Doctoral theses at NTNU, 2023:153

Evdokia Saiti

Deep Learning based frameworks for 3D registration of differential and multimodal data

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of Science and Technology

Evdokia Saiti

Deep Learning based frameworks for 3D registration of differential and multimodal data

Thesis for the Degree of Philosophiae Doctor

Trondheim, May 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Computer Science

ISBN 978-82-326-7010-9 (printed ver.)
ISBN 978-82-326-5814-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:153

Printed by NTNU Grafisk senter

Abstract

In recent decades, Visual Computing methodologies such as image processing and computer vision have addressed problems in the field of Cultural Heritage (CH) resulting in significant benefits. Specifically, accurate scanning methods have proved invaluable for documenting cultural heritage assets. However, such scans can also be used to track changes over time and to create holistic models of CH artefacts, resulting from multiple scan modalities. This in turn necessitates solving specific challenges in the task of registration, a classic problem in Visual Computing.

Informally, registration is the action of placing two geometric datasets with overlap (e.g. point clouds) in a common reference frame so that the areas of overlap match as closely as possible. This thesis focuses on two special cases of 3D registration: cross-time and multimodal. The first research area concerns the registration of differential 3D data, where the object of interest may have changed over time. The second research area concerns the registration of data from different modalities; specifically 3D point clouds and micro-CT volumes have been addressed. As both problems are too complex to address with direct algorithms while training instances exist or can be generated, it was chosen to apply deep learning methodologies to solve them and the results have been very encouraging.

Additionally, the cross-time registration solution has been extended into an automated framework for change monitoring and difference detection for CH objects, while the multimodal method was combined with the cross-time method in order to monitor changes on both the surface and inner structure of CH objects.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfilment of the requirements for the degree of Philosophiae Doctor (PhD).

The work was performed at the Department of Computer Science, NTNU, Trondheim and was supervised by professor Theoharis Theoharis, and co-supervised by professor Robert Sitnik of Warsaw University of Technology (WUT).

The conducted research was part of the CHANGE project and has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813789.

Acknowledgements

I would like to express my gratitude to all the people who supported me during the last three years of my Ph.D. journey. First and foremost, I would like to express my deepest gratitude to my supervisor Theoharis Theoharis for his unwavering support, guidance, dedication and for numerous beneficial discussions (both scheduled and impromptu). Thank you for all the enlightenment, both during the learning process as well as personally. Also, I would like to thank my co-supervisor Robert Sitnik for being there whenever I needed his assistance and feedback.

This research work was accomplished under the auspices of the CHANGE ('Cultural Heritage Analysis for new generations') project. CHANGE project provided me with the opportunity to broaden my knowledge in the area of cultural heritage and opened avenues for international collaboration. I would like to extend my sincere thanks to the coordinating team of the project for their support and for giving me the the opportunity to participate on the NO-CHANGE (CHANGE - Norwegian Network) project and disseminate my research throughout Norway. Many thanks to all early researchers of the CHANGE project for their friendship and support. Namely, I would like to thank Sunita, Sia and Amalia for the great and fun times we had collaborating and socializing together.

My biggest thanks should go to my family for all their support and endless love. In particular, I would like to thank my mother for her unconditional support and sacrifices but mostly for the countless times she has looked after all of us throughout the years. I could not have undertaken this journey without the support of my beloved husband, Antonios. Your encouragement and advice when times got rough are highly appreciated. Thank you for always being my motivation, my inspiration and my compass. Most importantly, I would like to thank my daughters, Poly and Elli, for being the shining light that beckoned me through my research journey.

My acknowledgments would not be complete without sending my deepest appreciation to my beloved father and grandpa-in-law who, although no longer with us, continue to inspire me. I really wish you could see this thesis completed.

Contents

List of Figures

Part I

Research Overview

Chapter 1

Introduction

This thesis addresses a classic problem in the field of Visual Computing, that of *registration*, which stems from disciplines such as robotics [54, 13], medical imaging [81, 44, 75] and cultural heritage analysis [79, 70], among many others. Informally, registration (aka *alignment*) is the action of placing two geometric datasets with overlap (e.g. point clouds) in a common reference frame so that the areas of overlap match as closely as possible. Registration of both unimodal and multimodal datasets is a first crucial step in many shape analysis tasks, such as 3D model reconstruction [45, 5], model fitting [20], 3D object recognition [8, 46] and retrieval [17] and semantic segmentation [87].

This research was part of the European Union's Horizon 2020 research and innovation program with the acronym 'CHANGE' [10] (Cultural Heritage Analysis for New Generations). The main objective of the project is to develop methodologies for monitoring and assessing changes in Cultural Heritage (CH) objects. The developed methodologies are intended to allow CH experts to track changes over time and develop effective strategies for conserving and preserving the CH objects. While this research's motivation has arisen from the context of the CHANGE project and the CH field, the presented techniques can potentially be applied to other areas as well.

1.1 Motivation

Mankind's Cultural Heritage (CH) has been inherited from the past and should be maintained for the future. CH resources are under constant threat due to natural or human-induced actions which gradually cause our heritage to vanish. Thus, protection, preservation, and promotion of our heritage are of importance.

Recent advances in imaging, computer vision and computational methods, have greatly aided CH by offering researchers a 'sixth sense' for understanding traces of the past. For example, an accurate, high-resolution digital model can reveal details and features of the object that might not be visible to the naked eye; accurate digital acquisitions, performed at regular intervals, could detect very small deformations and cracks before serious damage or decay. This knowledge can enable more efficient analysis and opportune interventions to prevent further deterioration and preserve the assets [48].

In this doctoral thesis, we addressed two open problems in registration: cross-time registration, which can facilitate the monitoring process of CH assets, and multimodal registration, which can merge independent scan modalities and generate enhanced 3D models of CH objects. By leveraging the power of deep learning networks, we overcame challenges such as finding accurate correspondence between modalities that do not share the same characteristics or between 3D models whose geometry has changed over time.

1.2 Research Objectives

Our objectives are to investigate cross-time and multimodal registration and their application to CH. We aim to achieve these objectives by trying to answser the following Research Questions (RQ) in this thesis.

Part 1: Cross-time 3D registration

- **RQ1:** How to design an automatic registration framework to assess the monitoring of CH objects?
- **RQ2:** How can an automatic cross-time registration method be applied to real CH datasets to facilitate monitoring and detection of fabricated CH objects?

Part 2: Multimodal 3D registration

- **RQ3:** What are the state-of-the-art techniques for registering multimodal 3D data and what are the open challenges regardless of the field of application?
- **RQ4:** How to design an automatic multimodal registration framework to fuse 3D data from point cloud and CT-volume modalities?

1.3 List of Contributing Papers

This section provides an overview of the publications included in this thesis and their relationship to the main topics of cross-time and multimodal registration as well as to the individual research questions. Five papers, labeled A-E, are included as the core contributions. Three of them have been published in peer-reviewed channels, while the other two are under review in a scientific journal and a conference. Additionally, two supplementary papers labeled F and G were written as part of the conducted research. However, they are not considered part of this thesis since they did not make any significant contributions.

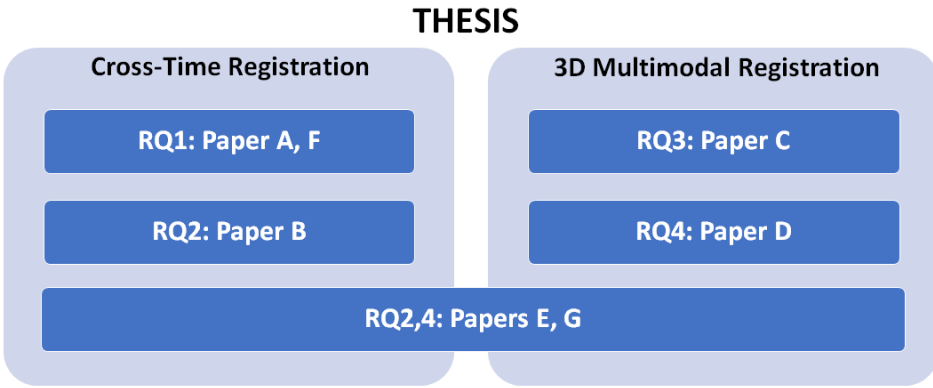


Figure 1.1: Overview of research papers and their relationship to research questions.

Figure 1.1 provides a visual representation of the research questions under the two main topics and how the research papers relate to them. Since the research and respectively the research questions are divided into two main directions, we devoted each paper in each different research question. Papers A and B focus on the first axis of the research, the cross-time registration problem. Paper A addresses the problem by proposing a deep learning framework for automated registration of cross time data and Paper B applies the aforementioned methodology in order to compare and detect differences between two similar museum's objects. The second axis, the 3D multimodal registration, is addressed with Paper C, which gives a review of the latest state of the art techniques and Paper D, where a novel framework for aligning 3D point clouds and 3D volumes is presented. Finally, Paper E bridges the two axes by proposing a pipeline for monitoring the external and internal structure of CH objects.

The list of the included papers and their main contributions is given below. The papers can be found in full in part II.

The list of the core contributing papers:

Paper A **Cross-time registration of 3D point clouds**
Saiti Evdokia, Antonios Danelakis, Theoharis Theoharis
In *Computer & Graphics*, Volume 99, Elsevier, 2021

Paper B **An automated approach for change and difference detection on Cultural Heritage Applications**
Saiti Evdokia, Sunita Saha, Eryk Bunsch, Robert Sitnik, Theoharis Theoharis
In *Digital Applications in Archaeology and Cultural Heritage*, under review as of March 2023

Paper C **An application independent review of multimodal 3D registration methods**
Saiti Evdokia, Theoharis Theoharis
In *Computer & Graphics*, Volume 91, Elsevier, 2020

Paper D **Multimodal registration across 3D point clouds and CT-volumes**
Saiti Evdokia, Theoharis Theoharis
In *Computer & Graphics*, Volume 106, Elsevier, 2022

Paper E **A pipeline for monitoring the external and inner structure of cultural heritage objects**
Saiti Evdokia, Theoharis Theoharis
In *Archiving Conference 2023*, submitted on March 2023

The list of the supplementary papers:

Paper F **Adaption of imaging techniques for monitoring cultural heritage objects**
Siatou Amalia, Athanasia Papanikolaou, **Evdokia Saiti**
In *Advanced Nondestructive and Structural Techniques for Diagnosis,* Redesign and Health Monitoring for the Preservation of Cultural Heritage: Selected work from the TMM-CH 2021
Cham: Springer International Publishing, 2022

Paper G **Automated 3D registration techniques for applications in cultural heritage monitoring**
Saiti Evdokia, Theoharis Theoharis
Chapter under review in the CHANGE project's Final Book: *Cultural Heritage Analysis for New GEnerations (CHANGE)*

1.4 Dissertation Structure

This thesis consists of two parts and is structured as follows:

- **Part I: Research Overview**

  This part provides a general overview of the work carried out. Chapter 1 covers the motivation of the research, the research objectives to be reached and the research questions to be answered. Chapter 2 gives the required background knowledge at a rather high level so that a reader can comfortably skim through it. Chapter 3 summarizes the research contributions for each of the core papers, while Chapter 4 outlines the main contributions of this thesis. The thesis is concluded in Chapter 5 with remarks and directions for future research.

- **Part II: Publications**

  This part contains the collection of full-length research papers submitted as an element of this thesis.

Chapter 2

Background in 3D Registration

This chapter provides a brief overview of the relevant context in which this thesis was conducted, focusing on 3D registration and its applications in Cultural Heritage. A detailed survey of 3D registration methods applicable to multimodal and differential data can be found in **paper C** (which is a survey) and **paper A** respectively. Rather than repeating these, the aim of this chapter is to provide the reader with the contextual understanding.

Registration is a challenging task for various problems in computer vision and computer graphics. The general aim is to bring together two or more geometric datasets by finding the transformation that optimally aligns them in a common reference frame. The datasets may, for example, represent the shape of the same object, two or more partial but overlapping instances of the same object or two similar 3D objects (as for example castings from the same mould). In this thesis we shall use the alternative term *alignment* as a synonym for registration.

Although our research extends to multimodal registration, the registration problem is classically defined across two 3D point clouds. After defining the 3D registration problem in Section 2.1 and a brief review of the main methods for 3D point cloud registration in 2.2, Section 2.3 presents an overview of the application of registration techniques in CH.

2.1 Problem Statement

Given two 3D point clouds, the source $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, ..., N\}$ and the target $\mathbf{Q} = \{q_j \in \mathbb{R}^3 |\ j = 1, 2, ..., M\}$, the objective is to recover the unknown rigid transformation $\mathbf{T}$ subject to minimizing a distance function between the source $\mathbf{P}$ and the target $\mathbf{Q}$ point clouds.

A rigid transformation in 3D can be represented by a transformation matrix $\mathbf{T}$ which consists of two components; a rotation submatrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, where $\mathbf{T}$ is a homogeneous 4×4 matrix:

$$\mathbf{T} = \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline 0 & 1 \end{array}\right] \tag{2.1}$$

where $\mathbf{T} \in SE(3)$, $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$. $SE(3)$ is the special Euclidean group of rigid transformations in 3D space (rotations and translations), while $SO(3)$ is the special orthogonal group of rotations in Euclidean Space $\mathbb{R}^3$.

Then the problem of rigid registration between two discrete point clouds can be formulated as [43] :

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} d(\mathbf{R}\mathbf{p_i} + \mathbf{t}, \mathbf{Q}) \tag{2.2}$$

where $d(\mathbf{p}, \mathbf{Q})$ is the distance of an arbitrary point $\mathbf{p} \in \mathbf{P}$ to the point cloud $\mathbf{Q}$. A common definition of this distance is:

$$d(\mathbf{p}, \mathbf{Q}) = \min_{\mathbf{q}\in\mathbf{Q}} d'(\mathbf{p}, \mathbf{q}) \tag{2.3}$$

where $d'(\mathbf{p}, \mathbf{q})$ is the distance between two points in space.

Equation (2.3) is referred to as the distance or error metric. Many methods [88, 42] use the squared Euclidean norm as the distance metric and optimize Equation (2.2) using least squares:

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \|\mathbf{R}\mathbf{p_i} + \mathbf{t} - \widehat{\mathbf{q}}_\mathbf{i}\|^2 \tag{2.4}$$

where $\widehat{\mathbf{q}}_\mathbf{i}$ is the closest point in $\mathbf{Q}$ to each point $\mathbf{p_i} \in \mathbf{P}$ based on the transformation $\mathbf{T}(\mathbf{R}, \mathbf{t})$.

2.2 3D Registration Methods

The 3D registration problem is a broad research topic and advances have been made over the years, resulting in a plethora of different strategies and algorithms.

Based on the level of precision in the alignment result, registration methods can be roughly classified into two broad categories, *coarse* and *fine*. Coarse registration techniques try to find a rough initial alignment between the two models without any prior information about their relative pose, whereas fine registration methods try to find a more precise alignment given a course initial alignment transformation.

The classic method for fine alignment is the **Iterative Closest Point (ICP)** [6]. Starting with an initialization of the alignment, ICP iteratively alternates between solving two sub-problems: finding point-to-point correspondences and estimating an accurate transformation based on current correspondences. ICP is a local registration method, meaning that correspondences are found using the nearest neighbor search based on Euclidean distance.

Over the course of almost three decades, ICP has been considered the golden standard, has been extensively used in various applications and has been commonly employed as a standard benchmark for evaluating new alignment techniques. However, ICP has several limitations:

- ICP's accuracy is highly dependent on the initial position of the models. Being a local method, ICP is effective only when the initial pose of the input geometries is close to the global optimum, otherwise it can converge to a local minimum.
- ICP is computationally expensive. The iterative nature of the algorithm and its point-to-point correspondence nature result in high computational complexity.
- ICP is sensitive to outliers as it assumes that each point of the source model corresponds to its closest point in the target model. Thus, if there are many outliers or the models do not overlap sufficiently, ICP can lead to biased transformations and erroneous alignments.

The simplicity of ICP combined with its drawbacks have inspired the development of a plethora of alternative methods. A large number of variations of ICP have appeared, which attempt to address one or more of the aforementioned limitations [60]. Conceptually different approaches have also arisen, proposing other strategies to improve robustness and complexity. Several of these methods attempt to tackle challenges such as complexity, but the final result is lacking in accuracy, requiring a refinement by a fine registration approach.

Feature-based methods identify feature-level similarities and correspondences, rather than finding point-to-point correspondences. After identifying the proper

features, these techniques use robust fitting or optimization techniques such as RANSAC [16] and Fast Global Registration [94] to achieve registration. These techniques are more efficient than point-level methods, but their accuracy is highly dependent on the quality of the features selected. Feature-based techniques generally involve three steps: feature detection, feature description and correspondence estimation. Features are a small group of interest points that can be detected on both objects, due to their distinctiveness or geometric stability under different transformations. Each feature can be delineated by a descriptor that characterizes its geometric information. Two main categories of descriptors exist: global and local. Global descriptors represent the geometric information of an entire 3D object, whereas local descriptors [33, 36, 25] encode local information at each feature point [24]. Specifically, for 3D registration local descriptors are more commonly used, because they can identify similar localities between the two surfaces to be aligned by exploiting the geometric properties around a certain point and its neighborhood. Feature-based methods do not require a good initial pose to converge, but they are generally less accurate since they rely on a few keypoints instead of dense point correspondences.

Probabilistic methods transform the point clouds into probability distributions and match their statistical properties. In particular, GMM-based methods such as [80, 32, 15], represent point clouds as Gaussian Mixture Models (GMMs), thereby reformulating the problem in a lower dimension. These methods have gained popularity due to their robustness to noise and outliers. Moreover, GMMs represent an straightforward way to formulate distributions and by lowering their dimension, these methods are computationally efficient [58].

Learning-based methods utilize machine learning techniques to achieve faster and more robust results than classical methods. Neural networks can be integrated into different stages of the registration pipeline. For instance, there are methods that generate data-driven features by machine learning approaches, and then use traditional approaches to calculate the final registration [2, 35]. The extracted features can be more detailed and invariant than the hand-crafted ones. Additionally, there are methods that use neural networks for the entire registration process, replacing both the feature extraction step and the registration estimation, by deep networks [68, 35].

2.3 3D Registration in Cultural Heritage

A complete digital recording of CH is a multi-step process with many challenges. CH objects require special care due to their value, articulation and fragility. Acquisition protocols are often much stricter than the ones from other applications. The surface of the CH asset cannot be touched or physically altered. Furthermore,

the physical access to the CH assets can be limited by their size, shape or location. Digital documentation of CH objects requires precise measurements at different scales and resolutions to ensure that they are accurately recorded. For example, large sites or statues require less accurate acquisitions than small delicate objects with fine details. There has been significant growth in research on registration in several applications, resulting in multiple methods and surveys [78, 37, 31, 65]. Specifically, 3D registration is an essential task in many CH applications ranging from reconstruction [22] and reassembly [45] to digital documentation analysis [48], preservation [21] and monitoring [62]. The applications of registration in cultural heritage can be divided into three broad categories, based on the type of the data that is aligned:

- **Multiview registration** aligns models of the same object captured by the same sensor from different viewpoints during a single session in order to form a complete 3D model of the object.
- **Cross-time registration** aligns temporally different models of the same object, taken at different times. This is usually done for the purpose of monitoring the object and to reveal any changes.
- **Multimodal registration** aligns models of the same object captured by different types of sensors, resulting in data of different modalities. After registration, the information of the different modalities can be fused in order to obtain a more complete and detailed description of the object.

Multiview registration is a main component of the 3D modeling workflow which obtains the raw data from the acquisition system and creates the final digitized 3D model. Due to the visibility constraints of laser scanning acquisition and the complexity of an object's surface, it is usually not possible to obtain the complete information of an object in one go. Therefore, it is necessary to scan a segment of the object or site of interest at a time, resulting in multiple scans from different points of view, each with its own distinct coordinate system. In order to reconstruct the final complete object, the several partial scans need to be combined, which is achieved by aligning the different partially overlapping scans into a single coordinate system (Figure 2.1(a)).

In many applications related to CH documentation, a coarse registration is first performed by using external reference points. But such external reference points are not always available or are not sufficiently accurate. The automatic registration of surface scans of 3D objects without the use of external reference points is an active research area.

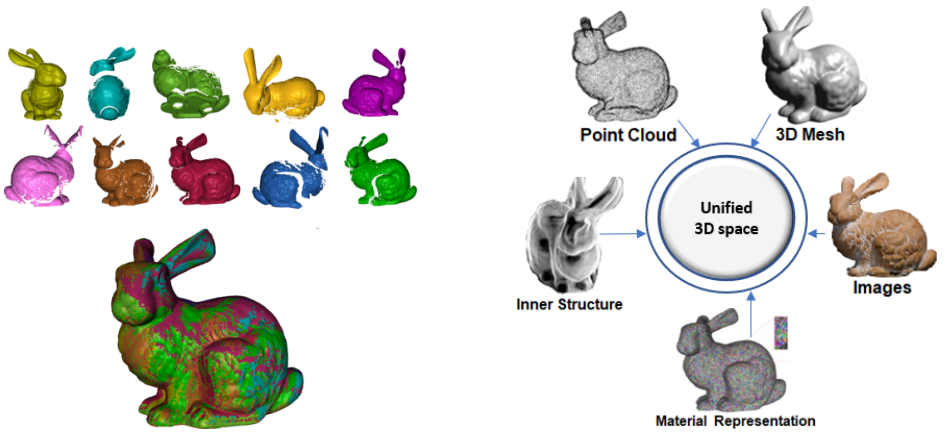


Figure 2.1: (a) Multiview 3D data registration as presented in [19]. Partial 3D views are brought into a common coordinate system and merged to build a complete 3D model of the object. (b) Multimodal 3D registration. Different modalities of the same object are aligned and merged into a complete holistic 3D model.

Multiview registration procedure consists of two stages, an initial course alignment followed by a refinement to achieve an accurate solution.

The coarse registration aims to align the scans without any prior information about their relative pose. Correspondence-based methods are often used, where initial keypoints of the geometry are detected, then correspondences are assigned between the input objects and, finally, a registration is achieved by computing the transformation that best aligns the corresponding points with respect to minimizing a specific distance function. A large number of both traditional [4, 92] and deep learning [83, 30] techniques have been proposed, each having both strengths and limitations. In the fine registration step, the transformation result is obtained through local search algorithms [60], with the most common one again being Iterative Closest Point (ICP) [6, 11] and its variants.

Cross-time registration refers to the process of aligning 3D models of the same object acquired at different points in time (see Figure 2.2). Digital models do not degrade themselves and can thus be used as a reference for monitoring the structural health of CH assets in a reliable and nondestructive manner. Periodic scanning and analysis can identify possible accidental or man-made alterations of the objects (e.g. through conservation actions). Moreover, microgeometric changes over time are measured and analyzed in order to support conservation strategies or to identify patterns of degradation [51].

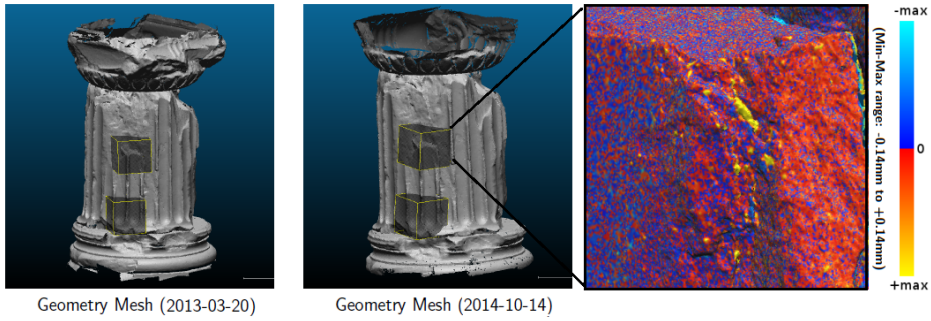


Figure 2.2: Two geometry meshes of the Elefsis column acquired at different times 20 months apart (data from the PRESIOUS project [55]) and a close-up view of their difference. One can observe that erosion is taking place by the removal of small randomly distributed parts (yellow).

Most methods for cross-time registration use general 3D registration algorithms, which identify keypoints on both models and align them [38]. Case studies like [40] and [73], have used physical targets as reference points, in order to provide a reliable way to track changes over time. However, this may not be feasible or practical for many invaluable and fragile CH objects. Moreover, if registration is performed without any external reference points, these methods ignore any temporal misalignment. However, due to deterioration effects, local details may have been eroded away over time or, worse, objects may have fragmented or lost important parts of their shape. Methods such as ICP [6], try to bring the surfaces as close as possible, overlooking the global (but known) effect of surface-recession. For example, dense parts or large 'flat' corresponding areas are often registered directly on top of each other, which can result in an incorrect alignment if the surface has globally receded between the different time points [47]. In order to overcome such challenges and obtain a more stable result, a common practice is to exclude from the registration process the areas where the maximal changes take place [9], but this requires manual effort by the CH experts and specific knowledge of each case study.

