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Lecture Notes in Computer Science*. Springer International Publishing, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

Rosten, E. and T. Drummond (2006). "Machine Learning for High-Speed Corner Detection". In: *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, pp. 430–443. DOI: 10.1007/11744023_34.

Rothe, R., M. Guillaumin and L. V. Gool (2015). "Non-maximum Suppression for Object Detection by Passing Messages Between Windows". In: *Computer Vision – ACCV 2014*. Springer International Publishing, pp. 290–306. DOI: 10.1007/978-3-319-16865-4_19.

Rupnik, E., M. Daakir and M. P. Deseilligny (2017). "MicMac – a free, open-source solution for photogrammetry". In: *Open Geospatial Data, Software and Standards* 2, 1. DOI: 10.1186/s40965-017-0027-2.

Sarlin, P.-E., D. DeTone, T. Malisiewicz and A. Rabinovich (2020). "SuperGlue: Learning Feature Matching with Graph Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947. DOI: arXiv:1911.11763[.

Sattler, T., W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl and T. Pajdla (2018). "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 8601–8610. DOI: 10.1109/cvpr.2018.00897.

Savinov, N., A. Seki, L. Ladicky, T. Sattler and M. Pollefeys (2017). "Quad-Networks: Unsupervised Learning to Rank for Interest Point Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3929–3937. DOI: 10.1109/cvpr.2017.418.

Schaffalitzky, F. and A. Zisserman (2002). "Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?"" In: *Computer Vision — ECCV 2002*. Springer Berlin Heidelberg, pp. 414–431. DOI: 10.1007/3-540-47969-4_28.

Schindler, G. and F. Dellaert (2012). "4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections". In: *Journal of Multimedia* 7, 2, pp. 124–131. ISSN: 1796-2048. DOI: https://doi.org/10.4304/jmm.7.2.124-131.

Schönberger, J. L. and J.-M. Frahm (2016). "Structure-from-Motion Revisited". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 4104–4113. DOI: 10.1109/cvpr.2016.445.

Schütz, M. (2016). "Potree: Rendering large point clouds in web browsers". MA thesis.

Shen, S. (2013). "Accurate Multiple View 3D Reconstruction Using Patch-Based Stereo for Large-Scale Scenes". In: *IEEE Transactions on Image Processing* 22, 5, pp. 1901–1914. DOI: 10.1109/tip.2013.2237921.

Shi, J. and Tomasi (1994). "Good features to track". In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*. IEEE Comput. Soc. Press, pp. 593–600. DOI: 10.1109/cvpr.1994.323794.

Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer (2015). "Discriminative Learning of Deep Convolutional Feature Point Descriptors". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 118–126. DOI: 10.1109/iccv.2015.22.

Simonyan, K. and A. Zisserman (2014). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv preprint arXiv:1409.1556*, pp. 1–14.

Sinkhorn, R. and P. Knopp (1967). “Concerning nonnegative matrices and doubly stochastic matrices”. In: *Pacific Journal of Mathematics* 21, 2, pp. 343–348. DOI: 10.2140/pjm.1967.21.343.

Snavely, N., S. M. Seitz and R. Szeliski (2006). “Photo tourism: exploring photo collections in 3D”. In: *ACM Transactions on Graphics* 25, 3, pp. 835–846. DOI: 10.1145/1141911.1141964.

Snavely, N., S. M. Seitz and R. Szeliski (2008). “Skeletal graphs for efficient structure from motion”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8. DOI: 10.1109/cvpr.2008.4587678.

Stafford, S. (2004). *The new Nikon compendium : cameras, lenses & accessories since 1917*. New York: Lark Books. ISBN: 1579905927.

Stewart, C. V. (1999). “Robust Parameter Estimation in Computer Vision”. In: *SIAM Review* 41, 3, pp. 513–537. DOI: 10.1137/s0036144598345802.

Strecha, C., A. Lindner, K. Ali and P. Fua (2009). “Training for Task Specific Keypoint Detection”. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 151–160. DOI: 10.1007/978-3-642-03798-6_16.