Ultimately, the alignment of differential 3D data does require manual effort by CH experts in current practice, making it time-consuming and bringing subjectivity to the final result. Alignment errors can lead to incorrect interpretations and conclusions about the alterations and deteriorations that CH objects have undergone. To ensure the accuracy of the alignment process, it is important to use a combination of reliable automatic registration techniques followed by validatation of the results using the knowledge of CH experts.

Multimodal registration is the process of aligning and merging multiple datasets acquired from different modalities, such as 3D models, CT scans, or thermal images to create a complete 3D representation of an object or scene (see Figure 2.1(b)). By combining different modalities, it is possible to create a detailed 3D model of the original object which can provide a more complete representation of the object's geometry, texture and physical properties. Such a comprehensive model can be used to study the object in detail, to guide restoration strategies before any physical intervention is carried out, or to identify areas of damage or deterioration [49, 57].

Multimodal registration is commonly applied in many research areas such as medicine [64] but it has great value in CH studies also. There are many projects that integrate data from different 3D acquisition techniques in order to create a complete way to document CH artifacts [29]. Pitzalis et al. [53] used a user-assisted process to combine data from different sources and create an enhanced 3D model of the Cylinder seal of Ibni-Sharrum. They combined photogrammetry with X-ray CT, μ-topography, and high-resolution images. Neutron and X-ray imaging have been combined in [39] and [72]. In [82], X-ray fluorescence (XRF) Imaging and X-ray CT data were combined, and created a 3D model with elemental composition information which assisted the analysis of baroque sculptures.

The combination of information about the geometry and the inner structure can be advantageous for understanding the object's history, construction, and condition. As mentioned in [67], most case studies convert the data from one modality to another, and then use a unimodal registration method to align the data. For example, [69] combined a photogrammetric 3D model with the micro-CT data by first reconstructing the 3D surface model from the micro-CT data and then using ICP to align the two 3D point clouds.

Chapter 3

Research Contributions

In this thesis, two special problems in 3D registration have been studied and two deep learning methods have been proposed to tackle them. This chapter provides an overview of the research papers that form the core of the thesis. The papers have been grouped into two categories based on the addressed registration problem: cross-time registration and multimodal registration. **Papers A and B** address the cross-time registration problem, while **Papers C and D** focus on 3D multimodal registration. **Paper E** provides a unified pipeline of both cross-time and multimodal registration.

A summary of each paper is presented, highlighting the objectives, main ideas, and contributions. In addition to the papers, non-peer-reviewed contributions, such as the source codes and datasets are publicly available to promote reproducibility and facilitate future studies.

3.1 Paper A: Cross-Time registration of 3D point clouds

Saiti Evdokia, Antonios Danelakis, Theoharis Theoharis
In *Computer & Graphics*, Volume 99, Elsevier, 2021

Paper A addresses RQ1, outlines the cross-time registration problem as a special case of 3D registration and proposes a deep learning registration framework to address it [64].

CH objects change over time, due to the interaction between the object and environmental factors (i.e. atmospheric oxygen, humidity, various pollutants) or repair processes (i.e. chemical or physical repair methods). Conservation science is constantly seeking efficient methods to preserve the CH objects by monitoring,

detecting and obviating any changes over time [52]. In monitoring, microgeometric changes over time must be measured and analyzed in order to detect any alterations that are occurring. In this matter, the geometric acquisition and measurements of a CH object produce snapshots of 3D models and can be used to track an object through time, in order to document different phases of its lifetime and make preservation decisions.

In general, data of the same object captured at different acquisitions is likely to contain geometrical differences. In order to facilitate the study of change, the data of the different captures need to be registered. After correct registration the process of monitoring can be automated in a non-invasive manner as even minute modifications on the object's surface or shape can be detected and measured. Motivated by the fact that traditional registration techniques are inadequate to address this case with the required high accuracy, this paper proposes *CrossTimeReg*, a deep learning framework which focuses on the pairwise cross-time registration of CH objects that have undergone erosion over time. The method overcomes limitations such as computational complexity of the iterative methods, the necessity for point-level correspondence and copes with large 3D models. It also exploits a known fact about erosion i.e. that areas of high curvature erode more.

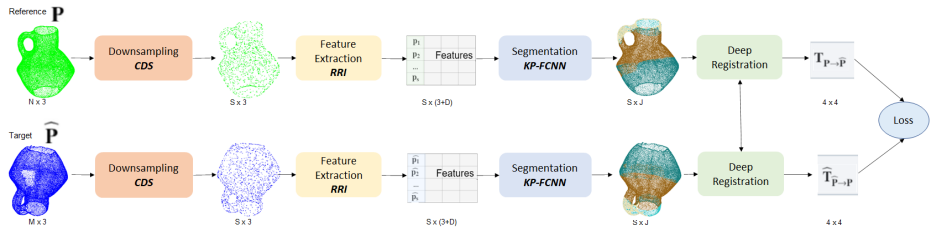


Figure 3.1: Overview of the presented *CrossTimeReg* registration framework.

CrossTimeReg consists of four components, as shown in Figure 3.1. The initial and eroded point clouds (also referred to as *source* and *target*) are first downsampled using the Curvature Downsampling (CDS) block, where the points that are less likely to be significantly altered by erosion are retained. These points are expected to be those with the minimum principal curvature [23]. The intuitive reason behind this is that such points are less exposed to erosion/degradation processes or conservation activities. Thus they are considered to be a more robust representation of the object across such processes or activities. Next, the Feature Extraction block (RRI) computes features that remain fixed under different orientations. The features along with the point clouds are then sent to a Siamese architecture of point cloud segmentation networks (KP-FCNN) (Figure 3.2a). Each point cloud is segmented into a specific number of components by estimating for each point

the component that it belongs to. Finally, the optimization module of DeepGMR [91], takes the segmented point clouds and calculates for each the Gaussian Mixture Model including mixture weights, means and covariances. Finally, the 3D rigid transformation parameters are computed by aligning the segment centroids (weighted by the covariances) using a deep registration block, a weighted version of the SVD solution (Figure 3.2b).

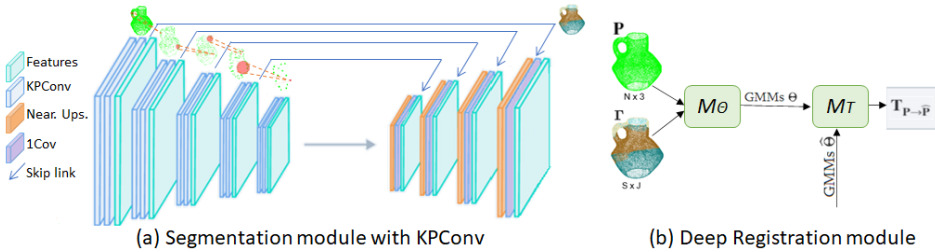


Figure 3.2: (a)The KP-FCNN segmentation network architecture. The encoder transforms the features by consecutive KPConv blocks and the decoder upsamples the features to the initial input resolution. Skip links are used to pass the features between intermediate layers. An example placed at the top of each layer shows the downsampling process and how the receptive field (red sphere) grows proportionally to the downsampling size. (b) The deep registration module with its two differentiable computing blocks M_Θ and M_T.

In order to benchmark and train cross-time 3D registration algorithms, the paper proposes ECHO; a dataset of eroded 3D models of CH objects along with the ground truth needed for evaluating cross-time registration algorithms.

The main contributions of the paper are as follows:

- The problem of cross-time 3D registration is formally defined and a framework for cross-time 3D registration is proposed.
- A down-sampling methodology that detects the most valuable points for cross-time registration is proposed.
- A benchmark for evaluating both traditional and cross-time registration algorithms, ECHO, is created and made publicly available [14].
- An extensive evaluation of both geometry-based and deep learning state-of-the-art approaches on 3D cross-time registration is performed.

3.2 Paper B: An automated approach for change and difference detection in cultural heritage applications

Saiti Evdokia, Sunita Saha, Eryk Bunsch, Robert Sitnik, Theoharis Theoharis
In *Digital Applications in Archaeology and Cultural Heritage*, submitted in March 2023

Paper B proposes an application of CrossTimeReg from **Paper A** to monitor and assess changes in CH objects (RQ2).

3D digital models can be durable and unalterable, and can thus be used as a reference for monitoring changes and identifying differences in a reliable and nondestructive way. An accurate, high-resolution digital model can reveal details and features of an object that might not be visible to the naked eye. The analysis of changes in an object's geometry, allows CH experts to track changes over time, develop effective conservation strategies, or even detect forgeries [18]. In this context, the analysis of geometric changes is a topic of great importance in CH, with three primary scopes of application:

- *Health monitoring and tracking of aging processes of CH objects*. Periodic scanning and monitoring can identify possible accidental or man-made alterations of the objects. Moreover, microgeometric changes over time are measured and analyzed in order to support conservation strategies or to identify patterns of degradation or deterioration [51].
- *Authenticity identification of a CH object*. 3D models constitute a digital archive that can be useful to identify replicas or forgeries. The automated comparison and analysis of the digital model can empower the object's authentication process. For example, the information provided by a geometric comparison can be used to verify a CH object's authenticity after returning from a loan.
- *Comparison of multiple CH objects of similar shape in order to analyse their history*. Throughout history, it was common practice for a workshop to create a series of artefacts, similar in shape and dimensions. By comparing the 3D models of multiple similar objects, the CH experts can verify a possible origin from the same workshop production and assess the technique of their manufacture.

The presented method can be applied in all of the three aforementioned applications but for Paper B, the selected case study was from the third type of application.

In particular, we study two ceramic sculptures from the collection of the Museum of King Jan III's Palace at Wilanów in Poland. The ceramic sculptures named 'Zephyr and Flora' were made at the Meissen workshop at different times, based on a master ancient composition. Even though the sculptures are similar in geometry, several changes can be detected and provide insights for establishing the time, place and process of the objects' production. The research was conducted by a multidisciplinary research group with representatives from computer science and cultural heritage.

This paper focuses on the automated data analysis for change and difference detection on CH objects. The novelty of the proposed method lies in the following; the adaptation of two recently developed techniques for automatic registration [64] and change segmentation [62], in a common framework to detect differences between CH objects efficiently and robustly; the showcasing of the combined method on two instances of real CH objects that originated from the same workshop: to the best of our knowledge, this is the first time that such instances have been geometrically compared and analyzed.

3.3 Paper C: An application independent review of multimodal 3D registration methods

Saiti Evdokia, Theoharis Theoharis
In *Computer & Graphics*, Volume 91, Elsevier, 2020

This paper addresses RQ3 and presents a literature review to summarize the existing state-of-the-art 3D multimodal registration methods. The goal of this survey is to unify and categorize 3D multimodal registration techniques across application domains. The review was restricted to methods where one or both modalities are three-dimensional.

The paper reviews the methods used for aligning multimodal 3D data. We use the term *multimodal* to refer to two datasets with qualitative variability in shape or appearance; thus having different dimensions (e.g. 3D/2D images, X-ray / MRI), different data structures (e.g. 3D point cloud and an MRI volume) or different physical and anatomical principles (e.g. MRI and CT volumes). The methods arose in fields including medical, cultural heritage and urban mapping. We tried to identify common trends, applications and evaluation metrics for multimodal registration.

The paper explicitly defines the *3D multimodal registration* problem and categorizes the methods based on their algorithmic strategies, rather than their registration attributes. It also overviews the publicly available multimodal datasets and

finally evaluates and experimentally compares the registration methods using publicly available source code.

The main contributions of this work are:

- A comprehensive categorization of 3D multimodal registration across application domains.
- An overview of publicly available multimodal registration datasets.
- An extensive evaluation of comparable implementations of publicly available 3D multimodal registration methods is performed.
- Trends in strategies, applications and evaluation metrics for multimodal registration are identified.

3.4 Paper D: Multimodal registration across 3D point clouds and CT-volumes

Saiti Evdokia, Theoharis Theoharis
In *Computer & Graphics*, Volume 106, Elsevier, 2022

Paper D addresses RQ4 and specifically the multimodal registration of 3D models from 3D surface scanning and computed tomography (CT). These modalities were chosen because they contribute to and supplement each other in order to create a complete and accurate 3D virtual representation of a CH object.

Geometry acquired from 3D surface scanners is a core aspect of a digital model but is limited since only data from the surface are acquired and the inner structure of the object cannot be documented. The penetrative capabilities of tomographic scanning allow the digitization of the interior of an object without having to perform physically invasive actions. By combining these two modalities, it is possible to produce a holistic 3D representation of an object. This combination first requires the registration of the two modalities in a common reference frame.

The main challenge in aligning these modalities is finding accurate correspondence between them since these two modalities do not exhibit the same characteristics, structure, or physical principles. The main idea of the method presented in this paper is that since both modalities represent the same object, there will exist common features to guide a supervised deep registration network. In our methodology and experiments, we take advantage of a ground truth in order to train a neural network to properly align the two modalities.

The paper presents *PCD2VOL* [63], a deep learning framework capable of aligning two different modalities, without transforming either of them before feeding them

to the network. *PCD2VOL* aligns 3D surface data with 3D CT volume data and was, to the best of our knowledge, the first time that a deep learning network was trained to register such modalities.

The presented framework, as shown in Figure 3.3, consists of three main components. First, the 3D point cloud and the 3D CT volume are fed into two modality-specific feature extraction network blocks to identify regional and geometric features of each modality independently. Then, the modality-based features are passed to a siamese architecture of cross-modal attention blocks, in order to capture local features and their global correspondence across the modalities. Finally, the deep registration block processes the fused feature representations to extract the registration parameters.

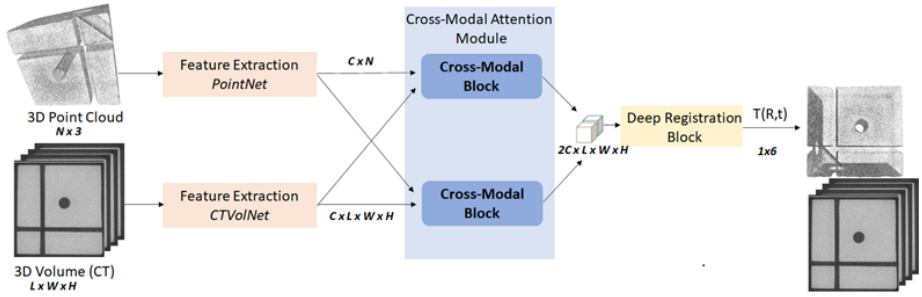


Figure 3.3: Overview of the proposed *Multimodal* 3D registration framework. 1. Each input modality (Point Cloud and 3D CT Volume) is fed into an independent feature extractor network that is suitable for that modality. 2. The captured features are fed to a siamese architecture of cross-modal attention blocks. 3. The registration block fuses the cross-modal features into the final registration parameters.

The main contributions of this paper are:

- The problem of multimodal 3D registration of CT volumes and 3D point clouds is formally defined.
- A deep learning framework for 3D registration of CT volumes and 3D point clouds is proposed, which employs a siamese architecture for a novel attention mechanism for effective multimodality fusion.
- A multimodal dataset for evaluating algorithms for aligning CT volumes and 3D point clouds is presented and made publicly available [1].

3.5 Paper E: A pipeline for monitoring the external and inner structure of cultural heritage objects

Saiti Evdokia, Theoharis Theoharis
In *Archiving Conference 2023*, submitted in March 2023

Paper E combines RQ2 and RQ4 and proposes a pipeline for monitoring both the surface and the inner structure of CH objects. The approach has the potential to facilitate the monitoring through time and change detection of CH objects in a more holistic way.

The pipeline proposed takes as input two different sets of 3D models and 3D volumes acquired at different times from 3D surface and CT scanning respectively, and registers both modalities in a multitemporal way. The results show the possibilities of this methodology for accurate multitemporal documentation of both surface and inner structure.

Figure 3.4 illustrates the presented pipeline, which consists of two parts, the cross-time and multimodal registration. The workflow takes as an input two sets of the 3D model and 3D volume of a CH object acquired at different times. Between these times, the surface and the interior may have altered due to environmental or human actions. The first part adapts the CrossTimeReg framework of **Paper A** for aligning 3D point clouds across time. The resulting aligned 3D models along with their respective 3D volumes, are subsequently forwarded to the next step. This step consists of two parallel multimodal registration frameworks as presented in **Paper D**. This stage is responsible for aligning and fusing each pair of 3D model and 3D volume.

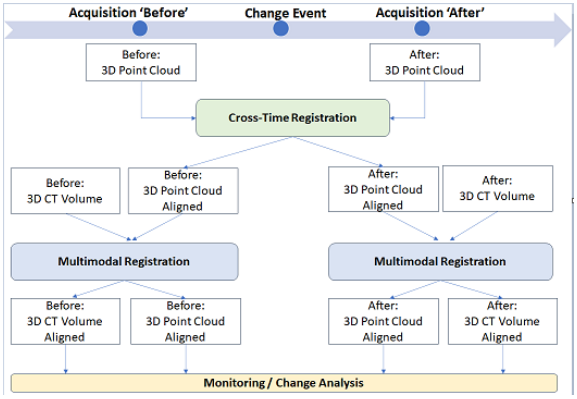


Figure 3.4: The pipeline of **Paper E**, for digital monitoring of external and internal structure of CH objects.

Chapter 4

Discussion

In this chapter, we highlight the contributions to the two main research areas; cross-time and multimodal registration.

4.1 Research Contribution

4.1.1 Cross-time Registration

Papers A, B and **E** address cross-time registration. **Paper A** proposes CrossTimeReg, a framework designed to accurately align 3D models that have undergone surface modifications. **Papers B** and **E** incorporate CrossTimeReg into two different pipelines, which facilitate two distinct case studies. **Paper B** addresses an intriguing topic of geometry analysis in the context of museums, where CrossTimeReg is applied to compare multiple CH objects of similar shapes in order to analyse their origins. In **Paper E**, the two special cases of registration discussed in this thesis, are combined into a unified pipeline, which monitors both the external and internal structure of CH objects. Below, we discuss further the contributions on cross-time registration, concentrating on the CrossTimeReg framework and the ECHO dataset.

The CrossTimeReg framework
Paper B focuses on the challenging problem of 3D cross-time registration and introduces *CrossTimeReg*, a deep learning method that can accurately align 3D point clouds that have undergone differential changes over time.

CrossTimeReg achieves state-of-art accuracy and robustness to large initial transformations while being computationally efficient. Several experiments have been conducted to evaluate the performance of CrossTimeReg in comparison to existing

state-of-the-art methods. These experiments included different levels of erosion, ranging from no erosion to 60 years of erosion with acid rain, and evaluation on the real data of the PRESIOUS project [77, 76].

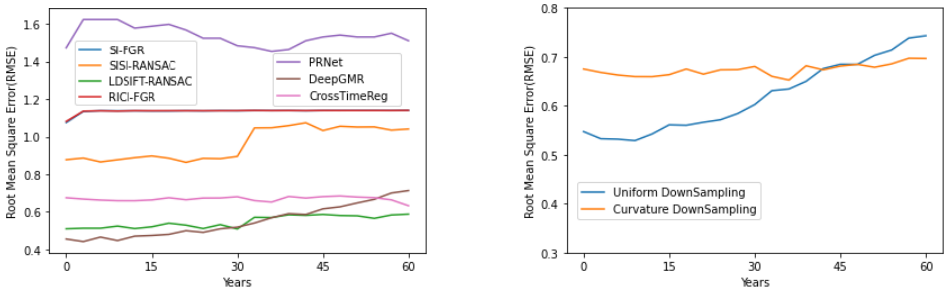
Table 4.1: Registration results on the ECHO dataset when only random rotations, translations and 60 years of erosion are performed on the initial objects. The metrics evaluated are rotation error, Error(R), translation error, Error(t), root mean square error, RMSE, root mean square distance, RMSD and Recall with threshold 0.2. Bold and dark gray denote best and second best performing methods for each measure respectively. For fairness reasons, we have not included in bold, cases where CrossTimeReg performs best when trained on the training partition of ECHO; instead such cases are in bold italics.

	Method	Registration		Error(R)	Error(t)	RMSE	RMSD	Recall$_a$(%)	Mean Exec. Time (sec)
		Local	Global						
Geometry-based	ICP [6]	✓		1.6992	42.5667	38.6065	42.583	0	34
	FPFH-RANSAC [61, 16]		✓	1.8314	29.2151	29.3316	29.2326	0	32
	SI-FGR [33, 94]		✓	1.8202	0.0629	1.1298	1.1344	21.91	32
	SISI-RANSAC [12, 16]		✓	0.9984	0.1044	0.6870	0.6877	96.88	67
	LD-SIFT -RANSAC [12, 16]		✓	**0.3496**	0.0793	**0.2789**	**0.2878**	98.79	68
	RICI-FGR [7, 94]		✓	1.1396	**0.0495**	1.1832	1.1396	20.77	38
Deep Learning	PRNet [84]	✓		1.7514	1.0184	1.4723	1.4858	43.12	14
	PointNetLK [2]	✓		1.7413	29.2389	29.2514	29.2561	0	11
	PCRNet[68]		✓	1.8095	49.3442	49.3603	49.3600	0	10
	RPM-Net [89]		✓	1.6993	29.2594	29.2784	29.2755	0	15
	DCP [83]		✓	1.6881	38.6109	38.6542	38.6133	0	15
	DeepGMR [91]		✓	1.0065	0.0673	0.9454	**0.6746**	**99.31**	4
	CrossTimeReg [64]		✓	**0.9942**	**0.0448**	**0.6764**	0.6812	**99.55**	6
	CrossTimeReg (trained on ECHO)		✓	***0.1397***	0.0714	***0.2606***	0.6928	***99.98***	6

Table 4.1 shows that CrossTimeReg generally outperforms the state-of-the-art in most performance metrics evaluated. The performance of geometry-based global registration methods like RANSAC [16] and FGR [94] is highly dependent on feature matching or keypoint detection from hand-crafted descriptors. It is also highlighted that the LDSIFT descriptor [12] performs considerably better that the rest of the state-of-the-art, since it is a rotation- and scale-invariant descriptor. As, in the simple erosion model used, erosion affects the surface of an object evenly, the scale invariant features result in better recovery of the correct transformation. However, LDSIFT suffers from large computation time and memory requirements, which limit its use in real-time applications or large datasets. In contrast, CrossTimeReg is more computationally efficient, making it a practical solution for real-time applications.

Furthermore, we conducted experiments on the ECHO dataset to evaluate the performance of CrossTimeReg and other state-of-the-art methods under different levels of erosion. The erosion model has been applied to the point clouds to simulate different levels of degradation, ranging from 1 year to 60 years. The results of these experiments (Figure 4.1(a)) indicate that CrossTimeReg exhibits stable performance across levels of erosion, with its performance even (curiously) increasing

slightly at the highest levels. This suggests that CrossTimeReg is robust to different levels of degradation and can effectively handle cross-time registration of point clouds with significant differences. The stable performance of CrossTimeReg under different levels of erosion further supports its potential as a reliable and robust solution for 3D cross-time registration.



(a) **(b)**

Figure 4.1: (a) Comparison between different registration methods on examples from the ECHO dataset for cross-time registration across different levels of erosion. (b) Ablation study of downsampling methodologies on different levels of erosion based on the ECHO dataset.

The contribution of the proposed downsampling method on the final performance of CrossTimeReg was evaluated by comparing it against the case where uniform sampling is used (Figure 4.1(b)). The results of this comparison showed that the proposed downsampling method behaved stably across different levels of erosion, demonstrating consistent and reliable performance. In contrast, uniform sampling had better RMSE results on small erosion values but the performance degraded as the level of erosion increased. This suggests that the proposed downsampling method is more robust and effective in handling point clouds with different levels of erosion, making it a more practical and reliable solution for cross-time registration.

The importance of these findings is that they demonstrate the practicality and usefulness of CrossTimeReg on real data, where point clouds may undergo differential changes over time due to environmental factors.

The ECHO dataset

One of the main challenges was the lack of a publicly available dataset with ground truth for cross-time 3D registration. In order to benchmark and train cross-time 3D registration algorithms, **paper A** proposed the ECHO dataset. Starting from the publicly available dataset of CH objects of [71], we first applied a random rigid transformation $(\mathbf{R}, \mathbf{t})$ to the objects; then we used an artificial erosion process to

erode the transformed objects [47]. Since erosion is performed in situ and the $(\mathbf{R}, \mathbf{t})$ parameters are known, we automatically have the ground truth for training and benchmarking cross-time registration algorithms. The process is outlined in Figure 4.2 and an example is given in Figure 4.3.

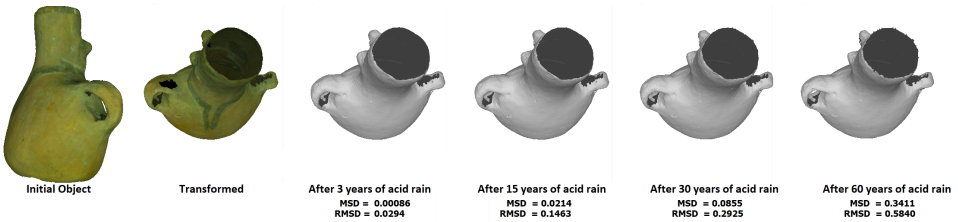


Figure 4.2: The steps of the ECHO dataset creation for an example object. The object is initially transformed rigidly and then the erosion simulator runs for 20 epochs of 3 years each. In this example, the initial model is shown degraded due to the effect of acid rain after 3, 15, 30 and 60 years. Below each eroded model, its point-wise MSD and RMSD differences from the transformed model are given.

Figure 4.3: ECHO dataset: example from the simulated dataset for weathering. Original CH object from SHREC2021 dataset (a), along with the transformed (reference) instance (b). In (c) the reference object is depicted in gray color and superimposed are the same object after 30 years of aging in blue and the after 60 years in red.

4.1.2 Multimodal Registration

3D Multimodal Registration: a review of the state-of-the-art
In **Paper C** the current state-of-the-art methods in 3D multimodal registration are reviewed, leading to several useful results.

This was the first review paper on multimodal 3D registration. Initially, the '3D multimodal registration' problem was defined and the methods were categorized based on their algorithmic strategies, rather than their registration attributes, and independently of application domain. Moreover, we overviewed the publicly available multimodal datasets and finally, evaluated and experimentally compared the registration methods with publicly available source code. During the review of

the methods, we identified that the term multimodal registration has largely been 'used' or 'abused' in the literature, referring to such aspects as the same object from different viewpoints, the same object at different moments in time or the same object scanned by different sensors. A definition therefore seemed appropriate in order to filter the methods that would be part of the survey.

A key question was what should be the characteristics of two modalities in order to be considered different? To answer this question, we have tried to locate what makes multimodal registration a more challenging task than unimodal registration. It has been observed that registration methods that perform well in the unimodal case, do not necessarily perform well in the multimodal. In unimodal registration, data have similar or correlated statistical properties and it is rather straightforward to recognize correspondences or a similarity metric. The core difficulty in multimodal registration is in identifying structure correspondences across modalities or in defining a general rule for identifying similarity across two modalities with different physical principles. Therefore, as mentioned before, we will use the term multimodal to refer to two datasets with qualitative variability in shape and appearance; thus having different data structure, different dimensionality or different physical and anatomical principles.

Some key findings of the review follow. Over the years many multimodal registration techniques have been proposed mainly related to the medical field. This is because, in the medical field, there are many body scanning modalities that need to be registered in order to acquire an integrated view of the patient. Registration of 3D models to 2D images is the most common case across applications from different fields. So, why not use registration methods created for another field also for our purposes in cultural heritage? There is a plethora of medically oriented algorithms, which align modalities such as MRI, TRUS, and X-ray images. But unfortunately, these were modalities that are not so popular in the CH field. Moreover, the physiology and characteristics of medical data differentiate them from cultural heritage data, as discussed below, and the need was therefore identified for a registration methodology mainly focused on the CH modalities at hand, i.e. surface scans and CT.

The modeling of cultural heritage objects faces technical challenges due to causes such as their shape, their articulation and size. For example, CH objects may be large. Statues or temples are far larger than a human brain or a human body part or many mechanical objects. In addition, CH objects are scanned in high resolution in order to depict any detail on their surface, resulting in large data which are in general hard to render and process.

The PCD2VOL framework

Paper D addresses the challenging problem of aligning 3D point clouds to 3D CT Volumes. *PCD2VOL* is a direct solution for registering different 3D modalities, without any prior conversion of one modality to the other. Both modalities are treated directly, so as to avoid information loss and time penalty. *PCD2VOL* employs a siamese architecture of cross-modal attention blocks that captures and fuses features of two structurally different modalities. An attention mechanism enables a model to focus on important information for a task; thus it has been applied widely to various computer vision problems, including image classification [90], object detection [85], image generation [93] and image captioning [86]. Recently this technique has also been used for multimodal registration. In [95], RGB images and point clouds were fused by learning feature interactions between the modalities with a cross-modal attention scheme while in [74] a self-attention mechanism was developed specifically for aligning 3D medical volumes of MRI and TRUS modalities.

Our problem is generic in that it concerns the alignment of 3D modalities that are complementary since they jointly describe the interior and the surface of a 3D object. The proposed network exploits cross attention for the challenging task of aligning 3D modalities of different geometric data structures. PCD2VOL is a combination of CNN for volume feature extraction [59], geometric deep learning for point cloud feature extraction [56] and a siamese architecture of cross modal attention network, trained to identify correspondences and fuse regular input data formats (like 3D voxels) and irregular 3D geometric data (like 3D point clouds). To the best of our knowledge, this is the first time that registration of such different modalities, without projecting one modality onto the other, is explored.

The biggest challenge was the lack of a benchmark for evaluating our method. An accurate and fair comparison between our method and previous approaches was not straightforward because we could not identify any previous registration methods that directly align point clouds and CT volumes. We thus opted to use ICP as a baseline. Since ICP works on point clouds, we pre-processed the CT volumes and converted them into point clouds. We then run the ICP algorithm between these point clouds and the point clouds of the '3DPCD-CT' dataset. In general, ICP fails when it comes to large rigid transformation differences. To succeed, ICP needs a good initial transformation estimation (which may not be the case in many applications). Thus, in most cases, ICP did not converge. Moreover, ICP and other state of the art registration techniques require inputs of the same modality (point clouds in general), necessitating the conversion of one of the inputs in order to address the modality gap. This conversion involves loss of information, which can significantly affect the registration result. In addition, such a conversion can be

expensive, especially when large 3D volumes are involved, as in CH applications.

Table 4.2: Performance comparison between multimodal registration methods.

Method	Data Modalities		Modalities Structure		Data	Runtime	Initial	TRE	Percent
	M1	M2	S1	S2	Conversion	(sec)	TRE		Change
[28]	MRI	US	3D volume	3D volume	No	20	6.76	2.12	68%
[74]	MRI	TRUS	3D volume	3D volume	No	0.003	8.00	3.63	54%
[3]	RGB	Depth Map	2D Image	2D Image	No	n/a	35.46	6.93	80%
[27]	MRI	CT	3D volume	3D volume	No	320.4	13.49	7.12	47%
[50]	RGB	Point Cloud	2D Image	3D model	Yes	9000	n/a	30.19	n/a
PCD2VOL [67]	CT	Point Cloud	3D volume	3D model	No	0.12	15.34	5.15	62%

We thus opted for a direct comparison of our method against other multimodal registration methods, even though they may deal with different modalities, as this was the nearest we could get to comparing against other methods. Table 4.2 shows quantitative registration results of the latest state-of-the-art 3D multimodal registration methods. Most of these methods align data of different modalities but of the *same structure*. However, the results are only indicative, since each method registers different modalities and the datasets that experiments were conducted on are different and oriented to the specific modalities and task. The table shows the TRE metric as it is considered to be a more generic measure of registration accuracy [41]. In general, TRE is the distance between the corresponding points of the inputs, but due to the fact that the modalities that each method fuses are different, the exact calculation of TRE may differ.

The methods that align different representations of data are [50] and the proposed one, PCD2VOL (Table 4.2). [50] aligns 2D images against a 3D model. However this method converts one modality to the other as a first step (the 2D images to a 3D model) and then executes a typical unimodal registration; the conversion involves the penalties of cost [50] and information loss, as also attested by its high TRE. The proposed method directly registers different data modalities and of different structure, which is a more challenging task compared to registering multimodal data of the same structure.

Interestingly the initial TRE, corresponding to the initial pose of the inputs of the compared methods, varies significantly. The results displayed in Table 4.2 show that the registration error is associated to the difference in initial pose of the inputs. When input modalities start with a pose close to the ideal solution, the initial TRE is lower and so is the registration (final TRE). However, many commonly used registration methods could produce non sufficient results if the modalities are not initialized properly [26].

In an attempt to measure the improvement in alignment of the compared methods, we also calculated the percentage change (PC) in TRE as [34]:

$$\mathrm{PC} = \frac{|TRE - InitTRE|}{InitTRE} 100\% \tag{4.1}$$

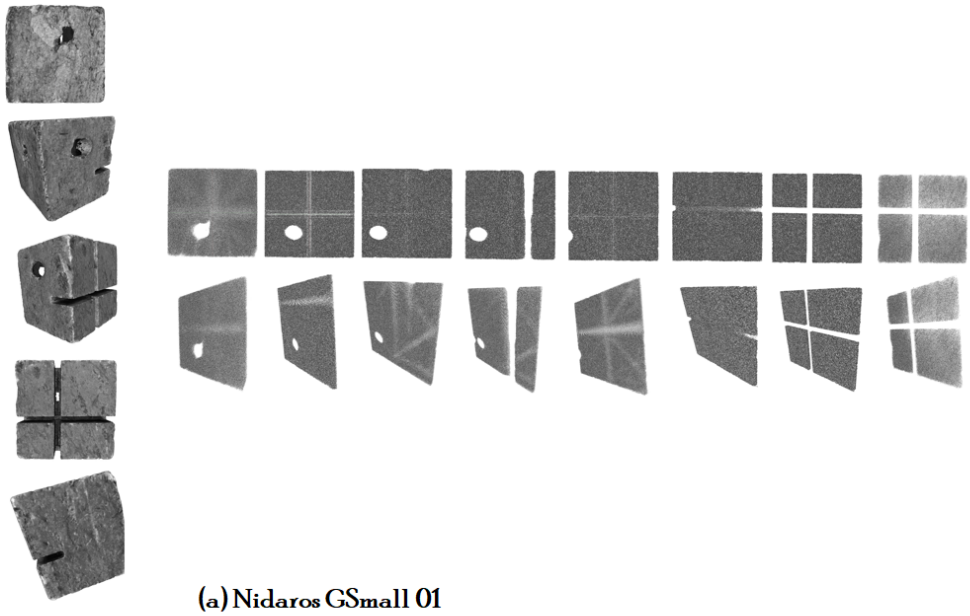
Higher values of PC denote a larger improvement on the initial pose. We chose a high initial TRE for the evaluation of our method in order to mimic real, challenging, situations. Taking into consideration the PC of the proposed method and the fact that it operates on modalities of different data structure, the results obtained can be considered as very competitive.

3D volume modalities (CT, MRI, TRUS) contain details about the inner structure of the object, like cracks, porosity and voids while 3D surface models contain a precise representation of the external surface of the object. A conversion from one modality to the other might result in information loss that will significantly affect the registration result. For example, a 3D model of the surface lacks information on the inner details, so a conversion will not contain any valuable contextual information of the interior and this is likely to affect the registration result. Conversely, a conversion of a 3D volume to a 3D model might add extra computational time without a respective benefit on registration accuracy.

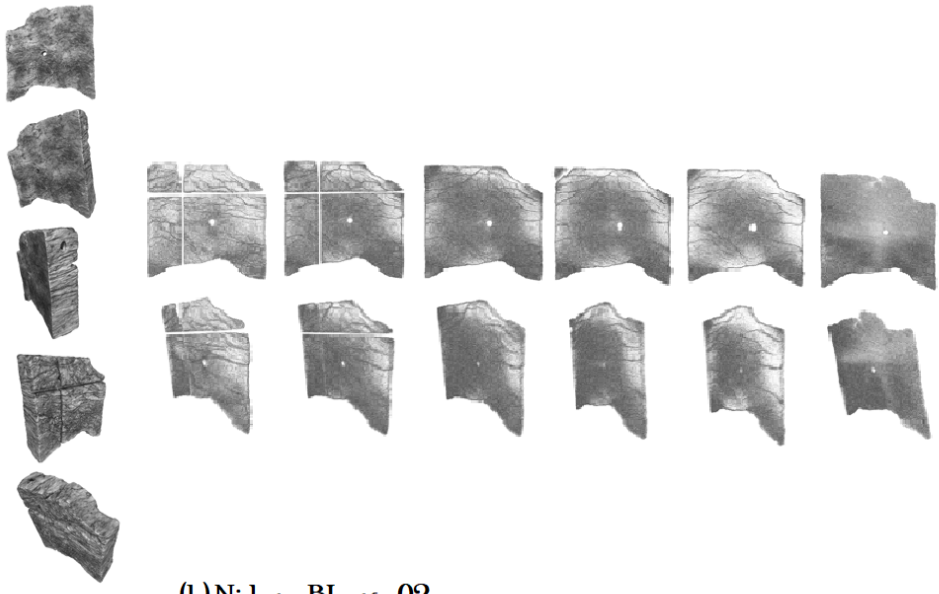
The modified Siamese registration network proposed in Paper D was, to the best of our knowledge, the first registration mechanism that attempted to align two different data modalities not only in terms of data type but data structure as well. In this light, the achieved results can be considered as satisfactory as well as promising.

The 3DPCD-CT dataset

The lack of a specialized dataset for training and benchmarking 3D multimodal registration methods also became evident in the case of *PCD2VOL*. As *PCD2VOL* is a fully supervised deep learning method, it is highly dependent on the availability of sufficient data for training and evaluation. **Paper D** highlights this need and proposes a synthetically generated dataset based on real data from the PRESIOUS project [76, 77]. The *3DPCD-CT* dataset is a multimodal dataset, containing both 3D point clouds and 3D CT volumes along with the ground truth of the best alignment between each pair of modalities. The dataset contains 636 pairs of modalities, divided into a training (80% of the dataset) and testing set (20% of the dataset). Each pair of modalities contains the CT Volume, the respective point cloud and their ground truth transformation for a perfect alignment. Figure 4.4 shows two examples from the *3DPCD-CT* dataset.



(a) Nidaros GSmall 01



(b) Nidaros BLarge 02

Figure 4.4: Example point clouds in the 3DPCD-CT dataset. Two different object cases are shown: a. the Nidaros GSmall 01 stone and b. the Nidaros BLarge 02 stone. The left images depict the 3D geometry of the stone from different viewpoints while the right images represent point clouds of the same slab generated from the respective CT-volume.

Chapter 5

Conclusions and future work

5.1 Future Perspectives

This thesis has identified several areas for future research. **Paper A** proposes CrossTimeReg, a deep learning method for aligning differential data from CH objects before and after possible alterations on their surface. The method focuses on the case of weathering erosion, assuming that the objects have been *uniformly* exposed to environmental effects, both spatially and temporally. The reason for this was the capabilities of the erosion simulator that we had at hand. This work can be extended to use a more realistic erosion model with non-uniform alterations on the surface's orientation, texture and shape.

Another limitation is that even though the CrossTimeReg method is deep learning based, the feature extraction is still hand-crafted. This specific task adds extra computational cost to the entire pipeline. In the future, it may be considered to replace the feature extraction part with features learned specifically for cross-time registration by the network.

While the CrossTimeReg framework was originally developed to register differential data that had been eroded by weathering, it was later integrated into a larger difference detection system as described in **Paper B**. The system efficiently compared similar CH objects, expanding the potential for future applications. Specifically, this framework could be applied to analyzing geometric changes and differences in CH objects for forgery detection and authenticity identification of an object after returning from a loan.

Paper D presents a registration method for 3D point clouds and 3D volumes that treats modalities directly without any prior conversion. However, due to the struc-

tural and physical differences between these modalities, finding an appropriate evaluation metric to validate registration accuracy can be challenging. Furthermore, visualizing both modalities in a clear and informative way is not trivial, and additional research is needed in order to improve the visualization. The framework proposed in **Paper D** consists of an adjustable feature extraction module, which allows generalization to different modalities. By using different feature extraction blocks that are specific to each modality, the method can be extended to fuse different modalities, such as voxel data. The method's adjustability enhances its potential application outside the CH field, as for example in medical imaging.

5.2 Conclusion

This thesis focuses on two specific visual computing problems: cross-time and multimodal registration. Two novel frameworks have been proposed, leveraging deep learning methodologies, along with the respective datasets for training and benchmarking them.

The frameworks proposed in this thesis overcome multiple challenges, such as finding accurate correspondences between modalities that do not share the same characteristics or between 3D models whose geometry has changed over time. The CrossTimeReg method was extended into an automated framework for difference detection and change monitoring for CH objects. Furthermore, both frameworks could be combined and applied for monitoring changes on both the surface and the inner structure of CH objects.

The presented frameworks offer CH experts a valuable tool for tracking changes over time and thus developing effective strategies for conserving and preserving CH objects, while being applicable elsewhere also.

The contributions of this thesis were made publicly available to the research community. They include efficient implementations of the cross-time [66] and multimodal registration [63] frameworks and two datasets for training and benchmarking each framework ([14, 1]). Additionally, all publications related to these contributions were made available open access.

Bibliography

[1] *3DPCD-CT Dataset*. https://doi.org/10.5281/zenodo.7061757. Accessed in March 2023.

[2] Yasuhiro Aoki et al. "Pointnetlk: Robust & efficient point cloud registration using pointnet". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7163–7172.

[3] Moab Arar et al. "Unsupervised multi-modal image registration via geometry preserving image-to-image translation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 13410–13419.

[4] Shahar Barnea and Sagi Filin. "Keypoint based autonomous registration of terrestrial laser point-clouds". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 63.1 (2008), pp. 19–35.

[5] Fausto Bernardini and Holly Rushmeier. "The 3D model acquisition pipeline". In: *Computer graphics forum*. Vol. 21. 2. Wiley Online Library. 2002, pp. 149–172.

[6] Paul J Besl and Neil D McKay. "Method for registration of 3-D shapes". In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. International Society for Optics and Photonics. 1992, pp. 586–606.

[7] Bart Iver van Blokland and Theoharis Theoharis. "Radial intersection count image: A clutter resistant 3D shape descriptor". In: *Computers & Graphics* 91 (2020), pp. 118–128.

[8] Kevin W Bowyer, Kyong Chang and Patrick Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition". In: *Computer vision and image understanding* 101.1 (2006), pp. 1–15.

[9] M Brunetti et al. “Using 3D scanning to monitor wood deformations and to evaluate preservation strategies”. In: *O3A: Optics for Arts, Architecture, and Archaeology*. Vol. 6618. SPIE. 2007, pp. 99–106.

[10] *CHANGE EU project 2019-2023*. https://change-itn.eu/. Accessed in April 2023.

[11] Yang Chen and Gérard Medioni. “Object modelling by registration of multiple range images”. In: *Image and vision computing* 10.3 (1992), pp. 145–155.

[12] Tal Darom and Yosi Keller. “Scale-invariant features for 3D mesh models”. In: *IEEE Transactions on Image Processing* 21.5 (2012), pp. 2758–2769.

[13] Jeffrey Delmerico et al. “The current state and future outlook of rescue robotics”. In: *Journal of Field Robotics* 36.7 (2019), pp. 1171–1191.

[14] *ECHO dataset*. https://doi.org/10.5281/zenodo.5347014. Accessed in March 2023.

[15] Benjamin Eckart, Kihwan Kim and Jan Kautz. “Hgmr: Hierarchical gaussian mixtures for adaptive 3D registration”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 705–721.

[16] Martin A Fischler and Robert C Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[17] Thomas Funkhouser and Michael Kazhdan. “Shape-based retrieval and analysis of 3D models”. In: *ACM SIGGRAPH 2004 Course Notes*. 2004, 16–es.

[18] A Galli and L Bonizzoni. “True versus forged in the cultural heritage materials: the role of PXRF analysis”. In: *X-Ray Spectrometry* 43.1 (2014), pp. 22–28.

[19] Natasha Gelfand et al. “Robust global registration”. In: *Symposium on geometry processing*. Vol. 2. 3. Vienna, Austria. 2005, p. 5.

[20] Corey Goldfeder et al. “Data-driven grasping with partial sensor data”. In: *2009 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2009, pp. 1278–1283.

[21] Leonardo Gomes, Olga Regina Pereira Bellon and Luciano Silva. “3D reconstruction methods for digital preservation of cultural heritage: A survey”. In: *Pattern Recognition Letters* 50 (2014), pp. 3–14.

[22] Robert Gregor et al. “Towards Automated 3D Reconstruction of Defective Cultural Heritage Objects.” In: *GCH*. 2014, pp. 135–144.

[23] Eyal Hameiri and Ilan Shimshoni. “Estimating the principal curvatures and the Darboux frame from real 3-D range data”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 33.4 (2003), pp. 626–637.

[24] Xian-Feng Han et al. “3D Point Cloud Descriptors in Hand-crafted and Deep Learning Age: State-of-the-Art”. In: *arXiv e-prints* (2018), arXiv–1802.

[25] C. Harris and M. Stephens. “A Combined Corner and Edge Detector”. In: *Proceedings of the Alvey Vision Conference*. doi:10.5244/C.2.23. Alvety Vision Club, 1988, pp. 23.1–23.6.

[26] Grant Haskins et al. “Learning deep similarity metric for 3D MR–TRUS image registration”. In: *International journal of computer assisted radiology and surgery* 14.3 (2019), pp. 417–425.

[27] Mattias P Heinrich et al. “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration”. In: *Medical image analysis* 16.7 (2012), pp. 1423–1435.

[28] Mattias Paul Heinrich et al. “Towards realtime multimodal fusion for image-guided interventions using self-similarities”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pp. 187–194.

[29] Michael Hess et al. “Fusion of multimodal three-dimensional data for comprehensive digital documentation of cultural heritage sites”. In: *2015 Digital Heritage*. Vol. 2. IEEE. 2015, pp. 595–602.

[30] Xiaoshui Huang, Guofeng Mei and Jian Zhang. “Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11366–11374.

[31] Xiaoshui Huang et al. “A comprehensive survey on point cloud registration”. In: *arXiv preprint arXiv:2103.02690* (2021).

[32] Bing Jian and Baba C Vemuri. “Robust point set registration using gaussian mixture models”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.8 (2010), pp. 1633–1645.

[33] Andrew E. Johnson and Martial Hebert. “Using spin images for efficient object recognition in cluttered 3D scenes”. In: *IEEE Transactions on pattern analysis and machine intelligence* 21.5 (1999), pp. 433–449.

[34] Lee Kaiser. “Adjusting for baseline: change or percentage change?” In: *Statistics in medicine* 8.10 (1989), pp. 1183–1190.

[35] Laurent Kneip, Zhou Yi and Hongdong Li. “SDICP: Semi-Dense Tracking based on Iterative Closest Points.” In: *BMVC*. 2015, pp. 100–1.

[36] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[37] Baraka Maiseli, Yanfeng Gu and Huijun Gao. “Recent developments and trends in point set registration methods”. In: *Journal of Visual Communication and Image Representation* 46 (2017), pp. 95–106.

[38] Marcello Manfredi et al. “A new quantitative method for the non-invasive documentation of morphological damage in paintings using RTI surface normals”. In: *Sensors* 14.7 (2014), pp. 12271–12284.

[39] D Mannes et al. “Combined neutron and X-ray imaging for non-invasive investigations of cultural heritage objects”. In: *Physics Procedia* 69 (2015), pp. 653–660.

[40] Emilio Marengo et al. “Development of a technique based on multi-spectral imaging for monitoring the conservation of cultural heritage objects”. In: *Analytica chimica acta* 706.2 (2011), pp. 229–237.

[41] Calvin R Maurer et al. “Registration of head volume images using implantable fiducial markers”. In: *IEEE transactions on medical imaging* 16.4 (1997), pp. 447–462.

[42] Pavlos Mavridis, Anthousis Andreadis and Georgios Papaioannou. “Efficient sparse ICP”. In: *Computer Aided Geometric Design* 35 (2015), pp. 16–26.

[43] Pavlos Mavridis, Anthousis Andreadis and Georgios Papaioannou. “Fractured Object Reassembly via Robust Surface Registration.” In: *Eurographics (Short Papers)*. 2015, pp. 21–24.

[44] Francisco PM Oliveira and Joao Manuel RS Tavares. “Medical image registration: a review”. In: *Computer methods in biomechanics and biomedical engineering* 17.2 (2014), pp. 73–93.

[45] Georgios Papaioannou et al. “From reassembly to object completion: A complete systems pipeline”. In: *Journal on Computing and Cultural Heritage (JOCCH)* 10.2 (2017), pp. 1–22.

[46] Georgios Passalis et al. “Using facial symmetry to handle pose variations in real-world 3D face recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.10 (2011), pp. 1938–1951.