Sun, J., Z. Shen, Y. Wang, H. Bao and X. Zhou (2021). “LoFTR: Detector-Free Local Feature Matching with Transformers”. In: *arXiv:2104.00680*, pp. 1–10.

Szeliski, R. (2010). *Computer Vision*. Springer-Verlag GmbH. ISBN: 9781848829350. URL: https://www.ebook.de/de/product/19111262/richard_szeliski_computer_vision.html.

Tian, Y., B. Fan and F. Wu (2017). “L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6128–6136. DOI: 10.1109/cvpr.2017.649.

Torr, P. and A. Zisserman (2000). “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry”. In: *Computer Vision and Image Understanding* 78, 1, pp. 138–156. DOI: 10.1006/cviu.1999.0832.

Torr, P., A. Zisserman and S. Maybank (1995). “Robust detection of degenerate configurations for the fundamental matrix”. In: *Proceedings of IEEE International Conference on Computer Vision*. IEEE Comput. Soc. Press, pp. 1037–1042. DOI: 10.1109/iccv.1995.466820.

Tuytelaars, T. and L. van Gool (2000). “Wide Baseline Stereo Matching based on Local, Affinely Invariant Regions”. In: *Procedings of the British Machine Vision Conference 2000*. British Machine Vision Association, pp. 1–14. DOI: 10.5244/c.14.38.

Tyszkiewicz, M. J., P. Fua and E. Trulls (2020). “DISK: Learning local features with policy gradient”. In: *arXiv:2006.13566*, pp. 1–15.

Ullmann, S. (1979). “The interpretation of structure from motion”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 203, 1153, pp. 405–426. DOI: 10.1098/rspb.1979.0006.

Umeyama, S. (1991). "Least-squares estimation of transformation parameters between two point patterns". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 4, pp. 376–380. DOI: 10.1109/34.88573.

Venkateswar, V. and R. Chellappa (1995). "Hierarchical stereo and motion correspondence using feature groupings". In: *International Journal of Computer Vision* 15, 3, pp. 245–269. DOI: 10.1007/bf01451743.

Verdie, Y., K. M. Yi, P. Fua and V. Lepetit (2015). "TILDE: A Temporally Invariant Learned DEtector". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5279–5288. DOI: 10.1109/cvpr.2015.7299165.

Ware, C. (2021). "Foundations for an Applied Science of Data Visualization". In: *Information Visualization*. Elsevier, pp. 1–29. DOI: 10.1016/b978-0-12-812875-6.00001-3.

Westoby, M., J. Brasington, N. Glasser, M. Hambrey and J. Reynolds (2012). "'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications". In: *Geomorphology* 179, pp. 300–314. DOI: 10.1016/j.geomorph.2012.08.021.

Wiedemann, A., M. Hemmleb and J. Albertz (2000). "Reconstruction of historical buildings based on images from the Meydenbauer archives". In: *International Archives of Photogrammetry and Remote Sensing* 33, B5/2; PART 5, pp. 887–893. ISSN: 0256-1840.

Williams, R. J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8, 3-4, pp. 229–256. DOI: 10.1007/bf00992696.

Winder, S. A. J. and M. Brown (2007). "Learning Local Image Descriptors". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8. DOI: 10.1109/cvpr.2007.382971.

Workman, S., C. Greenwell, M. Zhai, R. Baltenberger and N. Jacobs (2015). "DEEPFOCAL: A method for direct focal length estimation". In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 1369–1373. DOI: 10.1109/icip.2015.7351024.

Wu, C. (2013). "Towards linear-time incremental structure from motion". In: *3D Vision-3DV 2013, 2013 International conference on*. IEEE, pp. 127–134. ISBN: 0769550673. DOI: https://doi.org/10.1109/3DV.2013.25.

Wu, C. (2014). "Critical Configurations for Radial Distortion Self-Calibration". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 25–32. DOI: 10.1109/cvpr.2014.11.