[47] Panagiotis Perakis et al. “Simulating Erosion on Cultural Heritage Monuments”. In: *20th International Conference on Cultural Heritage and New Technologies (CHNT20)*. 2015.

[48] Massimiliano Pieraccini, Gabriele Guidi and Carlo Atzeni. “3D digitizing of cultural heritage”. In: *Journal of Cultural Heritage* 2.1 (2001), pp. 63–70.

[49] Roberto Pierdicca et al. “Virtual reconstruction of archaeological heritage using a combination of photogrammetric techniques: Huaca Arco Iris, Chan Chan, Peru”. In: *Digital Applications in Archaeology and Cultural Heritage* 3.3 (2016), pp. 80–90.

[50] Ruggero Pintus and Enrico Gobbetti. “A fast and robust framework for semiautomatic and automatic registration of photographs to 3D geometry”. In: *Journal on Computing and Cultural Heritage (JOCCH)* 7.4 (2015), pp. 1–23.

[51] Ruggero Pintus et al. “A survey of geometric analysis in cultural heritage”. In: *Computer Graphics Forum*. Vol. 35. 1. Wiley Online Library. 2016, pp. 4–31.

[52] Ruggero Pintus et al. “Geometric Analysis in Cultural Heritage.” In: *GCH*. 2014, pp. 117–133.

[53] Denis Pitzalis et al. “3D enhanced model from multiple data sources for the analysis of the Cylinder seal of Ibni-Sharrum”. In: *VAST 2008: The 9th International Symposium on Virtual Reality, Archaeology, and Cultural Heritage*. Eurographics Association. 2008, pp. 79–84.

[54] François Pomerleau, Francis Colas, Roland Siegwart et al. “A review of point cloud registration algorithms for mobile robotics”. In: *Foundations and Trends® in Robotics* 4.1 (2015), pp. 1–104.

[55] *PRESIOUS project*. https://www.ntnu.edu/presious. Accessed in May 2023.

[56] Charles R Qi et al. “Pointnet++: Deep hierarchical feature learning on point sets in a metric space”. In: *arXiv preprint arXiv:1706.02413* (2017).

[57] M Magda Ramos and Fabio Remondino. “Data fusion in cultural heritage-A review”. In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 40.5 (2015), p. 359.

[58] Abtin Rasoulian, Robert Rohling and Purang Abolmaesumi. “Group-wise registration of point sets for statistical shape models”. In: *IEEE transactions on medical imaging* 31.11 (2012), pp. 2025–2034.

[59] Olaf Ronneberger, Philipp Fischer and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[60] Szymon Rusinkiewicz and Marc Levoy. "Efficient variants of the ICP algorithm". In: *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*. IEEE. 2001, pp. 145–152.

[61] Radu Bogdan Rusu, Nico Blodow and Michael Beetz. "Fast point feature histograms (FPFH) for 3D registration". In: *2009 IEEE international conference on robotics and automation*. IEEE. 2009, pp. 3212–3217.

[62] Sunita Saha et al. "Segmentation of change in surface geometry analysis for cultural heritage applications". In: *Sensors* 21.14 (2021), p. 4899.

[63] Evdokia Saiti. *PCD2VOL:Multimodal registration framework for 3D point clouds and CT-volumes*. Aug. 2022. DOI: 10.5281/zenodo.7061757. URL: https://doi.org/10.5281/zenodo.7061757.

[64] Evdokia Saiti, Antonios Danelakis and Theoharis Theoharis. "Cross-time registration of 3D point clouds". In: *Computers & Graphics* 99 (2021), pp. 139–152.

[65] Evdokia Saiti and Theoharis Theoharis. "An application independent review of multimodal 3D registration methods". In: *Computers & Graphics* 91 (2020), pp. 153–178.

[66] Evdokia Saiti and Theoharis Theoharis. *CrossTimeReg_registration_algorithm*. Aug. 2021. DOI: 10.5281/zenodo.5347685. URL: https://doi.org/10.5281/zenodo.5347685.

[67] Evdokia Saiti and Theoharis Theoharis. "Multimodal registration across 3D point clouds and CT-volumes". In: *Computers & Graphics* 106 (2022), pp. 259–266.

[68] Vinit Sarode et al. "PCRNet: Point cloud registration network using PointNet encoding". In: *arXiv preprint arXiv:1908.07906* (2019).

[69] Cinzia Scaggion et al. "3D digital dental models' accuracy for anthropological study: Comparing close-range photogrammetry to μ-CT scanning". In: *Digital Applications in Archaeology and Cultural Heritage* 27 (2022), e00245.

[70] Mohsin M Shanoer and Fanar M Abed. "Evaluate 3D laser point clouds registration for cultural heritage documentation". In: *The Egyptian Journal of Remote Sensing and Space Science* 21.3 (2018), pp. 295–304.

[71] *SHREC 2021: Retrieval of Cultural Heritage Objects*. http://www.ivan-sipiran.com/shrec2021.html. Accessed in March 2021.

[72] Vaibhav Sinha, Ashish V Avachat and Hyoung K Lee. "Design and development of a neutron/X-ray combined computed tomography system at Missouri S&T". In: *Journal of Radioanalytical and Nuclear Chemistry* 296 (2013), pp. 799–806.

[73] R Sitnik et al. "Monitoring surface degradation process by 3D structured light scanning". In: *Optics for Arts, Architecture, and Archaeology VII*. Vol. 11058. SPIE. 2019, pp. 107–116.

[74] Xinrui Song et al. "Cross-Modal Attention for MRI and Ultrasound Volume Registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 66–75.

[75] Aristeidis Sotiras, Christos Davatzikos and Nikos Paragios. "Deformable medical image registration: A survey". In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1153–1190.

[76] Theoharis T. and G. Papaioannou. *PRESIOUS 3D Cultural Heritage Fragments*. 2013. URL: https://www.ntnu.edu/presious/3d-data-sets.

[77] Theoharis T. and G. Papaioannou. *PRESIOUS 3D Cultural Heritage Fragments - Differential Erosion Scans*. 2013. URL: https://www.ntnu.edu/presious/3d-data-sets.

[78] Gary KL Tam et al. "Registration of 3D point clouds and meshes: A survey from rigid to nonrigid". In: *IEEE transactions on visualization and computer graphics* 19.7 (2012), pp. 1199–1217.

[79] Federico Tombari and Fabio Remondino. "Feature-based automatic 3D registration for cultural heritage applications". In: *2013 Digital Heritage International Congress (DigitalHeritage)*. Vol. 1. IEEE. 2013, pp. 55–62.

[80] Yanghai Tsin and Takeo Kanade. "A correlation-based approach to robust point set registration". In: *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III 8*. Springer Berlin Heidelberg. 2004, pp. 558–569.

[81] Mathias Unberath et al. "The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective". In: *Frontiers in Robotics and AI* 8 (2021), p. 716007.

[82] Daniel Vavřík et al. "Analysis of baroque sculpture based on X-ray fluorescence imaging and X-ray computed tomography data fusion". In: *7th Conference on industrial computed tomography, Leuven, Belgium*. 2017, pp. 7–9.

[83] Yue Wang and Justin M Solomon. “Deep closest point: Learning representations for point cloud registration”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 3523–3532.

[84] Yue Wang and Justin M Solomon. “Prnet: Self-supervised learning for partial-to-partial registration”. In: *arXiv preprint arXiv:1910.12240* (2019).

[85] Yiling Wu et al. “Learning fragment self-attention embeddings for image-text matching”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. 2019, pp. 2088–2096.

[86] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[87] Fan Xue et al. “From semantic segmentation to semantic registration: Derivative-Free Optimization–based approach for automatic generation of semantically rich as-built Building Information Models from 3D point clouds”. In: *Journal of Computing in Civil Engineering* 33.4 (2019), p. 04019024.

[88] Jiaolong Yang et al. “Go-ICP: A globally optimal solution to 3D ICP point-set registration”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.11 (2015), pp. 2241–2254.

[89] Zi Jian Yew and Gim Hee Lee. “RPM-net: Robust point matching using learned features”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11824–11833.

[90] Zhou Yu et al. “Rethinking diversified and discriminative proposal generation for visual grounding”. In: *arXiv preprint arXiv:1805.03508* (2018).

[91] Wentao Yuan et al. “DeepGMR: Learning Latent Gaussian Mixture Models for Registration”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 733–750.

[92] Yufu Zang et al. “An Efficient Probabilistic Registration Based on Shape Descriptor for Heritage Field Inspection”. In: *ISPRS International Journal of Geo-Information* 9.12 (2020), p. 759.

[93] Han Zhang et al. “Self-attention generative adversarial networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 7354–7363.

[94] Qian-Yi Zhou, Jaesik Park and Vladlen Koltun. “Fast global registration”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 766–782.

[95] Lu Zou et al. “CMA: Cross-modal attention for 6D object pose estimation”. In: *Computers & Graphics* 97 (2021), pp. 139–147.

Part II

Selected Publications

Chapter 6

Paper A - Cross-time registration of 3D point clouds

Authors

Evdokia Saiti, Antonios Danelakis, and Theoharis Theoharis.

Published in

Computer & Graphics, Volume 99, Pages 139-152, Elsevier, 2021

Computers & Graphics 99 (2021) 139–152

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on 3DOR 2021

Cross-time registration of 3D point clouds

Evdokia Saiti*, Antonios Danelakis, Theoharis Theoharis

NTNU Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Article history:
Received 31 March 2021
Revised 25 June 2021
Accepted 2 July 2021
Available online 9 July 2021

Keywords:
3D registration
Alignment
Cross-time
Retrieval
Cultural heritage
Erosion

ABSTRACT

Registration is a ubiquitous operation in visual computing and constitutes an important pre-processing step for operations such as 3D object reconstruction, retrieval and recognition. Particularly in cultural heritage (CH) applications, registration techniques are essential for the digitization and restoration pipelines. Cross-time registration is a special case where the objects to be registered are instances of the same object after undergoing processes such as erosion or restoration. Traditional registration techniques are inadequate to address this problem with the required high accuracy for detecting minute changes; some are extremely slow. A deep learning registration framework for cross-time registration is proposed which uses the DeepGMR network in combination with a novel down-sampling scheme for cross-time registration. A dataset especially designed for cross-time registration is presented (called ECHO) and an extensive evaluation of state-of-the-art methods is conducted for the challenging case of cross-time registration.

1. Introduction

Geometric registration (or alignment) is a crucial tool in visual computing with applications in robotics, medical imaging and cultural heritage (CH) analysis, among many others. Registration of datasets and particularly point clouds, has become a key operation in many shape analysis tasks, such as 3D object retrieval [1,2], semantic segmentation and classification [3,4], 3D mapping [5–7], 3D object scanning [8] and 3D model reconstruction [9–11].

Registration aims to find the transformation that optimally aligns two or more similar objects or two or more instances of the same object taken at different times (cross-time data), from different viewpoints (multi-view data) or by different sensors (multi-sensor data), in order to bring the data into a common reference frame [12]. The surface alignment problem is a broad research topic and advances have been made over the years, resulting in a plethora of different strategies and algorithms. However, there are still open problems to be addressed, especially in the context of CH. Archaeological objects differ from mechanical or medical objects in their shape and size (some CH objects can be quite large), articulation and fragility. Moreover, the number of objects digitized and available for experimentation is limited in CH.

Computing has greatly aided the CH field over the last decades, including the restoration, preservation and monitoring processes [13]. In monitoring, microgeometric changes over time are measured and analyzed in order to support conservation strategies [14]. CH objects have been constantly undergoing changes or degradation over time. In this matter, geometric acquisition and measurements of a CH object produce snapshots of 3D models and can be used to track an object through time, in order to document different phases of the conservation pipeline and identify any destructive intervention, or to understand any damages that these modifications may indicate. 3D surface registration can automate the process of monitoring CH artefacts in a non-invasive manner by aligning the objects in such a way that even minute modifications on the object's surface or shape can be automatically detected and measured.

As CH digitization is becoming more widespread, CH object monitoring activities based on the digitized objects are increasingly relevant. Several methodologies have been proposed over the last years, but the contribution is limited due to the relatively small number of digitized CH objects than can be used in the experimentation with the monitoring process. The main reasons are that the conservation process is time consuming and needs to be planed properly so as not to harm the CH object and that the change detected from environmental erosion cannot be easily identified unless several decades pass. The lack of an adequate digital benchmark for deeper analysis and comparison is a major obstacle towards the development of automatic techniques for proper monitoring and documenting different phases of conservation. Such a benchmark is crucial for comparing methodologies and scenaria.

This work is focused on the pairwise cross-time registration problem. We introduce a registration methodology that copes with

* Corresponding author.
E-mail address: evdokia.saiti@ntnu.no (E. Saiti).

https://doi.org/10.1016/j.cag.2021.07.005

big data using a down-sampling scheme that is appropriate for objects that undergo erosion over time and overcomes limitations like the computational complexity of iterative methods, the necessity for point-level correspondence or a coarse pre-alignment step. Moreover, we address the absence of benchmarking data by contributing a dataset of artificially eroded CH objects, including their ground truth transformation. The initial models are taken from the SHREC 2021 dataset for retrieval of CH objects [15] and have been artificially eroded based on weathering conditions resulting from polluted environments and from naturally occurring climatic conditions [16].

The contributions of this paper are:

- The problem of cross-time 3D registration is formally defined and a framework for cross-time 3D registration is proposed. Publicly available upon publication.
- A down-sampling methodology that detects the most valuable points for cross-time registration is proposed. Publicly available upon publication.
- A benchmark for evaluating both traditional and cross-time registration algorithms is created. Publicly available upon publication.
- An extensive evaluation of both geometry-based and deep learning state-of-the-art approaches on 3D cross-time registration is performed.

The remainder of this paper is organized as follows: In Section 2 related works are discussed while in Section 3 the problem of cross-time 3D registration is defined. In Section 4 the proposed methodology for cross-time 3D registration is introduced while Section 5 presents the proposed evaluation benchmark. Experimental results on cross-time registration are presented in Section 6. The paper is concluded in Section 7.

2. Related work

Since surface registration is fundamental to many visual computing domains, there is a very extensive literature on the subject. However, to the best of our knowledge, there exists no methodology specifically for cross-time registration. Instead, standard point cloud or surface registration techniques have been used, but the results are sub-optimal as we shall see later. In this section, we review the methods that are most related to cross-time registration. For a comprehensive review of general registration methods, the interested reader is referred to [17] and for a survey oriented to cultural heritage applications to [18].

Registration methods can be roughly classified into two broad categories, local and global. Global registration techniques align the source and target objects without any prior information about their relative pose, whereas in local registration, a prior coarse transformation is known and the algorithm tries to refine the solution. In general, local approaches are more accurate but less robust to initial pose than global approaches. Examples of local approaches are the well-known Iterative Closet Point (ICP) [19] and its variants [20], while RANSAC [21] and Fast Global Registration [22] are examples of global methods. Further, registration approaches can rely on point-to-point correspondences between the data or be correspondence-free [23].

In addition to geometry-based registration techniques, there has been a recent wave of deep learning approaches, attempting to overcome the challenge of prolonged running time and aiming to boost accuracy further [12].

2.1. Geometry-based registration

Correspondence-based methodologies are based on the observation that computing the optimal alignment between two surfaces is equivalent to finding corresponding points and then computing the transformation that best aligns them with respect to minimizing a specific distance function. The Iterative Closest Point (ICP) [19] is the best-known and most applied such algorithm for solving rigid registration problems. ICP iteratively alternates between finding point-to-point correspondences and distance minimization to compute the optimal alignment. Given its popularity, a large number of variants have appeared [20,24] but there are some drawbacks. The method is local and, thus, is effective only when the initial pose of the input geometries is close to the global optimum, otherwise it can converge to a local minimum. Moreover, the iterative nature of the algorithm and its point-to-point correspondence nature result in high computational complexity. In addition, real-world data and particularly in the case of cross-time registration where erosion is involved, do not contain exact point level correspondences.

To overcome the issues of point-to-point matching, many strategies try to identify feature-level similarities and correspondences. Approaches like RANSAC [21] and Fast Global Registration (FGR) [22] use feature descriptors and matching combined with robust fitting or optimization techniques to achieve registration. These techniques are much more efficient than point-level methods but are highly dependent on the quality of features. Feature-based techniques generally involve three steps: feature detection, feature description and correspondence estimation. Features are a small group of interest points that can be detected on both objects, due to their distinctiveness or geometric stability under different transformations. Each feature can be delineated by a descriptor that characterizes its geometric information. Two main categories of descriptors exist: global and local. Global descriptors represent the geometric information of an entire 3D object, whereas local descriptors encode the local information at each feature point [25]. Specifically for 3D registration local descriptors are more commonly used, because they can identify similar localities between the two surfaces to be aligned by exploiting the geometric properties around a certain point and its neighborhood.

A large number of descriptors have been proposed. Diez et al. presented an analytical review in [26], however not every descriptor is suitable for cross-time registration. Some potentially applicable methodologies are next described. Fast Point Feature Histogram (FPFH) [27] consists of pose-invariant features and is generated as a simplified point feature histogram for each key point and its k-nearest neighbors. Johnson and Hebert introduced the Spin Image (SI) descriptor [28], a rigid transformation-invariant 2D characterization of the surface location around a support region of a specific point. This descriptor obtains competitive results in rigid registration, but is vulnerable to symmetries, noise and clutter. The Radial Intersection Count Image (RICI) descriptor [29], a variation of the SI, has been proposed to overcome the limitations of cluttered scenes and is a 2D histogram of integers that represent the number of intersections of circles centered over the point of interest with the 3D surface. Another variant of the SI is the Scale Invariant Spin Image mesh descriptor (SISI) [30], where the SI descriptor is computed over an estimated local scale at each interest point. The same authors also proposed the Local Depth SIFT (LD-SIFT) [30], a rotation and scale invariant descriptor based on the prior work of Lowe [31]. LD-SIFT represents the vicinity of the each interest point as a depth map by computing a local radial-angular histogram of the pixel value derivatives.

Another approach to registration is based on the branch-and-bound framework [32,33] where the low dimensionality (6DoF) is taken as an advantage to exhaustively search the Special Euclidean Group SE(3) space for the optimal alignment. Although, these methods can achieve a good matching regardless of initial conditions, they often have low efficiency. A popular methodology is the use of statistical models for outlier rejection and geometric

alignment. Specific methods include the use of the Expectation-Maximization (EM) [34] principle for finding accurately and efficiently the alignment transformation [35] and the use of Gaussian Mixture Models (GMMs) to reformulate the point-to-point correspondence problem in a lower dimension resulting in a computationally efficient solution, resistant to noise and outliers [36,37].

2.2. Learning-based registration

Significant recent advances of deep learning methodologies on 3D point clouds provide new opportunities for learning point cloud representations. Milestones like PointNet [38] and DGCNN [3] offer structured representations of 3D point clouds and even if originally designed for point cloud classification and segmentation, they have be transformed and applied to point cloud registration. Learning-based registration has recently shown robustness and efficiency gains over geometry-based techniques.

PointNetLK [39] integrates the Lucas & Kanade (LK) algorithm [40] with the PointNet network for aligning the global features produced by the latter. PointNetLK performs well on shapes unseen in training, but is not robust to noise. PCRNet [41], like PointNetLK, uses PointNet to encode the shape information of the input point clouds but replaces the Lucas-Kanade step by a deep network. DCP [42] is a non-iterative, one-shot algorithm that uses a Siamese DGCNN [3] network to extract the learned correspondences and a differentiable SVD method for registration. RPM-Net [43] tries to improve the robustness to partial visibility by inheriting the idea of the RPM algorithm [44] and incorporating it in a deep network. DeepGMR [45] integrates Gaussian Mixture Model (GMM) registration [36] with neural networks by extracting pose-invariant correspondences between raw point clouds and GMM parameters. Then, these correspondences are fed into the GMM optimization module to estimate the transformation matrix in a single step. The method is efficient and robust to arbitrary displacements and noise. Although, DeepGMR shows highly accurate results, it estimates the correspondence between all points and all components in the latent GMM, which is not suitable for real-life applications and especially in the case of 3D objects that are changing over time.

2.3. Partial registration

A more challenging sub-problem of 3D registration is partial registration, where only subsets of the source and the target object match to one another. Having partially overlapping areas, the alignment is performed by registering the mutually shared patches. Several methods attempt to find correspondences in the area of overlap by identifying keypoints that are common in both source and target. Super4PCS [46], is a variant of RANSAC which iteratively aligns congruent sets of four points taken from the source and the target object. The number of iteration is adaptive, so that when the partial overlap is low, more iterations are performed to reach an acceptable registration result, regardless of initial pose and overlap percentage. Other methods are variants of ICP that deal with noisy data and partial overlap by using general optimization algorithms, like Simulating Annealing [24] and Particle Swarm Optimization [47]. More recently, partial registration has been addressed by PRNet [48], which follows an iterative refinement strategy. It uses deep networks to detect the points of interest followed by estimating the correspondences iteratively in a coarse-to-fine manner to perform the final registration.

Cross-time registration and partial registration share a lot of characteristics. However, there is a crucial difference: in partial registration it is assumed that where overlaps exist, the shape has not changed, while in cross-time registration the objects may encounter considerable shape differences throughout their surface.

3. Problem statement

3.1. 3D registration

In 3D registration we are given two 3D point clouds, the source $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ and the target $\mathbf{Q} = \{q_j \in \mathbb{R}^3 |\ j = 1, 2, \ldots, M\}$ and the objective is to recover the unknown rigid transformation $\mathbf{T}$ so as to match the source $\mathbf{P}$ into the target point cloud $\mathbf{Q}$.

A rigid transformation in 3D can be represented by a transformation matrix $\mathbf{T}$ which consists of two components; a rotation submatrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. The rigid transformation $\mathbf{T}$ can then be represented by the following homogeneous 4×4 matrix:

$$\mathbf{T} = \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline 0 & 1 \end{array}\right] \tag{1}$$

where $\mathbf{T} \in SE(3), \mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$. $SE(3)$ is the special Euclidean group of rigid transformations in 3D space (rotations and translations), while $SO(3)$ is the special orthogonal group of rotations in Euclidean Space $\mathbb{R}^3$.

The problem of rigid registration between two discrete point clouds can be formulated as [49]:

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} d(\mathbf{R}\mathbf{p_i} + \mathbf{t}, \mathbf{Q}) \tag{2}$$

where function $d(\mathbf{p}, \mathbf{Q})$ measures the distance of an arbitrary point $\mathbf{p} \in \mathbf{P}$ to the point cloud $\mathbf{Q}$ and can be defined as:

$$d(\mathbf{p}, \mathbf{Q}) = \min_{\mathbf{q} \in \mathbf{Q}} d(p, q) \tag{3}$$

where $d(p, q)$ is the distance between two points in space.

Eq. (3) is referred to as the distance or error metric. Many methods [24,32] use the squared Euclidean norm as the distance metric and optimize Eq. (2) using least squares:

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \|\mathbf{R}\mathbf{p_i} + \mathbf{t} - \widehat{\mathbf{q}}_\mathbf{i}\|^2 \tag{4}$$

where $\widehat{\mathbf{q}}_\mathbf{i}$ is the closest point in $\mathbf{Q}$ to each point $\mathbf{p_i} \in \mathbf{P}$ based on the transformation $\mathbf{T}(\mathbf{R}, \mathbf{t})$.

3.2. The cross-time 3D registration problem

Methods that monitor the geometric variation of an object over time, must try to compare the 3D representations of the same object captured at different points in time. During these time intervals, several modifications like degradation from environmental erosion, cleaning and conservation actions, or even cracking may have occurred on the surface of the object. Therefore, it is not expected that the acquisition process will start at the exact same position at both times; thus the 3D point clouds will not have the same number of points and no perfect correspondences.

Various decay phenomena and alteration processes may occur to the surface of a CH object. Alterations can be due to weathering conditions, physical or chemical aging or human intervention [50]. The material alteration processes can cause local loss of the surface (bursting, chipping, peeling), change in shape (deformation, blistering, delamination, exfoliation, crumbling), cracks (splitting, hair cracks, star cracks) or changes in texture (discoloration, bleaching, staining). Moreover, any conservation process can be considered as an alteration operation to the object, even though it does not imply a worsening of its characteristics and shape (e.g. application of reversible coating, varnish removal or mechanical and chemical cleaning).

Let us define the initial CH object as a set of 3D points $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ and the altered object as $\widehat{\mathbf{P}} = \{\widehat{p_j} \in \mathbb{R}^3 |\ j =$

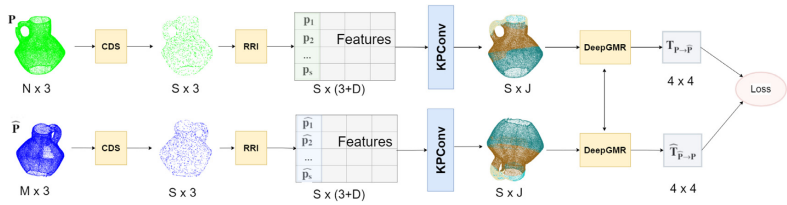


Fig. 1. Overview of the proposed *CrossTimeReg* cross-time 3D registration pipeline.

$1, 2, \ldots, M\}$. Without loss of generality, one can assume the existence of a change function f_{ch} that describes the modifications that the initial object has undergone, so that $\widehat{\mathbf{P}} = f_{ch}(\mathbf{P})$. f_{ch} may encompass various types of alterations.

In this framework, the 3D cross-time registration problem can be formulated as: given two 3D point clouds of the same object but captured at different time frames, the source $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ and the target $\widehat{\mathbf{P}} = \{\widehat{p}_j \in \mathbb{R}^3 |\ j = 1, 2, \ldots, M\}$ with $N \neq M$, the aim is to find the unknown rigid transformation $\mathbf{T}$ so as to align the source $\mathbf{P}$ onto the target $\widehat{\mathbf{P}}$ as well as possible for a specific distance metric d:

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} d(\mathbf{R}\mathbf{p_i} + \mathbf{t}, \widehat{\mathbf{P}}) \tag{5}$$

The problem of cross-time registration can be really challenging if all the different aspects of alterations that a CH object may experience are taken into account. In this work, we focus on the simplified but still challenging case of weathering erosion, where we assume that the objects have been uniformly exposed to environmental effects, both spatially and temporally. We are motivated from the observation of [51] that a typical registration algorithm like ICP [19], will align the source and target point clouds $\mathbf{P}$ and $\widehat{\mathbf{P}}$, so as to minimise the error (i.e. RMSD, Chamfer distance) between them. In doing this, the registration process will often bring the two point clouds close together in certain areas, most probably where the sampling density is higher. This is not ideal where objects have undergone uniform erosion across their surface as shown in the experiments of [52].

This problem has a number of interesting characteristics, especially when considered in the Cultural Heritage domain where the cross-time nature of $\mathbf{P}$ and $\widehat{\mathbf{P}}$ arises after erosion over a long time period:

- A classic registration algorithm will weigh more areas with dense sampling as more points are contributing to the error metric; however an erosion process is more likely to affect the surface of the object evenly and thus a resampling process is required.
- As $\mathbf{P}$ and $\widehat{\mathbf{P}}$ can be assumed to be the same object, we know that there exists an ideal registration $(\mathbf{R}, \mathbf{t})_{\mathbf{ideal}}$. However, as object scans are likely to have been taken across several years, probably with different scanner technology and without external reference points, $(\mathbf{R}, \mathbf{t})_{\mathbf{ideal}}$ is not known. In Section 5 we have created a synthetically eroded dataset where $(\mathbf{R}, \mathbf{t})_{\mathbf{ideal}}$ is known by definition and can be used for training and benchmarking cross-time registration algorithms.

4. Method overview

In this Section, the *CrossTimeReg* framework is presented, see Fig. 1 for the pipeline. The initial and eroded point clouds (also referred to as *source* and *target*) are denoted by $\mathbf{P}$ and $\widehat{\mathbf{P}}$ respectively. $\mathbf{P}$ and $\widehat{\mathbf{P}}$ are first down-sampled using the Curvature Down-Sampling (CDS) block and then rotation invariant features are computed by the Feature Extraction block (RRI). The features along with the point clouds are then sent to a Siamese architecture of KPConv networks. KPConv network is a segmentation network, which estimates for each point the component that it belongs to; it thus determines a point-to-component correspondence. Finally, the registration is performed by aligning the component centroids (weighted by the covariances) using the DeepGMR module, a weighted version of the SVD solution proposed in [42].

Curvature down-sampling (CDS): Registration algorithms often use a down-sampling pre-processing step on the input point clouds to accelerate the registration process. Some methods [30] detect the most interesting points and compute a descriptor for each of them while others [45,48] keep the nearest or farthest S points to the centroid of the object. In traditional registration, these methods may be sufficient as the local shape of the source and target object is not expected to vary. However in cross-time registration, the target object's local shape is expected to be modified due to erosion and other effects and the aforementioned down-sampling approaches may fail. To address this, we propose a down-sampling approach for cross-time registration that takes into consideration the points that are less likely to be significantly altered by erosion. We expect these points to be those with the minimum principal curvature [53,54]. The intuitive reason behind this is that such points are less exposed to erosion/degradation processes or conservation activities. Thus they are considered to be a robust representation of the object across such processes or activities [55]. We thus compute the principal curvature of each point of $\mathbf{P}$ and down-sample by retaining the S points with the minimum principal curvature values. We have selected $S = 1024$ (see Section 6).

To compute the principal curvature λ_i of a point $p_i \in \mathbf{P}$, the neighborhood covariance matrix $\mathbf{C_i}$ is first computed and then Eq. (6) is resolved with respect to scalar λ_i (eigenvalue of $\mathbf{C_i}$) and matrix $\mathbf{u}$ (eigenvectors of $\mathbf{C_i}$) [56]:

$$\mathbf{C_i}\mathbf{u} = \lambda_i \mathbf{u} \tag{6}$$

The symmetric 3×3 covariance matrix $\mathbf{C_i}$ of a point p_i is calculated based on its local neighborhood of κ nearest points q_j, $j = 1, 2, \ldots, \kappa$:

$$\mathbf{C_i} = \frac{1}{\kappa} \sum_{j=1}^{\kappa} \begin{bmatrix} q_j^x q_j^x & q_j^x q_j^y & q_j^x q_j^z \\ q_j^y q_j^x & q_j^y q_j^y & q_j^y q_j^z \\ q_j^z q_j^x & q_j^z q_j^y & q_j^z q_j^z \end{bmatrix} \tag{7}$$

where q_j^x, q_j^y, q_j^z correspond to the x,y and z coordinates of neighboring point q_j respectively.

Eigenvectors $\mathbf{u}$ represent the principal axes of the neighborhood:

$$\mathbf{u} = \begin{bmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{bmatrix} \tag{8}$$

and their eigenvalues λ are:

$$\lambda = \begin{bmatrix} \lambda_A & 0 & 0 \\ 0 & \lambda_B & 0 \\ 0 & 0 & \lambda_C \end{bmatrix} \tag{9}$$

Then the principal curvature λ_i of p_i is:

$$\lambda_i = \frac{min(\lambda_A, \lambda_B, \lambda_C)}{\lambda_A + \lambda_B + \lambda_C} \tag{10}$$

Feature extraction (RRI): We adopt the RRI (rigorous rotation invariant) descriptors for the point cloud [57] which creates features that remain fixed under different orientations. For each point $p_i \in \mathbf{P}$, the RRI module searches for its K-nearest neighbors and constructs a $K-NN$ graph. Then a combination of distance, angle, sin and cos features are computed for p_i based on the local neighborhood of the $K-NN$ graph.

Thus, the outcome of the RRI module is a feature matrix $\mathbf{F} = \{f_i \in \mathbb{R}^D |\ i = 1, 2, \ldots, S\}$ of dimension $S \times D$, where $D = 4 * K$ (K neighbors with 4 features each). The features $\mathbf{F}$ are then combined with the points $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, S\}$ that resulted from down-sampling and the concatenated matrix of dimension $S \times (3 + D)$ is output to the next stage.

Model segmentation (KPConv): We next estimate point-to-component correspondences, by segmenting each point cloud with the KPConv network [58]. We chose the deformable KPConv (KPFCNN) presented in the same work, as our segmentation backbone for its good performance in learning local shifts effectively by deforming the convolution kernels to make them fit into the point cloud.

Given the down-sampled point cloud $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, S\}$ and its D corresponding features at each point $\mathbf{F} = \{f_i \in \mathbb{R}^D |\ i = 1, 2, \ldots, S\}$, the convolution of a kernel g at a point $x \in \mathbb{R}^3$ is defined as:

$$g(x, \mathbf{P}, \mathbf{F}) = \sum_{x_i \in N_x} g(x_i - x) f_i \tag{11}$$

where $N_x = \{x_i \in \mathbf{P} |\ \|x_i - x\| \leqslant r \in \mathbb{R}\}$, is the radius neighborhood of point x [59]. This neighborhood creates a sphere $\mathcal{S}_r^3$ around the point of interest x, and K kernels are spread in this sphere. Let $\{\tilde{x}_k |\ k = 1, .., K\} \subset \mathcal{S}_r^3$ be the kernel points and $\{W_k |\ k = 1, .., K\}$ be their associated weight matrices; then the kernel g can be defined in association with the linear correlation h between the kernel points $\tilde{x}_k$ and any point $(x_i - x)$ of sphere $\mathcal{S}_r^3$, as:

$$g(x_i - x) = \sum_{k=1}^{K} h(x_i - x, \tilde{x}_k)\ W_k \tag{12}$$

where

$$h(x_i - x, \tilde{x}_k) = \max\left(0, 1 - \frac{\|(x_i - x) - \tilde{x}_k\|}{\sigma}\right) \tag{13}$$

and σ is the influence distance between the kernel point and the selected point of the sphere that is related on the input density.

Combining equations Eqs. (12) and (11) we get the standard KPConv layer:

$$g(x, \mathbf{P}, \mathbf{F}) = \sum_{x_i \in N_x} \left(\sum_{k=1}^{K} h(x_i - x, \tilde{x}_k)\ W_k \right) f_i \tag{14}$$

Even though the standard KPConv produces sufficiently good results, we concluded that the deformable KPConv [58] suits the cross-time registration even better, because the network learns the kernel point positions. Instead of defining the kernel g on the kernel points $\tilde{x}_k$, the network generates a set of K shifts $\Delta(x)$ for every point $x \in \mathbb{R}^3$. Then the deformable KPConv layer is defined as:

$$g(x, \mathbf{P}, \mathbf{F}) = \sum_{x_i \in N_x} \left(\sum_{k=1}^{K} h(x_i - x, \tilde{x}_k + \Delta(x))\ W_k \right) f_i \tag{15}$$

The KPConv module estimates the point-to-component correspondences of both source and target point clouds, essentially performing a segmentation. The registration is done by the GMM-based DeepGMR module, which learns a consistent GMM representation of J components in order to recover the optimal transformation between the segmented point clouds. Given the desired number of segmentation components J, KPConv produces a respective segmentation of the input points in the form of an $S \times J$ association matrix $\mathbf{\Gamma} = \{\gamma_{ij}\}$ whose elements represent the probability of a point p_i belonging to a component $j \in J$. These J components are used to express the point cloud as a Gaussian Mixture Model (GMM) of J Gaussian distributions.

Final alignment (DeepGMR): The association matrix $\mathbf{\Gamma}$, representing the point-to-component correspondence, is used to estimate the transformation matrix $\mathbf{T}$ that aligns $\mathbf{P}$ and $\tilde{\mathbf{P}}$. To this end, we employ the optimization module of the DeepGMR network [45], where two differentiable blocks M_Θ and M_T are used to calculate the Gaussian mixture model (GMM) parameters from the association matrix $\mathbf{\Gamma}$ and transformation matrix $\mathbf{T}$ respectively.

M_Θ block converts the given point cloud $\mathbf{P} = \{p_i \mid i = 1, .., S\}$ and its association matrix $\mathbf{\Gamma} = \{\gamma_{ij} \mid i = 1, \ldots, S\ \&\ j = 1, \ldots, J\}$ to GMM parameters Θ as:

$$\Theta_j = (\pi_j, \mu_j, \Sigma_j) \tag{16}$$

where: $\pi_j = \frac{1}{S} \sum_{i=1}^{S} \gamma_{ij}$ is a scalar mixture weight, μ_j is the mean vector and $\mathbf{\Sigma}$ the covariance matrix of the j-th component, computed by solving the equations:

$$S\pi_j \mu_j = \sum_{i=1}^{S} \gamma_{ij} p_i \tag{17}$$

$$S\pi_j \Sigma_j = \sum_{i=1}^{S} \gamma_{ij} (p_i - \mu_j)(p_i - \mu_j)^\top \tag{18}$$

Finally, the transformation matrix $\mathbf{T}^* = (\mathbf{R}, \mathbf{t})$ is computed by block M_T, which tries to minimize the KL-divergence between the transformed GMM parameters Θ of the source and the GMMs $\hat{\Theta}$ of the target:

$$\mathbf{T}^* = \arg\min_T KL(T(\hat{\Theta}) |\ \Theta) = \arg\min_T \sum_{j=1}^{J} \frac{\hat{\pi}_j}{\sigma_j^2} \|T(\hat{\mu}_j) - \mu_j\|^2 \tag{19}$$

where $\Sigma_j = diag([\sigma_j^2, \sigma_j^2, \sigma_j^2])$ due to the fact that the covariances are chosen to be isotropic. This computes the alignment of the components' centroids instead of the alignment of the point clouds themselves.

The loss function of the DeepGMR module is back-propagated to the KPConv module in order to fine-tune its parameters with respect to the segmentation into the desired J components.

Loss function: The training objective of the loss function is to minimize the registration error. Many previous methods try to minimize the actual distance between the corresponding points in source and target point clouds [41,60], but in the case of cross-time registration this may not be ideal. We employ the directed Hausdorff distance, which has been proposed before [51] as a suitable metric for erosion. Given the ground truth transformation $\mathbf{T_{ideal}} = (\mathbf{R}, \mathbf{t})_{ideal}$ that aligns the source $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ to the target $\tilde{\mathbf{P}} = \{\tilde{p}_j \in \mathbb{R}^3 |\ j = 1, 2, \ldots, M\}$ and the predicted transformation $\mathbf{T} = (\mathbf{R}, \mathbf{t})$ that CrossTimeReg estimates, the loss function that we aim to minimize is:

$$L = \sqrt{D_H + D_{MH}} \tag{20}$$

where D_H is the standard Hausdorff distance calculated as the maximum of the directed Hausdorff distances D_h,where $D_h =$

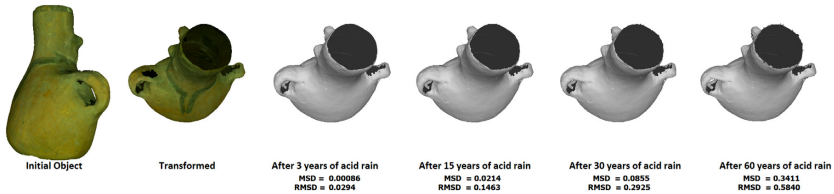


Fig. 2. The steps of the ECHO dataset creation for one object. The object is initially transformed and then the erosion simulator runs for 20 epochs of 3 years each. In this example, the initial model is shown degraded due to the effect of acid rain after 3, 15, 30 and 60 years. Below each step the point-wise MSD (Eq. (26)) and RMSD from the transformed model are given.

$\max_i(\min_j \|p_i - \widehat{p}_j\|)$:

$$D_H = \max\left(D_h(\mathbf{P}, \widehat{\mathbf{P}}) \;\; D_h(\widehat{\mathbf{P}}, \mathbf{P})\right) \tag{21}$$

and D_{MH} is the average directed Hausdorff distance:

$$D_{MH} = \frac{1}{N}\sum_{i=1}^{N} \min_j(\|p_i - \widehat{p}_j\|) \tag{22}$$

The average directed Hausdorff distance denotes the mean value of the minimum Euclidean distances $\|p_i - \widehat{p}_j\|$ between the initial source point cloud and the eroded target point cloud.

5. ECHO: a dataset of Eroded Cultural Heritage Objects

To the best of our knowledge, there is no publicly available dataset with ground truth for cross-time 3D registration. In order to benchmark and train cross-time 3D registration algorithms, we propose the ECHO dataset. Starting from a publicly available dataset of CH objects (see Section 5.1) we first applied a random rotation and translation $(\mathbf{R}, \mathbf{t})$ to the objects; then we used an artificial erosion process to erode the transformed objects. Since erosion is done in situ and the $(\mathbf{R}, \mathbf{t})$ parameters are known, we have the ground truth for benchmarking cross-time registration algorithms. The process is outlined in Fig. 2.

The ECHO dataset consists of three main parts, the original dataset, the transformed objects and the transformed-eroded objects. All three parts of the dataset along with the steps performed are explained thoroughly in the following subsections.

ECHO will be made publicly available with this paper.

5.1. Initial CH dataset

As a cornerstone, we selected the freely available *SHREC 2021: Retrieval of Cultural Heritage Objects* dataset [15] hereafter called SHREC2021. SHREC2021 dataset consists of 1575 3D scans of CH objects from pre-Columbian cultures captured in the Josefina Ramos de Cox museum in Lima, Peru. The SHREC2021 dataset is separated into two sub-datasets, considering two aspects, the shape and the culture. Each of the datasets is also divided into a collection set (70% of the dataset) and a query set (30% of the dataset) that can be used for training and testing respectively. The dataset regarding shape (referred as *datasetShape*) consists of 938 objects, 661 objects for training and 277 for testing. The other dataset is related to the retrieval-by-culture challenge of the SHREC competition, thus we will refer to it as *datasetCulture*. This dataset consists of 637 objects, 448 objects for training and 189 for testing. The objects of both sub-datasets are 3D meshes in.OBJ format, each consists of nearly 40,000 triangles, and they have been pre-processed so as to be centered in the origin of 3D space and with the up direction along the Y-axis. Figs. 3 and 4 show examples from SHREC2021.

Fig. 3. Original CH objects from SHREC2021 datasetShape.

Fig. 4. Original CH objects from SHREC2021 datasetCulture.

5.2. Building the ECHO dataset

Random transformation As a first step, we generated a variation of the initial dataset by applying a randomly calculated rigid 3D transformation; each object of the SHREC2021 dataset has been randomly rotated and translated. The rotation parameters were unrestricted while the translation vector was restricted to a maximum limit of 30 cm. The latter was decided based on the size of the objects. Fig. 5 shows examples of the initial objects along with their transformed instances. This dataset can of course be used as is for evaluating regular 3D registration algorithms. However, we extended it as per the next Section, in order to assess cross-time registration algorithms specifically.

Introduction of erosion Erosion due to atmospheric factors can affect the physiology of the object, resulting in alteration of its small-scale features that can challenge the registration process. We extended the aforementioned dataset by providing an eroded

Fig. 5. Original CH objects from SHREC2021 datasetShape (left), along with their transformed instances (middle). On the (right), a combination of the original and the transformed object is shown in order to demonstrate the translation value.

dataset of the transformed objects. The eroded set represents the erosion/degradation phenomenon that an object faces when exposed to the outdoor environment. Without loss of generality, we focus on chemical weathering of carbonate stone, i.e. the process that carbonate stone objects undergo when exposed to weather and especially to common atmospheric chemicals, such as carbon dioxide (CO_2) and nitrogen dioxide (NO_2).

The exact physico-chemical composition of the original objects of our source CH dataset is not known. Since our aim is the training and benchmarking of cross-time registration algorithms, rather than the simulation of realistic erosion for the objects' specific material, we have assumed that they consist of carbonate stone (a common material for CH objects, e.g. around the Mediterranean region) and applied weathering models that were available to us for this material (see below). Note that weathering models for other materials are not commonly available and, to the best of our knowledge, no other large publicly available dataset of eroding CH objects exists, for any type of material, that is suitable for training and testing deep networks (i.e. contains ground truth).

To this end, we adapted the simulation algorithm for the erosion of carbonate stone and marble presented in [51]. The simulator estimates the degradation of homogeneous marble or carbonate stone objects after their uniform exposure (spatially and temporally) to environmental conditions of polluted areas. We used the cases of chemical weathering in polluted atmosphere regions and the interaction of sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon dioxide (CO_2) with the material of the object. Specifically, we considered the effects of dry deposition of crust due to pollution and the recession by acid rain, which can result in gain or loss of material on the surface of an object. Dry deposition indicates the reaction of the material with SO_2 and NO_2 and manifests itself by the creation of crust upon the object's surface due to the transfer of chemical compounds from polluted industrial environments [61]. Recession by acid rain is the effect of surface loss of the object mainly due to its reaction with water and SO_2, NO_2 and CO_2 [62].

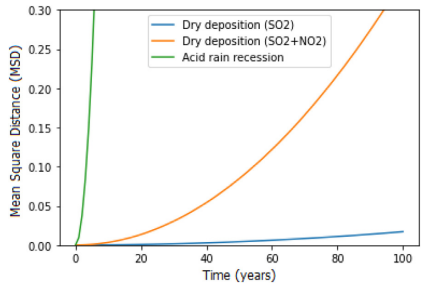


Fig. 6. Point-wise MSD between the initial and eroded carbonate stone objects over a period of 0–100 years, for different weathering cases.

These weathering processes describe the change of the surface geometry and can be formulated as follows. Assuming that the initial object is modelled as a set of 3D points $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ and $\mathbf{n} = \{n_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ are the normals per 3D point, the deposition/recession process relies on a computational and chemical model. The model can be formulated as a uniform offsetting procedure based on the diffusion equation:

$$\frac{\partial \mathbf{P}}{\partial t} = \mu \nabla^2 \mathbf{P} = \delta\ \mathbf{n} \tag{23}$$

so the target eroded surface $\widehat{\mathbf{P}} = \{\widehat{p_j} \in \mathbb{R}^3 |\ j = 1, 2, \ldots, N\}$ is then calculated as

$$\widehat{\mathbf{P}} = \mathbf{P} + \delta\ \mathbf{n}\ dt \tag{24}$$

which becomes:

$$\widehat{p_j} = p_i + \delta_i\ n_i\ dt \tag{25}$$

where n_i is the normal vector of point p_i, δ_i is the surface alteration at this point as predicted by the erosion model and dt is the time interval that the change is occurred. The above computation can be repeated for a number of epochs. Each epoch consists of time intervals of dt, where the environmental conditions are simulated. At the end of each epoch a new eroded surface is produced. The final surface produced after the total number of epochs reflects the changes that the initial surface faced when exposed to weathering conditions. If $\delta_i > 0$, the process simulates the surface recession due to dry deposition and when $\delta_i < 0$ it simulates the recession due to acid rain at a specific point i. The surface alteration offset δ derives from modeling the chemical processes according to the weathering models, described in [51,61–64].

In order to quantify the degradation that the CH objects experience under the above chemical models, we computed the Mean Square Distance (MSD) between each initial object and its eroded counterpart over a period of 0–100 years. Let $\mathbf{P_0} = \{p_{0i} \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ be the original transformed point cloud and $\widehat{\mathbf{P_e}} = \{\widehat{p_{ej}} \in \mathbb{R}^3 |\ j = 1, 2, \ldots, M\}$ be the eroded point cloud after e years of erosion ($\widehat{\mathbf{P_e}}$ has level of erosion = e); then the MSD at level e is calculated as:

$$\text{MSD} = \frac{1}{N} \sum_{i=1}^{N} \| p_{0i} - \widehat{p_{ej}} \|^2 \tag{26}$$

where $\widehat{p_{ej}}$ is the nearest neighbor of p_{0i}.

Fig. 6 shows how considerably high is the degradation due to acid rain, compared to the respective degradation due to crust.

Fig. 7. Original CH objects from SHREC2021 datasetShape (left), along with their transformed instances (middle). The transformed object after application of acid rain erosion simulation is shown on the (right).

The erosion simulation is performed on each transformed object, for the time interval of 60 years, divided into 20 epochs (of 3 years each). During the simulation, the object's rigid parameters do not change, so we can argue that the ground truth random transformation matrix still holds.

Fig. 7 shows examples of the three main steps of the creation of ECHO dataset, the original object, the transformed one and the eroded instance.

6. Experiments

Our experiments are divided into five parts. First, we evaluate the proposed registration algorithm against the relevant state-of-the-art methods using the proposed ECHO dataset for the challenging problem of cross-time registration. Second, we compare the methods across multiple levels of erosion on ECHO. Third, we evaluate them on the task of traditional registration on two datasets; the ECHO dataset (containing only random rigid transformations, no erosion) and the SHREC2016 dataset [65]. Fourth, we evaluate CrossTimeReg on real erosion data, by performing cross-time registration on data from the PRESIOUS project [16]; these data are derived from erosion accelerators that simulate acid rain, salt and freeze-thaw effects on marble and soapstone slabs. Fifth, we attempt to measure the contribution of the proposed curvature downsampling on the cross-time registration task.