Xiong, Y., Z. Lin, G. Li, C. Xian and C. Peng (2021). "Camera focal length from distances in a single image". In: *The Visual Computer* 37, 9-11, pp. 2869–2881. DOI: 10.1007/s00371-021-02233-z.

Yao, Y., Z. Luo, S. Li, T. Fang and L. Quan (2018). "MVSNet: Depth Inference for Unstructured Multi-view Stereo". In: *Computer Vision – ECCV 2018*. Springer International Publishing, pp. 785–801. DOI: 10.1007/978-3-030-01237-3_47.

Yi, K. M., E. Trulls, V. Lepetit and P. Fua (2016). "LIFT: Learned Invariant Feature Transform". In: *Computer Vision – ECCV 2016*. Springer International Publishing, pp. 467–483. DOI: 10.1007/978-3-319-46466-4_28.

Zagoruyko, S. and N. Komodakis (2017). “Deep compare: A study on using convolutional neural networks to compare image patches”. In: *Computer Vision and Image Understanding* 164, pp. 38–55. DOI: 10.1016/j.cviu.2017.10.007.

Zawieska, D. and J. Markiewicz (2016). “Development of Photogrammetric Documentation of the Borough at Biskupin Based on Archival Photographs - First Results”. In: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection.* Springer International Publishing, pp. 3–9. DOI: 10.1007/978-3-319-48974-2_1.

Zhang, J., M. Marszałek, S. Lazebnik and C. Schmid (2006). “Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study”. In: *International Journal of Computer Vision* 73, 2, pp. 213–238. DOI: 10.1007/s11263-006-9794-4.

Zhang, L., E. Rupnik and M. Pierrot-Deseilligny (2021). “Feature matching for multi-epoch historical aerial images”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 182, pp. 176–189. DOI: 10.1016/j.isprsjprs.2021.10.008.

List of Figures

List of Tables

List of Abbreviations

2D two-dimensional

3D three-dimensional

4D four-dimensional

6-DoF 6-degrees of freedom

AKAZE Accelerated KAZE

AP average precision

AR Augmented Reality

ASIFT Affine-SIFT

CBIR Content-based image retrieval

CNN Convolutional Neural Network

CSIFT Colored SIFT

CVPR Computer Vision and Pattern Recognition

DEGENSAC Degeneracy Sample Consensus

DELF Deep Local Feature

DISK Discrete Keypoints

DLT Direct Linear Transformation

DoG Difference-of-Gaussian

DSP-SIFT Domain-Size Pooled SIFT

FLANN Fast Library for Approximate Nearest Neighbors

FOV field of view

FP false positive

FSC Fast Sample Consensus

GIS Geographic Information Systems

GUI graphical user interface

HMD head-mounted display

IR image retrieval

LEA Layer Extraction Approach

LO-RANSAC Locally-Optimized RANSAC

LOD level of detail

LoG Laplacian-of-Gaussian

MAGSAC Marginalizing Sample Consensus

mAP mean average precision

MD metadata

MLESAC Maximum Likelihood Estimation Sample Consensus

MODS Matching On Demand with View Synthesis

MR Mixed Reality

MSER Maximally Stable Extremal Regions

NN Nearest Neighbor

NRD nonlinear radiation distortions

OoI Object of Interest

OpARA TU Dresden Open Access Repository and Archive

ORB Oriented FAST and Rotated BRIEF

PCA principal component analysis

PR precision-recall

PROSAC Progressive Sampling Consensus

RANSAC Random Sample Consensus

RIFT Radiation-Invariant Feature Transform

RMS root mean square

ScaDS Competence Center for Scalable Data Services and Solutions

ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig

SfM Structure-from-Motion

SIFT Scale Invariant Feature Transform

SLAM Simultaneous Localization and Mapping

SLUB Saxon State and University Library Dresden

SPA single-page application

SURF Speeded-Up Robust Features

SVD singular value decompositon

TP true positive

VP vanishing point

VPD Vanishing Point Detection

VR Virtual Reality

VRML Virtual Reality Modeling Language