We compare against both geometry-based and deep-learning, local and global, as well as correspondence-based and non-correspondence-based algorithms. Regarding geometry-based methods, we compare against ICP [19], RANSAC [21] and Fast Global Registration (FGR) [22]. For RANSAC and FGR, we evaluated several variants with different feature extraction methods: FPFH [27], Spin Images (SI) [28], SISI, LD-SIFT [30], and RICI [29]. We tested each feature descriptor, as a pre-process step for both RANSAC and FGR algorithms. However, we kept and present the combinations that gave the best result in the Recall_α metric. Regarding deep-learning methods, we evaluated PRNet [48], PointNetLK [39], PCRNet [41], RPM-Net [43], DCP [42] and DeepGMR [45]. For ICP, RANSAC, FGR and FPFH we used the python implementations from the Open3D library [66], while for SISI, LD-SIFT and RICI we adapted the code bases released by the authors, which were implemented in MATLAB and C++. For the deep-learning methods, we used the pre-trained models provided by the open-source library Learning3D [67]. To ensure a fair comparison, all deep learning methods (including the proposed one) have been trained on the ModelNet dataset [68]. We have trained the complete CrossTimeReg pipeline using the ModelNet dataset with the annotated data provided in [45]. The first 20 classes of ModelNet have been used, as only those have been annotated by the authors (the rest were used for testing). Random translations and rotations are generated on the fly during the training/validation process for each annotated input point cloud of the ModelNet dataset. Based on the ablation study [45] regarding the ideal number J of Gaussian distributions, we use $J = 16$ for all experiments.

All tests were run on a PC with an i7-7700K CPU at 4.20 GHz, NVIDIA GeForce GTX 1080 Ti GPU and 32 GB of RAM.

6.1. Evaluation metrics

The rotation and translation errors are the absolute errors in Euler angles and translation vectors with respect to the ground truth. Ideally, both should be zero:

If $\mathbf{T}_{GT} = (\mathbf{R}_{GT}, \mathbf{t}_{GT})$ and $\mathbf{T}_{pred} = (\mathbf{R}_{pred}, \mathbf{t}_{pred})$ are the ground truth and predicted transformations respectively, the rotation and translation errors are measured as:

$$Error(R) = \|\mathbf{I} - \mathbf{R}_{GT}^{-1}\ \mathbf{R}_{pred}\| \tag{27}$$

$$Error(t) = \|\mathbf{t}_{GT} - \mathbf{t}_{pred}\| \tag{28}$$

where $\mathbf{I}$ is the identity matrix.

We next measure the *root mean square error (RMSE)* in Euclidean space against the ground truth solution. For the case of cross time registration, it is not sufficient to consider the registration error between the transformed source and the target, since there may not exist exact correspondence between them. Further, the commonly eroded surfaces of the objects may erroneously be measured as registration error, even though they represent the actual degradation of material. We thus measure the effect of the predicted transformation $\mathbf{T}_{pred}$ against the ground truth transformation $\mathbf{T}_{GT}$ on the source object based on [69]:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{T}_{pred}\, p_i - \mathbf{T}_{GT}\, p_i)\|^2} \tag{29}$$

where N is the total points of the source object.

Moreover, following [12,70], we use the *root mean square distance (RMSD)* metric as a distance function employing Euclidean distance. It measures the similarity across the post-registration point cloud and the target point cloud (ground truth). This metric often appears in the literature as RMSE, but we decided to differentiate it from the aforementioned RMSE of Eq. (29) in order to highlight the difference of measuring the distance between the target and transformed point clouds from the error based only on the ground truth transformation. This results from the observation that the source and the target are not the same or parts of the same point cloud. The target object is eroded, which means that even if we perform the ground truth transformation on the source object, the result will not coincide with the target object. Thus, the RMSD which measures the distance between the point clouds, will not present the real registration success or error. We estimate RMSD as:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{T}_{pred}\, p_i - \widehat{p}_i)\|^2} \tag{30}$$

Table 1
Registration results on the ECHO dataset when only random rotations, translations and 60 years of erosion are performed on the initial objects. The metrics evaluated are rotation error, Error(R), translation error, Error(t), root mean square error, RMSE, root mean square distance, RMSD and Recall with threshold 0.2. Bold and *dark gray* denote best and second best performing methods for each measure respectively. For fairness reasons, we have not included in bold, cases where CrossTimeReg performs best when trained on the training partition of ECHO; instead such cases are in bold italics.

	Method	Registration		Error(R)	Error(t)	RMSE	RMSD	Recall$_\alpha$(%)	Mean Exec. Time (sec)
		Local	Global						
Geometry-based	ICP [19]	✓		1.6992	42.5667	38.6065	42.583	0	34
	FPFH-RANSAC [21,27]		✓	1.8314	29.2151	29.3316	29.2326	0	32
	SI-FGR [22,28]		✓	1.8202	0.0629	1.1298	1.1344	21.91	32
	SISI-RANSAC [21,30]		✓	0.9984	0.1044	0.6870	0.6877	96.88	67
	LD-SIFT -RANSAC [21,30]		✓	**0.3496**	0.0793	**0.2789**	**0.2878**	98.79	68
	RICI-FGR [22,29]		✓	1.1396	*0.0495*	1.1832	1.1396	20.77	38
Deep Learning	PRNet [48]	✓		1.7514	1.0184	1.4723	1.4858	43.12	14
	PointNetLK [39]	✓		1.7413	29.2389	29.2514	29.2561	0	11
	PCRNet[41]		✓	1.8095	49.3442	49.3603	49.3600	0	10
	RPM-Net [43]		✓	1.6993	29.2594	29.2784	29.2755	0	15
	DCP [42]		✓	1.6881	38.6109	38.6542	38.6133	0	15
	DeepGMR [45]		✓	1.0065	0.0673	0.9454	*0.6746*	*99.31*	4
	CrossTimeReg		✓	*0.9942*	**0.0448**	*0.6764*	0.6812	**99.55**	6
	CrossTimeReg (trained on ECHO)		✓	***0.1397***	0.0714	***0.2606***	0.6928	***99.98***	6

Since exact point correspondences do not exist in cross-time registration, we approximate the computation using the nearest neighbor $\widehat{p}_i$ of the respective point.

Finally, we compute the success rate across the dataset *recall*$_\alpha$, i.e. the percentage of tests for which the RMSE is below a certain threshold α:

$$recall_\alpha = \frac{|S_\alpha|}{|S|} \times 100\% \tag{31}$$

where $|S|$ is the total number of tests performed and $|S_\alpha|$ is the number of tests that achieve RMSE less that the threshold α.

In previous literature, more metrics have been proposed and used to evaluate registration techniques, such as Chamfer distance or Earth Mover's Distance [71]. However, these metrics are less robust and have the same problem as RMSD in the case of erosion, i.e. they do not take into consideration the common erosion that may have occurred on all points of the surface. They have thus not been considered further.

6.2. Experimental results and analysis

Synthetic data - ECHO dataset In Table 1 we summarize the quantitative registration results on the challenging ECHO dataset for cross-time 3D registration; Fig. 8 illustrates some qualitative results. CrossTimeReg generally outperforms the state-of-the-art under most performance metrics.

Since cross-time registration involves point clouds with non exact point-level correspondences, methods like FPFH and DCP fail to converge in every run of this experiment. In addition, the initial poses of corresponding objects are generally far apart, both in terms of translation and rotation, and thus local methods like ICP, PointNetLK and PRNet fail to converge for many objects.

The performance of geometry-based global registration methods RANSAC and FGR rely on feature matching or keypoint detection from hand-crafted descriptors. Such descriptors face unusual challenges in the case of eroded objects. When SI is used as the local descriptor, its instability in the presence of noise and non-uniform sampling of the object's surface, result in many failed registrations. SISI and RICI, being derivatives of the SI, face similar challenges. RICI fails to properly identify the keypoints across the source and the target because it counts the intersections of homocentric circles with the surface. A target object which is evenly eroded produces different intersections to the corresponding source object. LD-SIFT, being both rotation and scale invariant, performs considerably better than the rest of the state-of-the-art; since erosion may affect the surface of the object evenly, the scale invariant features result in better recovery of the correct transformation. However, in terms of translation, the errors are larger and this is reflected in the Recall$_\alpha$ metric which is not as good as that of CrossTimeReg. A significant disadvantage of LD-SIFT is the large computation time and memory requirements which precludes its use in real time applications and on large scale datasets.

Interestingly, most deep learning methods perform significantly worse on the cross-time registration problem than geometry-based methods. This can be due to the fact that the networks have been trained on a different dataset and task than the related test ones. As mentioned before, to ensure a fair comparison, all deep learning methods have been trained on the ModelNet dataset for the traditional 3D registration problem. Thus, methods like PointNetLK that are trained on feature detection for specific object categories, fail to recognize useful features in different objects categories, like the pots and figurines of cultural heritage datasets. The generalization to unseen data, unrestricted rotation and significant translation results in poor performance for many deep learning methods. However, PRNet, DeepGMR and CrossTimeReg seem to overcome this obstacle and produce accurate registration results. The fact that PRNet was designed to perform partial-to-partial registration, can explain why the method converges on the cross-time registration problem. Cross-time registration shares a lot of common with the partial-to-partial registration, since the source and the target may have different surfaces but share common parts of their geometry. Still, PRNet is a local method and does not converge under large transformations.

Both DeepGMR and CrossTimeReg learn latent correspondences between the point clouds and GMM components, which are pose-invariant. Thus, the registration result is invariant to the magnitude of transformation or the density of the input geometries. However, DeepGMR estimates the correspondence between all points and all components in the latent GMM, meaning that its performance degrades when the point clouds partially overlap or if the points of the source and the target point clouds have been shuffled and randomly sampled. CrossTimeReg overcomes this drawback with the addition of the curvature based sub-sampling step. Moreover, with the addition of the KPConv network, CrossTimeReg learns local shifts effectively, implying that it learns the erosion part of the cross-time registration. The CrossTimeReg model has been trained in the same dataset as the rest of the deep

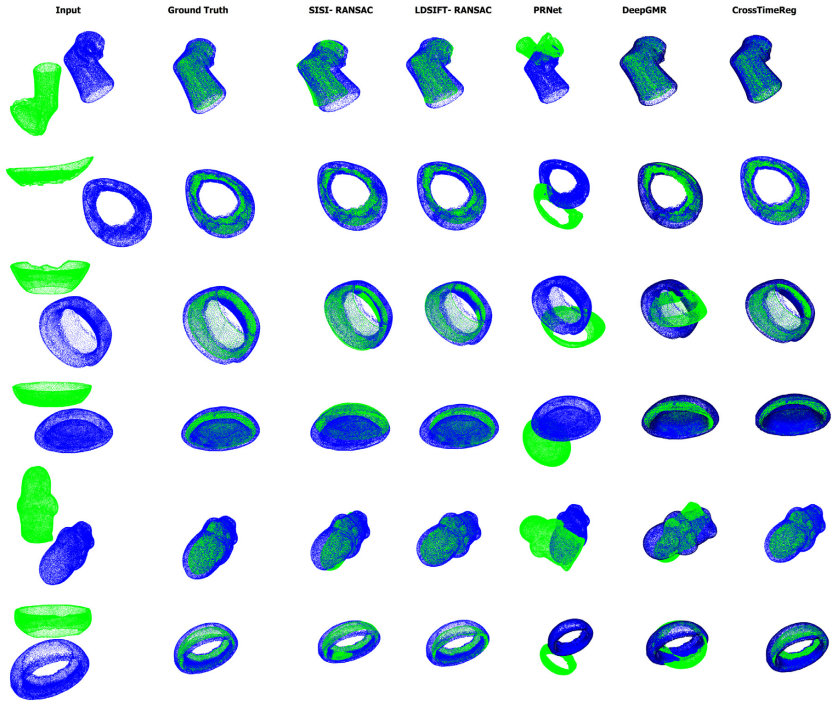


Fig. 8. Comparison between different registration methods on examples from the ECHO dataset for cross-time registration. Methods with the highest recall rates ($Recall_\alpha$ > 40%) are included.

networks (ModelNet), so that it is fairly comparable against the state-of-the-art. In order to investigate the effect of using eroded models in training, we have also fine-tuned the CrossTimeReg model on the training partition of the ECHO dataset. As can be seen from the last line of Table 1, the performance increases spectacularly.

ECHO Dataset - Multiple levels of erosion: In order to detect how the registration methods perform on different levels of degradation, in this section we evaluate the registration methods on the ECHO dataset against different levels of erosion. We have performed experiments for 20 different levels of erosion; from 1 year up to 60 years. Fig. 9 shows the RMSE metric for the most accurate methods. For clarity of illustration, we have excluded the methods which had average RMSE greater than 20, for every erosion level. It can be deduced that most methods tend to perform worse as the level of erosion increases. This is because when there is no or small amount of degradation, the geometry of the objects remains basically the same, so the identified keypoints and subsequent registration are accurate enough. However, as degradation increases, the target shape differentiates more and more from the source shape and most traditional registration methods tend to lose accuracy. Across all levels of erosion, CrossTimeReg appears to have stable performance, which even increases slightly at the highest levels.

In Table 2 we summarize the quantitative registration results on the ECHO dataset when only random rotations and translations are performed (no erosion).

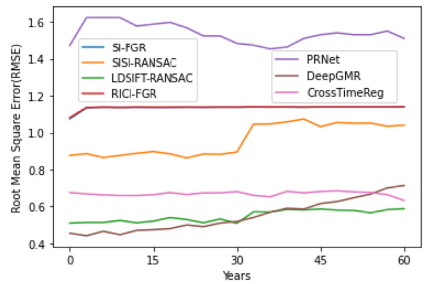


Fig. 9. Comparison between different registration methods on examples from the ECHO dataset for cross-time registration on different levels of erosion.

By comparing Tables 1 and 2 we can see that, relative to other methods, CrossTimeReg performs better when erosion is involved.

SHREC2016 dataset - traditional registration: In this section we evaluate the methods on traditional registration using the SHREC2016 dataset [65]. We chose this dataset, consisting of 383 models, as it is related to the cultural heritage domain. The 3D objects are pottery models originating from the Virtual Hampson Museum collection [72]. Again, we performed random rotations and

Table 2
Registration results on the ECHO dataset when only random rotations and translations are performed on the initial objects (no erosion). Bold and *dark gray* denote best and second best performing methods for each measure respectively.

	Method	Error(R)	Error(t)	RMSE	Recall$_a$(%)
Geometry-based	ICP [19]	1.6453	40.3423	30.0234	0
	FPFH-RANSAC [21,27]	1.8315	29.2325	29.4201	0.07
	SI-FGR [22,28]	1.7274	**0.0247**	1.0814	92.52
	SISI-RANSAC [21,30]	1.2945	0.2363	0.8774	85.89
	LD-SIFT -RANSAC [21,30]	**0.7021**	0.1661	**0.5102**	*99.01*
	RICI-FGR [22,29]	1.7392	*0.0272*	1.0814	93.73
Deep Learning	PRNet [48]	1.7368	0.9868	1.4728	49.35
	PointNetLK [39]	1.7346	29.2016	29.2192	0
	PCRNet[41]	1.8054	49.4641	49.4701	0
	RPM-Net [43]	1.6779	29.1860	29.2018	0
	DCP [42]	1.7219	39.7070	39.7200	0
	DeepGMR [45]	0.9578	1.9203	*0.5192*	98.34
	CrossTimeReg	*0.9456*	1.0821	0.6751	**99.43**

Table 3
Registration results on SHREC2016 dataset when only random rotations and translations are performed (no erosion). Bold and *dark gray* denote best and second best performing methods for each measure respectively.

	Method	Error(R)	Error(t)	RMSE	Recall$_a$(%)
Geometry-based	ICP [19]	1.2593	20.8497	91.60	9.14
	FPFH-RANSAC [21,27]	1.2295	18.2610	89.56	9.39
	SI-FGR [22,28]	*0.0062*	**0.0014**	**0.002**	*99.67*
	SISI-RANSAC [21,30]	0.0118	1.6920	3.40	99.47
	LD-SIFT -RANSAC [21,30]	**0.0021**	*0.0418*	**0.002**	**99.74**
	RICI-FGR [22,29]	0.0088	1.6819	2.66	98.47
Deep Learning	PRNet [48]	1.4985	71.6138	119.12	1.30
	PointNetLK [39]	1.4044	22.0351	100.32	7.57
	PCRNet[41]	1.3341	35.9334	98.40	2.61
	RPM-Net [43]	1.3670	98.0969	143.25	0.26
	DCP [42]	1.5797	123.5982	160.85	0
	DeepGMR [45]	0.7650	0.2756	5.47	98.56
	CrossTimeReg	0.0341	0.2328	*2.64*	98.13

Fig. 10. Some of the stone slabs used in PRESIOUS accelerated erosion experiments [16]. The stones named Elefsis consist of pentelic marble, while stones names Nidaros consist of grytdal soapstone, representing the material of two monuments that were considered in the PRESIOUS project, the Demeter Sanctuary in Elefsis (Greece) and the Nidaros Cathedral in Trondheim (Norway) respectively.

translations to the objects (no erosion is present) and the results are shown in Table 3. The behaviour of the methods is similar to that on the ECHO dataset, a positive indication for the quality of ECHO; CrossTimeReg demonstrates high accuracy without achieving the top results. Interestingly, in both Tables 2 and 3 where no erosion is involved, CrossTimeReg is one of the top 2 deep learning methods, while non-learning methods perform extremely well.

PRESIOUS dataset - real erosion data from accelerated erosion experiments: To demonstrate how CrossTimeReg performs in the case of real erosion data, we employed data from the PRESIOUS EU project [16]. These data consist of three accelerated erosion experiments on two different types of stone slabs; pentelic marble and grytdal soapstone, see Fig. 10.

The erosion effects that were studied in the accelerated erosion experiments were acid rain weathering, salt and freeze-and-thaw effects. Table 4 gives details on the 3D scanned slabs across the erosion experiments. We tested CrossTimeReg and LD-SIFT with RANSAC, see Table 5. For each stone slab, we register the initial scan with the scan after the first period of erosion (Round 1 - Round 2), the scan after the first period with the final scan (Round 2 - Round 3) and the initial scan with the final scan after 2 periods of experiments (Round 1 - Round 3). Both methods have been run on the same hardware. CrossTimeReg's execution time increased with the number of object points, but this was only due to the curvature downsampling component; the execution of the rest of the modules of CrossTimeReg remained constant, irrespective of the number of object points. LD-SIFT had to be interrupted after 4 hours on the same data. To overcome this, we uniformly down-sampled the 3D models, so that the down-sampled meshes would contain 50K points. For this reason, we omitted the execution time of LD-SIFT in the above Table. Since, the dataset had no ground truth of the transformations performed on the objects, we measured the RMSD based on eq. (30), which measures also the distance due to the degradation of the objects. Table 5 and Fig. 11 show that CrossTimeReg behaves favourably compared to LD-SIFT on real data.

Evaluating curvature downsampling (CDS) for cross-time registration: In order to get a better intuition of the contribution of the proposed downsampling method on final performance, we carried out an ablation study on the ECHO dataset. The erosion level was varied from no erosion (only random rotations and translations) up to 60 years of erosion due to deposition of crust and acid rain. We compare against the case where uniform sampling, based on

Table 4
Details of the slabs used in the PRESIOUS dataset. Round 1 contains the initial scanned object, Round 2 after the first period of erosion and Round 3 after the second period of erosion.

Experiment	Material	Stone Slab Label	Round	# points
Freeze and Thaw	Pentelic Marble	Elefsis Large 01	1	998621
			2	790553
			3	847791
	Grytdal Soapstone	Nidaros Bad Large 01	1	1904088
			2	2671989
			3	2778924
Salt	Pentelic Marble	Elefsis Large 02	1	1236236
			2	1050038
			3	1336365
	Grytdal Soapstone	Nidaros Bad Large 02	1	2023069
			2	3978584
			3	4250544
Acid Rain	Pentelic Marble	Elefsis Large 03	1	983698
			2	976587
			3	612447
	Grytdal Soapstone	Nidaros Good Large 03	1	3009981
			2	2858613
			3	3130228

Table 5
Registration results on the PRESIOUS dataset of real erosion data from accelerated experiments. The best performance between LD-SIFT and CrossTimeReg is in bold. CrossTimeReg was run on the original datasets; LD-SIFT had not completed execution after 4h on the original datasets and was run on subsampled versions.

Exp.	Stone Slab Label	Rounds	LD-SIFT	Cross Time Reg	
			RANSAC [21,30] RMSD	RMSD	Exec.Time (sec)
Freeze and Thaw	Elefsis Large 01	1 - 2	0.03854	**0.03817**	250
		2 - 3	0.05358	**0.03858**	196
		1 - 3	0.05916	**0.03867**	264
	Nidaros Bad Large 01	1 - 2	0.05087	**0.03617**	450
		2 - 3	0.03605	**0.03132**	511
		1 - 3	0.03840	**0.03838**	461
Salt	Elefsis Large 02	1 - 2	0.04682	**0.03888**	252
		2 - 3	**0.03667**	0.03719	250
		1 - 3	0.04079	**0.03676**	239
	Nidaros Bad Large 02	1 - 2	0.03831	**0.03564**	663
		2 - 3	0.04702	**0.03793**	900
		1 - 3	0.03537	**0.03033**	699
Acid Rain	Elefsis Large 03	1 - 2	**0.03552**	0.03799	174
		2 - 3	0.04253	**0.03544**	118
		1 - 3	0.05479	**0.03667**	150
	Nidaros Good Large 03	1 - 2	**0.03896**	0.03931	331
		2 - 3	**0.03460**	0.03819	343
		1 - 3	0.03794	**0.03253**	406

voxel size, is used; in our experiments we considered voxel size = 0.05 (5 cm in metric scale) which gives the best results. We calculated the RMSE based on Eq. (29). As can be seen from Fig. 12, the proposed downsampling scheme behaves stably across levels of erosion. On the contrary, uniform downsampling has better RMSE results on small erosion values but increases with the level of erosion.

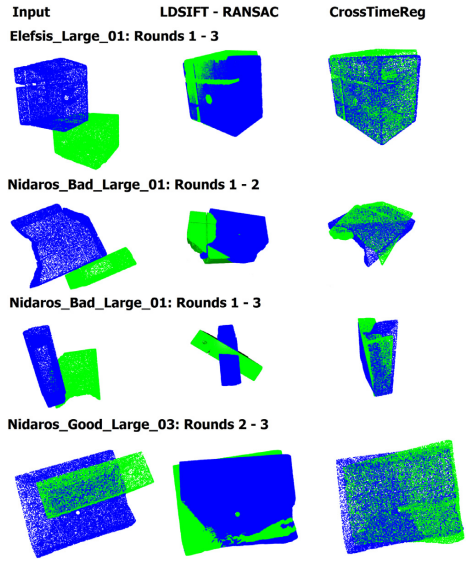


Fig. 11. Qualitative comparison between CrossTimeReg and LD-SIFT on examples from the PRESIOUS dataset of real eroded data.

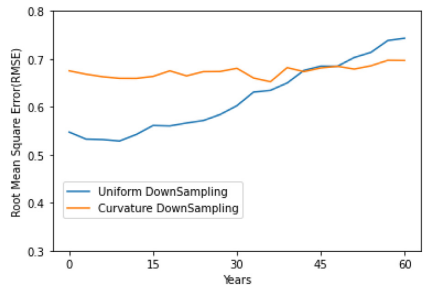


Fig. 12. Ablation study of downsampling methodologies on different levels of erosion on the ECHO dataset.

7. Conclusions and future work

The challenging problem of cross-time 3D registration has been defined and CrossTimeReg, a deep learning method for cross-time 3D registration, has been proposed. CrossTimeReg achieves state-of-art accuracy and robustness to large initial transformations while being computationally efficient. Indeed computational efficiency is a main advantage of the proposed method against previous very accurate geometry based methods. The proposed method is also very stable as the level of erosion increases. Its computational efficiency can be further optimized, especially the subsampling step. A new dataset, ECHO, has been created to facilitate the evaluation of techniques on cross-time registration with high quality models and ground truth. We anticipate that the public availability of ECHO will facilitate future experiments for cross-time related tasks (registration, retrieval, recognition). In ad-

dition, an implementation-based comparison of both deep learning and geometry-based object registration algorithms has been made, with some interesting observations.

As a future step, we consider to replace the hand-crafted features with features learned specifically for cross-time registration by the network. We plan to integrate our network into larger systems, for tasks such as monitoring and segmentation of changes that CH objects undergo and extend the framework to register and fuse multiple modalities, in addition to 3D point clouds. Moreover, we intend to study partiality in conjunction with degradation on CH objects.

If extreme erosion values are applied on surface meshes, then mesh folding can occur. This can also arise on boundary and fragile edges (e.g. Fig. 2). Mesh folding is a challenging but worthwhile problem to address (see for example [73,74]).

Another potential avenue of future work is to apply a non-uniform, more realistic erosion model in the simulation process, which could take into consideration the surface's orientation, shape and texture.

Credit Author Statement

Evdokia Saiti: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization; Software; Writing - original draft. **Antonios Danelakis:** Conceptualization; Methodology; Validation; Visualization; Software; Writing review & editing. **Theoharis Theoharis:** Conceptualization; Funding acquisition; Project administration; Methodology; Resources; Supervision; Validation; Visualization; Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation program CHANGE under the Marie Skłodowska-Curie grant agreement No. 813789.

References

[1] Fonseca MJ, Ferreira A, Jorge JA. Towards 3D modeling using sketches and retrieval. Proceedings of the Eurographics workshop on sketch-based interfaces and modeling 2004. Citeseer; 2004. p. 127.
[2] Gojcic Z, Zhou C, Wegner JD, Guibas LJ, Birdal T. Learning multiview 3D point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 1759–69.
[3] Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM. Dynamic graph CNN for learning on point clouds. ACM Trans Graph (tog) 2019;38(5):1–12.
[4] Wu W, Qi Z, Fuxin L. Pointconv: deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9621–30.
[5] Kim P, Chen J, Cho YK. SLAM-driven robotic mapping and registration of 3D point clouds. Autom Constr 2018;89:38–48.
[6] Weinmann M, Leitloff J, Hoegner L, Jutzi B, Stilla U, Hinz S. Thermal 3D mapping for object detection in dynamic scenes. ISPRS Ann Photogramm Remote Sens Spat Inf Sci 2014;2(1):53.
[7] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2013. p. 2100–6.
[8] Mellado N, Dellepiane M, Scopigno R. Relative scale estimation and 3D registration of multi-modal geometry using growing least squares. IEEE Trans Vis Comput Graph 2015;22(9):2160–73.
[9] Papaioannou G, Schreck T, Andreadis A, Mavridis P, Gregor R, Sipiran I, Vardis K. From reassembly to object completion: a complete systems pipeline. J Comput Cult Herit (JOCCH) 2017;10(2):1–22.
[10] Chang W, Zwicker M. Global registration of dynamic range scans for articulated model reconstruction. ACM Trans Graph (TOG) 2011;30(3):1–15.
[11] Zollhöfer M, Stotko P, Görlitz A, Theobalt C, Nießner M, Klein R, Kolb A. State of the art on 3D reconstruction with RGB-D cameras. In: Computer graphics forum, vol. 37. Wiley Online Library; 2018. p. 625–52.
[12] Saiti E, Theoharis T. An application independent review of multimodal 3D registration methods. Comput Graph 2020;91:153–78.
[13] Pintus R, Pal K, Yang Y, Weyrich T, Gobbetti E, Rushmeier HE. Geometric analysis in cultural heritage.. In: Proceedings of the GCH; 2014. p. 117–33.
[14] Saha S, Foryś P, Martusewicz J, Sitnik R. Approach to analysis the surface geometry change in cultural heritage objects. In: Proceedings of the international conference on image and signal processing. Springer; 2020. p. 3–13.
[15] SHREC 2021: retrieval of cultural heritage objects. http://www.ivan-sipiran.com/shrec2021.html; Accessed on March2021.
[16] PRESIOUS fp7-600533 eu project -finalevaluationreport. http://presious.eu/file_downloads/PRESIOUS-D5.8-FinalEvaluationReport.pdf, Accessed on May2021.
[17] Tam GK, Cheng Z-Q, Lai Y-K, Langbein FC, Liu Y, Marshall D, Martin RR, Sun X-F, Rosin PL. Registration of 3D point clouds and meshes: a survey from rigid to nonrigid. IEEE Trans Vis Comput Graph 2012;19(7):1199–1217.
[18] Tombari F, Remondino F. Feature-based automatic 3D registration for cultural heritage applications. In: Proceedings of the digital heritage international congress (DigitalHeritage), vol. 1. IEEE; 2013. p. 55–62.
[19] Besl PJ, McKay ND. Method for registration of 3D shapes. In: Sensor fusion IV: control paradigms and data structures, vol. 1611. International Society for Optics and Photonics; 1992. p. 586–606.
[20] Rusinkiewicz S, Levoy M. Efficient variants of the ICP algorithm. In: Proceedings of the third international conference on 3-D digital imaging and modeling. IEEE; 2001. p. 145–52.
[21] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 1981;24(6):381–95.
[22] Zhou Q-Y, Park J, Koltun V. Fast global registration. In: Proceedings of the European conference on computer vision. Springer; 2016. p. 766–82.
[23] Zhang Z, Dai Y, Sun J. Deep learning based point cloud registration: an overview. Virtual Real Intell Hardw 2020;2(3):222–46.
[24] Mavridis P, Andreadis A, Papaioannou G. Efficient sparse ICP. Comput Aided Geom Des 2015a;35:16–26.
[25] Han X.-F., Sun S.-J., Song X.-Y., Xiao G.-Q.. 3D Point cloud descriptors in hand-crafted and deep learning age: state-of-the-art. arXiv e-prints2018;:arXiv–1802.
[26] Diez Y, Roure F, Lladó X, Salvi J. A qualitative review on 3D coarse registration methods. ACM Comput Surv (CSUR) 2015;47(3):1–36.
[27] Rusu RB, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3D registration. In: Proceedings of the IEEE international conference on robotics and automation. IEEE; 2009. p. 3212–17.
[28] Johnson AE, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans Pattern Anal Mach Intell 1999;21(5):433–49.
[29] van Blokland BI, Theoharis T. Radial intersection count image: a clutter resistant 3D shape descriptor. Comput Graph 2020;91:118–28.
[30] Darom T, Keller Y. Scale-invariant features for 3D mesh models. IEEE Trans Image Process 2012;21(5):2758–69.
[31] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91–110.
[32] Yang J, Li H, Campbell D, Jia Y. Go-ICP: a globally optimal solution to 3D ICP point-set registration. IEEE Trans Pattern Anal Mach Intell 2015;38(11):2241–54.
[33] Li H, Hartley R. The 3D-3D registration problem revisited. In: Proceedings of the IEEE 11th international conference on computer vision. IEEE; 2007. p. 1–8.
[34] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 1977;39(1):1–22.
[35] Granger S, Pennec X. Multi-scale EM-ICP: a fast and robust approach for surface registration. In: Proceedings of the European conference on computer vision. Springer; 2002. p. 418–32.
[36] Jian B, Vemuri BC. Robust point set registration using gaussian mixture models. IEEE Trans Pattern Anal Mach Intell 2010;33(8):1633–45.
[37] Eckart B, Kim K, Kautz J. HGMR: hierarchical gaussian mixtures for adaptive 3D registration. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 705–21.
[38] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 652–60.
[39] Aoki Y, Goforth H, Srivatsan RA, Lucey S. Pointnetlk: robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 7163–72.
[40] Lucas BD, Kanade T, et al. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on artificial intelligence, vol. 2. Vancouver, British Columbia; 1981. p. 674–9.
[41] Sarode V., Li X., Goforth H., Aoki Y., Srivatsan R.A., Lucey S., Choset H.. Pcrnet: point cloud registration network using pointnet encoding. arXiv preprint arXiv:1908079062019;.
[42] Wang Y, Solomon JM. Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019a. p. 3523–32.
[43] Yew ZJ, Lee GH. Rpm-net: Robust point matching using learned features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 11824–33.
[44] Gold S, Rangarajan A, Lu C-P, Pappu S, Mjolsness E. New algorithms for 2D and 3D point matching: pose estimation and correspondence. Pattern Recognit 1998;31(8):1019–31.

[45] Yuan W, Eckart B, Kim K, Jampani V, Fox D, Kautz J. Deepgmr: Learning latent gaussian mixture models for registration. In: Proceedings of the European conference on computer vision. Springer; 2020. p. 733–750.
[46] Mellado N, Aiger D, Mitra NJ. Super 4PCS fast global pointcloud registration via smart indexing. In: Computer graphics forum, vol. 33. Wiley Online Library; 2014. p. 205–15.
[47] Chen Y-W, Mimori A, Lin C-L. Hybrid particle swarm optimization for 3-d image registration. In: Proceedings of the 16th IEEE international conference on image processing (ICIP). IEEE; 2009. p. 1753–6.
[48] Wang Y., Solomon J.M.. Prnet: self-supervised learning for partial-to-partial registration. arXiv preprint arXiv:1910122402019b;.
[49] Mavridis P, Andreadis A, Papaioannou G. Fractured object reassembly via robust surface registration.. In: Proceedings of the Eurographics (Short Papers); 2015b. p. 21–4.
[50] Bromblet P, Vallet J, Verges-Belmin V. Illustrated glossary on stone deterioration patterns, 3. Monuments and sites; 2008.
[51] Perakis P, Schellewald C, Kebremariam KF, Theoharis T. Simulating erosion on cultural heritage monuments. Proceedings of the 20th international conference on cultural heritage and new technologies (CHNT20); 2015.
[52] Barbosa IB, Gebremariam KF, Perakis P, Schellewald C, Theoharis T. Establishing parameter values for the stone erosion process. In: Proceedings of the CAA2015; 2015. p. 347.
[53] Hameiri E, Shimshoni I. Estimating the principal curvatures and the Darboux frame from real 3-D range data. IEEE Trans Syst Man Cybern Part B (Cybern) 2003;33(4):626–37.
[54] Agapaki E, Nahangi M. Chapter 3 - scene understanding and model generation. In: Brilakis I, Haas C, editors. Infrastructure computer vision. Butterworth-Heinemann; 2020. p. 65–167. ISBN 978-0-12-815503-5, 10.1016/B978-0-12-815503-5.00003–6
[55] Yang Y-L, Lai Y-K, Hu S-M, Pottmann H, et al. Robust principal curvatures on multiple scales. In: Proceedings of the symposium on geometry processing; 2006. p. 223–6.
[56] Patrikalakis NM, Maekawa T. Shape interrogation for computer aided design and manufacturing. Springer Science & Business Media; 2009.
[57] Chen C, Li G, Xu R, Chen T, Wang M, Lin L. Clusternet: deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 4994–5002.
[58] Thomas H, Qi CR, Deschaud J-E, Marcotegui B, Goulette F, Guibas LJ. Kpconv: flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 6411–20.
[59] Thomas H.. Rotation-invariant point convolution with multiple equivariant alignments. arXiv preprint arXiv:2012040482020;.
[60] Khazari AE, Que Y, Sung TL, Lee HJ. Deep global features for point cloud alignment. Sensors 2020;20(14):4032.
[61] Yerrapragada SS, Jaynes JH, Chirra SR, Gauri K. Rate of weathering of marble due to dry deposition of ambient sulfur and nitrogen dioxides. Anal Chem 1994;66(5):655–9.
[62] Baedecker PA, Reddy MM. The erosion of carbonate stone by acid rain: laboratory and field investigations. J Chem Educ 1993;70(2):104.
[63] Gauri KL, Bandyopadhyay JK. Carbonate stone: chemical behavior, durability and conservation; 1999.
[64] Yerrapragada SS, Chirra SR, Jaynes JH, Li S, Bandyopadhyay JK, Gauri K. Weathering rates of marble in laboratory and outdoor conditions. J Environ Eng 1996;122(9):856–63.
[65] Pratikakis I, Savelonas M, Arnaoutoglou F, Ioannakis G, Koutsoudis A, Theoharis T, Tran M, Nguyen V, Pham V, Nguyen H, et al. SHREC'16 Track: partial shape queries for 3D object retrieval Proceedings of the 3DOR, 1; 2016.
[66] Zhou Q.-Y., Park J., Koltun V.. Open3D: a modern library for 3D data processing. arXiv:1801098472018;.
[67] Learning3d: Learning3D: a modern library for deep learning on 3D point clouds data.https://github.com/vinits5/learning3d, Accessed on March2021.
[68] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1912–20.
[69] Choi S, Zhou Q-Y, Koltun V. Robust reconstruction of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 5556–65.
[70] Yang J, Cao Z, Zhang Q. A fast and robust local descriptor for 3D point cloud registration. Inf Sci 2016;346:163–79.
[71] Fan H, Su H, Guibas LJ. A point set generation network for 3D object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 605–13.
[72] Virtual Hampson Museum collection. https://hampson.cast.uark.edu/, Accessed on May2021.
[73] Kim S-J, Lee D-Y, Yang M-Y. Offset triangular mesh using the multiple normal vectors of a vertex. Comput Aided Des Appl 2004;1(1–4):285–91.
[74] Chen Y, Wang CC. Uniform offsetting of polygonal model based on layered depth-normal images. Comput-Aided Des 2011;43(1):31–46.

Chapter 7

Paper B - An automated approach for change and difference detection on cultural heritage applications

Authors

Saiti Evdokia, Sunita Saha, Eryk Bunsch, Robert Sitnik, and Theoharis Theoharis.

Under review in

Digital Applications in Archaelogy and Cultural Heritage, 2023

Chapter 8

Paper C - An application independent review of multimodal 3D registration

Authors

Evdokia Saiti, and Theoharis Theoharis.

Published in

Computer & Graphics, Volume 91, Pages 153-178, Elsevier, 2020

Computers & Graphics 91 (2020) 153–178

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

An application independent review of multimodal 3D registration methods☆

E. Saiti*, T. Theoharis

NTNU Department of Computer Science Gløshaugen, Sem Sælands vei 9 Trondheim 7034 Norway

ARTICLE INFO

Article history:
Received 1 May 2020
Revised 20 July 2020
Accepted 27 July 2020
Available online 31 July 2020

Keywords:
Registration
Multimodal
Survey
3D

ABSTRACT

Registration is a ubiquitous operation in Visual Computing, with applications in 3D object retrieval among others. Registration is the process of overlaying two or more datasets taken from different viewpoints, at different times or by different sensors into a common reference frame. Multimodal registration is a special case where the data to be matched do not belong to the same modality and is challenging due to the diverse nature of the modalities involved which makes the creation of a distance function harder. Due to the large number of possible modality combinations and application fields, a considerable number of multimodal registration techniques have been proposed in diverse fields, including medicine and archaeology. This survey aims to unify 3D multimodal registration techniques (i.e. where at least one of the modalities is in 3D) across application domains, with the hope of providing an application-independent view and the potential for cross-fertilization. The problem of 3D multimodal registration is explicitly defined and the various methods are systematically categorized and described in terms of a number of important properties. Methods with publicly available source code have been compared on common datasets. A discussion on trends, observations and challenges for further research concludes the review.

1. Introduction

The technological progress of the last decades has led to an explosion in volume, variety and complexity of data. There is a massive amount of highly heterogeneous 2D and 3D datasets consisting of multimodal samples acquired by a variety of different sensors. 3D data can exist in different domains, in different types of format, characteristics and possess different sources of error. For such data to be exploited, the proper alignment in a common coordinate system is often essential.

This alignment, or *registration*, has become a fundamental task in computer vision and computer graphics and a host of applications use alignment techniques before visualizing, comparing or processing data. Registration techniques are utilized in multiple operations, such as 3D object retrieval [1], 3D mapping [2–4], 3D object scanning [5], 3D model reconstruction [6,7], which are basic components of applications such as cultural heritage [8–10] and medical imaging [11,12].

Registration is the process of aligning two or more similar objects or two or more instances of the same object taken at different times (multi-temporal data), from different viewpoints (multi-view data) or by different sensors (multi-sensor data) into a common reference system. Given a target and source/reference dataset, a registration technique can be described by three components: the transformation which relates the two datasets, the similarity metric that evaluates the similarity of the datasets and an optimization method which determines the optimal transformation parameters as a function of the similarity metric. Thus, a registration method geometrically aligns two datasets by finding an optimal transformation that minimizes the error of a similarity metric.

Multimodal registration is a special category of registration, where the data to be aligned are of the same object but of different modality (Fig. 1). Multimodal data may have different data structure, dimension, density, noise and types of error in their geometry. Multi-modality is also referred in the literature as inter-modality or cross-modality. Compared to unimodal registration, the multimodal case is more challenging because it is not straightforward to define a general registration framework for relating the different modalities.

☆ This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813789.

* Corresponding author.
E-mail addresses: evdokia.saiti@ntnu.no (E. Saiti), theotheo@ntnu.no (T. Theoharis).

https://doi.org/10.1016/j.cag.2020.07.012

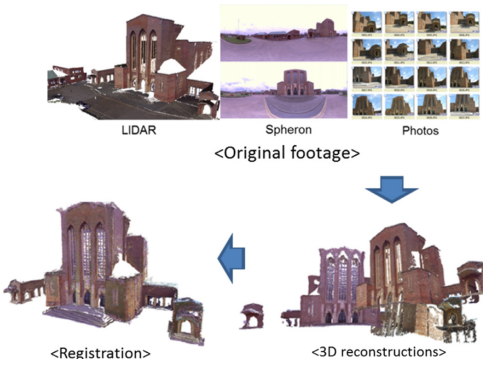


Fig. 1. Multimodal data registration as presented in [13].

There has been significant growth in research on registration of 3D data both unimodally and multimodally. Several surveys have been published covering aspects of image registration [14,15] and 3D unimodal registration [16–19]. Registration of images has been extensively researched in the medical imaging domain, resulting in multiple reviews, focused on medical applications [11] or modalities [20]. Refer to [21–23] for surveys covering the main issues and methods related to medical image registration techniques. Recently a lot of attention has been directed into utilizing deep learning for registration of medical images, also leading to some surveys [24–27].

Due to the breadth of the registration research field and the volume of research performed and published each year, we focus this review on methods for multimodal 3D registration as defined below, a topic that has not been covered by a survey before to the best of our knowledge. At the same time, we strive to be open to all application areas where such techniques have been developed with the aim of showing commonalities as well as potential for cross-fertilization. We restrict ourselves to techniques where one or both modalities are three dimensional as this is arguably the most common and useful dimensionality; such techniques are either concerned with different 3D modalities or work across 2D and 3D. We take as starting point the work of Kotsas et al. [28] for registration techniques of different dimensionality (2D/3D) as well as the review of Andrade et al. [21], both specifically for medical image registration.

The remainder of this paper is organized as follows: In Section 2 the 3D multimodal registration problem is defined and analyzed. Section 3 presents applications of 3D multimodal registration while Section 4 presents multimodal registration attributes. In Section 5, public datasets and performance evaluation measures are presented. Section 6 overviews the multimodal registration methods; optimization-based registration techniques in subsection 6.1 and learning-based approaches in subsection 6.2. Section 7 compares methods with publicly available source code on common datasets while, finally, in Section 8 we reflect on the past and anticipate on future perspectives for multimodal 3D registration.

2. Multimodal 3D data registration

The term multimodal registration has largely been 'abused' in the literature, referring to such aspects as the same object from different viewpoints, the same object at different moments in time or the same object scanned by different sensors. Thus the data may share the same geometric characteristics and even the same data structure (e.g. registering dense 3D point clouds produced by terrestrial laser scanners at different times and from different views [29] or registering CT and cone-beam CT (CBCT) spine images which have different fields of view [30]). Although, different sensors can produce variations in terms of density, scale, noise and deformation, the data are often geometrically similar and within the same family of data structure (e.g. a low resolution 3D point cloud and a high resolution 3D mesh generated from 3D scanning [5]).

What should then be the characteristics of two modalities in order to be considered different? To answer this question, we have tried to locate what makes multimodal registration a more challenging task than unimodal registration. It has been observed that registration methods that perform well in the unimodal case [31,32], do not necessarily perform well when they are applied to multimodal datasets [33]. In unimodal registration, data have similar or correlated statistical properties and it is rather straightforward to recognize correspondences or a similarity metric. The core difficulty in multimodal registration is in identifying structure correspondences across modalities or defining a general rule to identify similarity between two modalities with different physical principles.

Therefore, we will herein use the term *multimodal* to refer to two datasets with qualitative variability in shape and appearance; thus having different dimension (e.g. 3D/2D images, X-ray / MRI), different data structure (e.g. 3D point cloud and an MRI volume) or different physical and anatomical principles (e.g. MRI and CT volumes). We shall thus not include methods that register the same modalities generated by different acquisition devices (e.g. [34]), same modalities with different resolutions (e.g. alignment of a low resolution point cloud/mesh with high resolution point cloud/mesh [35]) or the same modalities with different imaging parameters (e.g. registration of T1 and T2 weighted MRI volumes [36]). Moreover, challenges like missing data, varying scaling factors and densities, variation due to different viewpoints, noise and outliers are considered difficulties confronting both unimodal and multimodal registration, and thus will not be included.

The spectrum of modalities that need to be aligned is large. In general purpose registration, the most popular modality in two dimensions is the 2D image and in three dimensions the 3D point cloud and 3D mesh. The 2.5D RGB-D image (i.e. 2D color image plus depth) is also a common modality; such images are often referred as being 2.5D since they are essentially an image with depth information per point. A variety of modalities are derived from medical imaging applications. Anatomical images such as ultrasound (US), X-ray, magnetic resonance (MR) and computed tomography (CT) expose the structure of entire areas. Functional images like single-photon emission computed tomography (SPECT) and positron emission tomography (PET) show the physiological activity of certain body areas. Some of the most common data representations for 3D and 2D data (the most common dimensionalities) are:

- 3D Data
  - 3D point clouds
  - 3D meshes
  - 2.5D RGB-D images
  - Computed Tomography (CT) scans
  - Magnetic Resonance Imaging (MRI) scans
  - Single Photon Emission Tomography (SPECT) volumes
  - Positron Emission Tomography (PET) volumes
- 2D Data
  - Images
  - Points
  - X-rays
  - Ultrasounds (US)

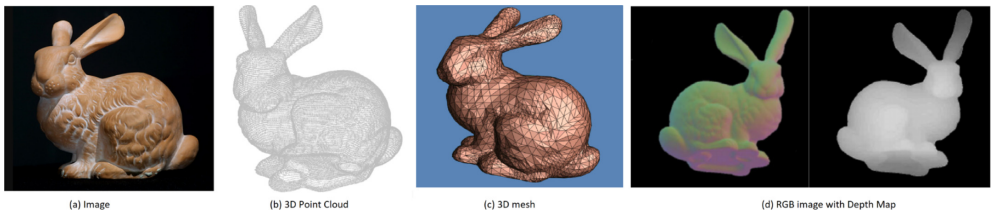


Fig. 2. The Stanford Bunny in Different Modalities as presented in [37–40].

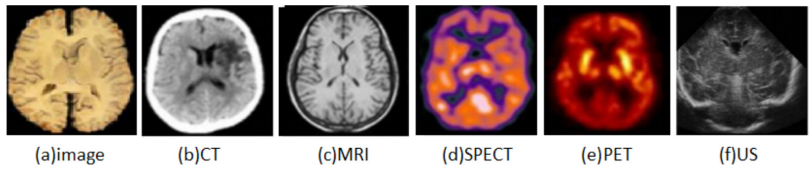


Fig. 3. Different Modality Representations of Brain Anatomy [44].

- 2D slices of a 3D volume (i.e. slice of CT)
- Painting
- 2D Projections of 3D models

Figs. 2 and 3 present examples of different modalities for the Stanford bunny and brain anatomy respectively.

Multimodal 2D/3D Registration

The most common case of multimodal registration across different dimensions is 3D to 2D, e.g. 3D mesh to 2D image. Thus, the problem can also be found with the terms model-to-image or volume-to-slice registration [41]. This is a challenging task with a variety of applications. Its complexity arises from both the different dimensionality and different visual sensors that the data are obtained from, but also from differences in structure, format, and noise characteristics of the data.

The aim of registering a 3D model against a 2D image is to localize the acquired image in the 3D scene and/or to compare the two. Another aspect of the 2D/3D registration problem is the camera localization problem: estimating the pose of a calibrated camera that produces the 2D image, from 3D-to-2D point correspondences between a 3D model and the 2D image. 2D/3D registration can be solved by aligning the visual correspondences extracted from the 3D model and the 2D image. A set of correspondences is usually obtained from features which are extracted from both data models and matched. When the set of correspondences is known, the problem is the well studied perspective-n-point (PnP) problem [42]. However, more challenging is when the correspondences are not known, and the registration method needs to find simultaneously the correspondences and the pose of the data. This review is focused on algorithms for solving the more challenging problem of the correspondence-free registration; for more details on the PnP problem, we refer the reader to a recent survey on the topic [43].

3. Applications of multimodal 3D registration

Multimodal 3D registration has proved vital to many *applications* as well as generalized *operations* within multiple application areas.

By far the largest application area is *medical imaging* where CT, MRI, 3D Rotational X-ray and other modalities are used [45–47]. Clinical practice can benefit from the integrated visualization and analysis of different modalities of the same anatomy in order to make the diagnostic and treatment process more efficient. Multimodal registration is an essential tool in image-guided minimally invasive therapy, image-guided radiation therapy and image-guided surgery [41], to name a few. The different modalities involved, such as CT and MRI are based on different physical principles and capture complementary but non-overlapping information. By fusing the different modalities, all related information can be presented in a consistent way, in order to ease the functional analysis and diagnosis and obtain complete information about the patient [48,49] (Fig. 4). Furthermore, multimodal registration is an important step in the majority of computer-aided surgery (CAS) systems, where the main goal is to align pre-operative and intra-operative data sets so that they can be used in the operating room for image-guided navigation and robot positioning.

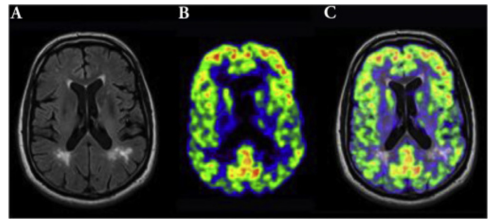


Fig. 4. Medical fusion of MRI and PET modalities. (A) MRI and (B) PET images are registered and fused (C) [50].

Another important application domain is *cultural heritage*. Here multimodal 3D registration is used in visualization, where 2D and 3D sensing modalities are combined (e.g. multispectral images and 3D models) [8,10]. Also in the reconstruction of 3D models from range and color images which must be aligned with the 3D mesh/point cloud derived from 3D scanning; this is applied to digital preservation [51], restoration [52], or to create Virtual Reality (VR) environments (e.g. a museum for multimedia exhibitions or a historical building) [53,54].

Other application areas include *remote sensing* where aerial or satellite data are registered onto maps and *urban mapping* where accurate registration between panoramic images, laser scanning data (LiDAR) or radio detection and ranging (Radar) is crucial for autonomous navigation [55–57], 3D building and terrain modelling [58], 3D city change detection [59], etc.

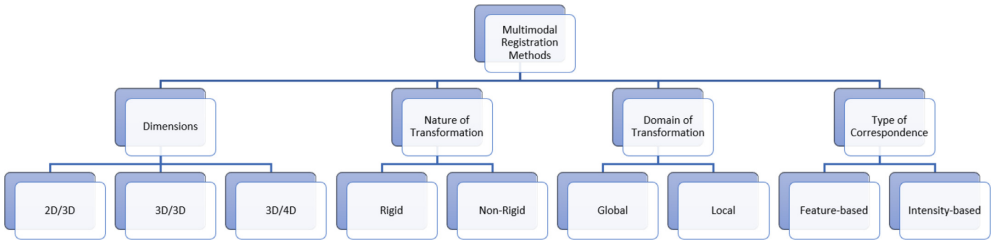


Fig. 5. Attributes of Registration Methods.

Generalized operations that exploit multimodal 3D registration include *3D object retrieval* with the query being of different modality to the 3D object gallery [1,60,61], the *visualization of big multimodal datasets* [13,62], *object recognition* [9,63,64], *motion segmentation* [65] and *camera localization* [66] and tracking [67–69].

4. Registration attributes

In the vast literature of registration methods, some attributes can be identified that characterize such methods. Earlier schemes used subsets of these attributes to classify registration algorithms [11,70,71]; we diverge by proposing a classification mainly based on their algorithmic strategy, see Fig. 5.

Dimensionality

Based on the dimensionality of the data involved, registration techniques can be distinguished into 2D/2D, 2D/3D and 3D/3D. An exhaustive amount of research has been conducted on 2D/2D registration of two images or slices taken from 3D volumes (e.g. slices from tomographic datasets). 3D/3D registration techniques most commonly involve the registration of 3D point clouds or meshes. 3D/3D registration has many applications in medical imaging where most of the modalities used for alignment are 3D volumes. A special case of registration is 2D/3D registration or, as it is known in the medical imaging community, 'slice-to-volume' alignment. 4D image registration is the process of aligning sequences of 3D images, i.e. 3D meshes or point clouds across time (3D+t). 4D image registration is utilized in medical health treatments [72].

Nature of Transformation

Registration techniques usually fall into two categories: rigid or non-rigid, depending on the underlying transformation model. Rigid approaches assume a rigid environment such that the transformation can be modeled using only 6 Degrees of Freedom (6DOF), i.e. translations and rotations only. If the objects can be of different shape or deformable, then non-rigid transformations are used. Non-rigid methods can cope with articulated objects or soft bodies that change shape over time.

Domain of Transformation

Two types of registration algorithms can be recognized based on the proportion of data that is used during the registration process. An algorithm is global if it applies to the entire data set (image, voxels, etc.) and local if registration is applied to only a part of the data set.

Type of Correspondence

Recognizing the correspondence between the datasets is crucial for any registration technique. As correspondence we refer to the explicit relation between parts of the data (elements), structure or context. According to the type of correspondence, registration methods may be feature-based or intensity-based. Feature-based methodologies extract feature correspondences based on local appearance and utilize them to determine the misalignment between datasets. Intensity-based methodologies try to identify context similarity between the datasets by utilizing a similarity metric that is a function of the transformation parameters and then search the extrema of this function.

1. **Feature-based Registration** methods aim to find the transformation that minimizes the distance between the features extracted from the datasets to be aligned. The features are geometrical entities, with the most commonly used ones being points, lines or contours. Due to the significant differences between multimodal datasets, it is non trivial to detect features that are common across different modalities.
2. **Intensity-based Registration** utilizes statistical intensity patterns within the datasets to compute similarity. These methods are based on the assumption that the datasets will be most similar at the optimal alignment. The main goal is to define a measure of intensity similarity between the datasets and adjust the transformation until the value of the measure is maximized. Commonly used similarity metrics that perform well in unimodal registration (e.g. Mean Squared Difference (MSD), Normalized Correlation (NC)), do not give the same results in the multimodal case. For multimodal registration, statistical similarity measures based on minimizing the distance between intensity probability distributions give better results. Mutual information (MI) and Normalized Mutual Information (NMI) are the most popular metrics due to their robustness, accuracy and universality. **Mutual information (MI)** [73,74] is considered as the gold standard similarity measure for multimodal alignment. It is a statistical measure of similarity between two sets of data, which measures the mutual dependence of the underlying image intensity distributions by catching the non-linear correlations between them. MI assumes that the co-occurrence of the most probable values in the two datasets is maximized when they are aligned. **Normalized Mutual Information (NMI)** improves the robustness of MI by avoiding some mis-registrations by being independent of overlapping areas of the two datasets. An interesting use of NMI was proposed by Zhao et al. [75] who used similarity measurements between a chosen set of 2D/3D attribute-pairs which could be dominant in a specific scene. The method has a preliminary training phase where the attribute-pairs are chosen and then combined into NMI. Other variations of MI have been applied for multimodal registration of urban scenes, like Weighted Normalized Mutual Information (WNMI) [76] and Normalised Combined Mutual Information (NCMI) [77].
   The **Mutual Correspondence (MC)** approach, proposed by [78], combines sparse correspondences and Mutual Information (MI)

measures. Mutual Correspondence is simply defined as the weighted sum of the average distance in pixels between the 2D image point and the corresponding 3D point projected in 2D, and the MI. The method combines the correspondence based method with Mutual Information maximization in order to benefit from both, be robust and flexible but also automatic and fast.

5. Public datasets and performance evaluation

5.1. Public datasets

Techniques tested on the same datasets can be compared more reliably. However, the lack of a 'golden standard' large-scale publicly available multimodal dataset makes the comparison of the state-of-the-art approaches non-trivial. In recent years, there has been some progress towards the creation of benchmark multimodal datasets, as outlined below.

KITTI Vision Benchmark [79]: This dataset contains scan sequences of different objects and was presented in 2013 [75,80]. Five different object categories are defined and 3D range scans, as well as 2D images, are provided for each frame of a sequence. The 2D images are stored in PNG [81] format while the 3D range scans as binary float matrices (BFM).

Data61/2D3D Dataset [82]: Data61 / 2D3D dataset was introduced in 2015 [83] and consists of a series of 2D panoramic images (in TIFF format) with corresponding 3D LIDAR point clouds (in LAR [84] format). There are ten outdoor scenes, each of which includes a block of 3D point clouds together with several panoramic images. The number of 3D points in the scenes varies from 1 to 2 million, and each scene is accompanied by 11 to 21 panoramic images.

RGB-D 7-Scenes Dataset [85]: This dataset was introduced in 2013 [86]. It involves 7 different indoor scenes given as RGB-D images. The extracted images are in PNG format. Each scene was captured using an RGB-D Kinect camera with 640x480 resolution. The scenes were recorded in several sequences each one containing from 500 to 1000 frames. The dataset provides a dense 3D model per scene in TSDF format [87] and the 'ground truth' was obtained by an implementation of the KinectFusion system [88,89].

Cambridge Landmarks Dataset [90]: This dataset was created in 2017 and contains the 3D models of 6 Cambridge University landmarks [91]. The data for each landmark includes its 3D model and a number of corresponding images from different points of view. The images are in PNG format while 3D reconstructions are stored in NVM [92] format.

Stanford 3D Scanning Repository [37]: It contains nine different objects as 3D models captured either by various 3D scanners or by the XYZ-RGB [93] auto-synchronized camera. The data are stored in the form of PLY [94] files. There are a variable number of scans for each model. The dataset also contains 2D photographs of selected models along with CT scans of the famous Stanford bunny. It was initially constructed in 1996 [87,95,96] but was further enhanced in 2003 [97].

BrainWeb [98]: The BrainWeb dataset consists of 3D brain volumes (MRI scans) of 270 simulated subjects and was introduced back in 1997 [99]. There are three different MRI image sequences (T1-w, T2-w, and PD-w) for healthy as well as subjects with Multiple Sclerosis. The technical characteristics of the produced sequences (slice thickness, noise) are determined by the user. The data are given in MINC [100] format.

NLM-NIH-VHP [101]: The National Library of Medicine (NLM) Visible Human Project (VHP) is a dataset containing complete, anatomically detailed, 3D Volumes (CT and MRI) and 2D anatomical images of high resolution obtained from one male and one female cadaver [102]. The dataset was introduced back in 1994 for the male and was extended in 1995 for the female. For the male, there are more than 1800 anatomical slices, while for the female there are more than 5000. PNG format is used.

RIRE Dataset [103]: The Retrospective Image Registration Evaluation (RIRE) project delivered a dataset specifically designed to compare 3D volume (CT-MR and PET-MR) registration techniques. The data were acquired from seven different patients and have been available since 2007. It was previously called "Retrospective Registration Evaluation Project (RREP)" [104]. The data format is DICOM [105].

IXI Dataset [106]: The Information eXtraction from Images (IXI) dataset was presented in 2018 [107]. It utilizes 3D volumes of MRI, MRA and Diffusion-Weighted (DW) images in 15 directions. For the data gathering, 600 healthy subjects were recruited. The data is in NIFTI [108] format.

VIPS Dataset: The Virtual Implant Planning System (VIPS) dataset was also introduced in 2018 [109]. It contains a CAD [110] model of a volar plate implant, accompanied by seven X-ray images (in PNG format). Thus, the dataset can be used for applying 2D/3D registration to match the 3D virtual implant with the real one.

SmartTarget Dataset [111]: The SmartTarget [112] is a recent dataset (introduced in 2019) which contains 3D volumes of MRI and US images. The data were recorded from 129 male patients. The initial purpose of this dataset was to compare the two imaging methods for analyzing prostate cancer, but it turned out to be useful for assessing registration methods as well. The data is encoded in the DICOM format.

RESECT Dataset [113]: The RESECT dataset also includes MRI and US scans in the form of 3D volumes. The data were acquired from 23 patients. In addition, anatomical landmarks were identified across US images and between MRI and US. These landmarks can be used to validate image registration algorithms. The dataset was introduced in 2017 [114] and the data is stored in NIFTI format.

Table 1 provides an overview of the aforementioned publicly available datasets.

5.2. Evaluation measures

To evaluate registration methods, one needs to define how accurately two objects coincide after a registration technique has been applied. This can be done by determining the difference between the predicted values of the transformation that the registration method finds and the actual values that are provided by the dataset ground truth. This difference can be computed using a distance measure for the registration error. Several such measures exist in the literature; in general, the lower the registration error is, the better the accuracy of the registration method. Commonly used registration error measures are listed below:

- **Target registration error (TRE)**: measures alignment deviation [115] as the distance of a certain point P under the ground-truth (GT) registration transformation T_{ground} and the estimated registration T_{reg} [116]. Real units (e.g. mm) are often used. Based on the modalities to be registered, methods choose different distance equations, with the Euclidean, Maximum Symmetric (MSD) and Average Symmetric (ASD) being the most common.

$$\mathrm{TRE} = \|T_{reg}(P) - T_{ground}(P)\| \tag{1}$$

- **Mean Target registration error (mTRE)**: is the average distance between the points in the ground truth and the estimated registration. mTRE is calculated by averaging the values of Eq. 1 over all the N points P_i of the dataset.

$$\mathrm{mTRE} = \frac{1}{N}\sum_{i=1}^{N} \|T_{reg}(P_i) - T_{ground}(P_i)\| \tag{2}$$

Table 1
Publicly available datasets for multimodal 3D registration.

Dataset Name	Modality	Data Format	# Subjects	Year
The KITTI Vision Benchmark	2D Images / 3D Range Scans	PNG / BFM	5	2013
Data61/2D3D	2D Images / 3D Point Clouds	TIFF / LAR	10	2015
RGB-D 7-Scenes	RGB-D Images / 3D Models	PNG / TSDF	7	2013
Cambridge Landmark	2D Images / 3D Models	PNG / NVM	6	2017
Stanford Scanning Repository	3D Models/ CT scan / 2D images	PLY	9	1996 2003
BrainWeb	3D Volume MRI/2D slices	MINC	270	1997
NLM-NH-VHP	3D Volume MRI, CT / 2D Images	PNG	2	1994 - 1995
RIRE	3D Volume CT-MR and PET-MR	DICOM	7	2007
IXI	3D Volume MRI, MRA and DW	NIFTI	600	2018
VIPS	2D Images / 3D Models	PNG / CAD	1	2018
SmartTarget	3D Volume MRI and US	DICOM	129	2019
RESECT	3D Volume MRI and US	NIFTI	23	2017

- **Mean Target Registration Error in the projection direction (mTREproj)**: is used when registration is between 2D and 3D modalities; it is the mean distance between re-projected 3D points P_i into 2D [46]. mTREproj is computed as the average across all points of the angle between the displacement vector and the normal to the projection plane $\hat{n}$.

$$\text{mTREproj} = \frac{1}{N}\sum_{i=1}^{N} \| (T_{reg}P_i - T_{ground}P_i) \cdot \hat{n} \| \tag{3}$$

- **Root Mean Square Distance (RMSD)**: is a measure of the average distance between two or more structures. It measures the similarity between the after-registration transformation parameters and the transformation that is provided from the ground truth data.

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \| (T_{reg}P_i - T_{ground}P_i) \|^2} \tag{4}$$

- **Dice similarity coefficient (DSC)**: is a spatial overlap index and is a useful evaluation measure between volumes where the ground truth data is unknown. DSC ranges from 0, indicating no spatial overlap between the two datasets, to 1, indicating complete overlap and thus a successful registration. Given two datasets X, Y to be registered, the DSC is defined as in Eq. 3, where $|X|$ and $|Y|$ refere to the cardinalities of the respective datasets [117].

$$\text{DSC} = \frac{2|X \bigcap Y|}{|X| + |Y|} \tag{5}$$

- **Success Rate (SR)**: is defined as the overall percentage of successful registrations. As successful is considered a registration which has a registration error below a certain threshold. The success rate can be determined using various registration error measures, with mTRE being the most popular. According to the application and the modalities involved, each method defines an explicit criterion for measuring the success rate.
- **Failure Rate (FR)**: is defined as the percentage of aligned cases having registration error greater that a certain value. In [118] the FR is calculated as the proportion of cases with TRE greater than 10mm.
- **Convergence Rate (CR)**: is defined as the range of starting positions from which an algorithm finds a sufficiently accurate registration transformation [46]. It is defined as the number of initial guesses that converge to a success relative to the total number of initial guesses. A method with high CR is generally more efficient, as it converges quickly to correct transformations.

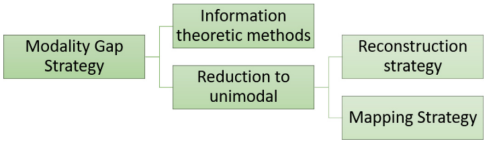


Fig. 6. Modality Gap Strategies.

6. Multimodal 3D registration techniques

Dealing with data from different modalities is a challenging task due to the lack of a general rule for measuring similarity across different modalities. There have been two main approaches to bridge the multimodality gap [11]: (a) use of information theoretic measures, and (b) reduction to a unimodal registration problem by reconstructing one modality to the other or by mapping both modalities to another common representation (Fig. 6).

Information theoretic approaches try to use statistical measures, like MI or NMI in order to identify similarity across modalities and maximize their statistical dependency to achieve registration [74]. Alternatively, there are methods that instead of finding correspondences between the different modalities, try to simplify the multimodal registration into unimodal, and then solve it with the respective state-of-the-art unimodal techniques [119]. In order to achieve this, two strategies have been followed. The first one converts one modality to the other. The most straightforward such operation is in 2D/3D registration, where the 3D modality is mapped into 2D by projection, or the 2D points are back-projected into 3D space. The other tactic is to map both modalities into a common representation, in an initial step before the registration technique is performed [120].

To solve the multimodal registration problem without prior knowledge of the correspondences, two major algorithmic strategies can be identified: optimization-based and learning-based. In the former case, the value of a function that quantifies the alignment quality between the two datasets is maximized while in the latter case, a neural network is typically utilized to find the best alignment. At the top level, we shall base our categorization on this distinction which is presented in Fig. 7.

6.1. Optimization-based registration

Optimization-based methods iteratively optimize the alignment transformation parameters over a scalar-valued metric function representing the quality of the registration. Particularly for 2D/3D registration, the problem can be subdivided into two subproblems: finding correspondences and estimating the pose (align-

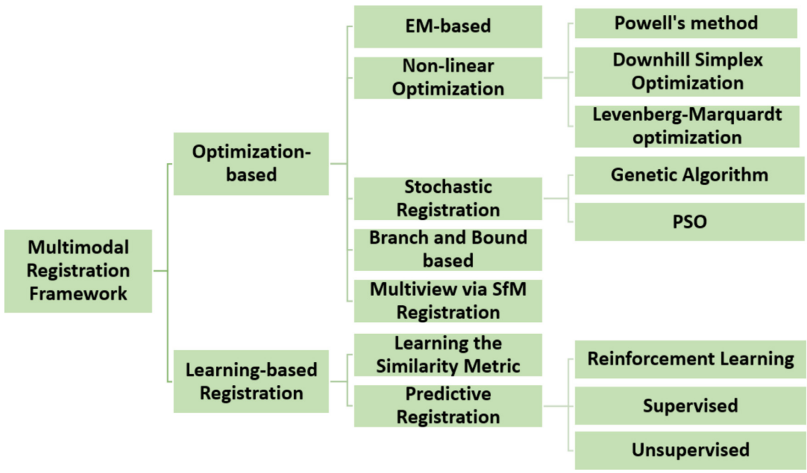


Fig. 7. A classification of presented multimodal registration strategies.

ment transformation) given the correspondences. These two subproblems are intertwined, and the solution of one depends on the other. A mathematical function based on the transformation parameters is optimized using an optimization technique. Optimization plays a fundamental role in registration because it determines the accuracy, robustness and convergence. We therefore further classify optimization-based registration methods in the subsections below based on the optimization technique that they use. Table 2 provides an overview of optimization-based multimodal 3D registration methods.

6.1.1. Expectation-Maximization (EM)-based registration

EM-based Registration is the most popular methodology for multimodal registration and is a local deterministic method which attempts to find the best alignment with an iterative optimization strategy. It starts from an initial solution (a guess/computation of pose/point correspondence) and iteratively tries to find a solution that optimizes an objective function locally. Although such methods are generally accurate, they depend on initialization in order to converge to the best solution and finding the global minimum cannot be guaranteed. One more limitation of these methods is their heavy computation cost.

An early solution to the 2D/3D registration problem is proposed by Beveridge [163], where a random-start local search procedure is used to arrive at a local optimum. The method uses a hybrid pose estimation algorithm with both full-perspective and weak-perspective camera models. The weak-perspective pose algorithm ranks neighbor points in the search space and the full-perspective pose algorithm updates the object's pose after moving to a new set of correspondences. The authors investigated how easy this problem is by evaluating expected run-time as a function of the number of lines and the amount of clutter. A more restrictive approach was proposed by Christmas et al. [168], where the detected lines are viewed as edges on a graph, leading to a graph matching problem. However, using a graph structure cannot guarantee an optimal registration for 2D/3D registration.

The most effective algorithm to solve the correspondence-free registration problem is the SoftPosit algorithm [142], which is one of the best approaches to correspondence-free registration using points. It locally searches the transformation space while simultaneously determining the correspondences between the 2D and 3D points. At each iteration, it first uses the SoftAssign technique to determine the point correspondences [169]; multiple weighted correspondences are hypothesized based on the pose. Then, the Posit [170] algorithm is used to iteratively estimate the pose. The SoftPosit algorithm stands out due to its accuracy, but it cannot guarantee a global minimum and tends to fail in the presence of large amounts of clutter, occlusions or repetitive patterns. Moreover, it is quite slow because it needs to randomly try hundreds of different initial poses.

An extension of the SoftPosit algorithm with line features was proposed by David et al. [164]. The method is iterative and, in each step the given 2D to 3D line correspondence problem is mapped to a new 2D to 3D point correspondence problem and the SoftPOSIT algorithm is utilized to find the registration parameters. In [143] the same authors assumed that all lines are orthogonal in order to speed up the algorithm in high-clutter environments.

More recently, Dong et al. presented an iterative algorithm inspired by SoftPosit, named SoftOI [152]. Like SoftPosit, the SoftAssign algorithm [169] is used for determining the correspondences, but for computing the pose another pose estimation algorithm, named OI (Orthogonal Iteration) [171], is employed. The SoftOI algorithm first introduces an assignment matrix that describes the correspondences for the OI algorithm. The pose and correspondences are then evolved iteratively from an initial pose to an optimum value by minimizing the objective function based on the weighted object space collinearity error and by applying a deterministic annealing technique. The method exhibits efficiency and accuracy even in cases with occlusions.

Moreno-Noguer et al. proposed another Expectation-Maximization algorithm, the BlindPnP [119], where local optimality is alleviated in each iteration. The method models an initial set of poses as a Gaussian mixture model from which a Kalman filter is initialized and progressively refined by hypothesizing correspondences. Each new candidate is incorporated in a Kalman filter, which reduces the number of potential 2D matches for each 3D point and makes it possible to search the pose space sufficiently fast. Eventually, the method determines a solution with high con-

Table 2
Overview of Optimization-based Registration Methods, grouped by evaluation measure and dataset used.

Optimization-based Registration Methods												
Method	Modality A	Modality B	Nature of Transform.	Domain of Transform.	Type Of Correspondence	Modality Gap Strategy	Optimization-based Strategy	Dataset	Initial Application	Evaluation Measure	Value of Eval.Measure	Execution time (sec)
Parmehr et al [77].	3D model (LIDAR point cloud)	2D image (aerial photograph)	rigid	local	intensity: NCMI	mapping one modality to another	intensity-based distance optimization	private	urban navigation	TRE	0.12m - 0.0051°	n/a
Sottile et al [78].	3D model	2D image	rigid	local	intensity: MC	mapping one modality to another	intensity-based distance optimization	private	general	TRE	4.8pixels	2sec
Wachowiak et al [121].	3D volume MRI	2D US	rigid	global	intensity: NMI	mapping one modality to another	Stochastic/ HPSO	BrainWeb [98,122], NLM-NIH VHP[101]	medical	TRE	2.36mm	350sec
Wachowiak et al [121].	3D volume MRI	2D CT	rigid	global	intensity: NMI	mapping one modality to another	Stochastic / HPSO	BrainWeb [98,122], NLM-NIH VHP[101]	medical	TRE	2.14	230sec-500sec
Schwab et al [116].	3D volume MRI	3D volume CT	rigid	global	intensity: NMI	learning multimodal similarity measure	Stochastic / PSO	RIRE [104]	medical	TRE SR	9.57mm 78%	n/a
Chen et al [123,124].	3D volume MRI	3D volume CT	rigid, non rigid	global		learning multimodal similarity measure	Stochastic /HPSO	RIRE [104]	medical	TRE	2.36mm	n/a
Lin et al [125].	3D volume MRI	3D volume CT	rigid, non rigid	global		learning multimodal similarity measure	Stochastic /HPSO	RIRE [104]	medical	TRE	2.36mm	1893.637sec
Liu et al [126].	3D model	2D image	rigid	global	features: points	mapping one modality to another	BnB	[52]	general	TRE SR	14.18mm - 1.55° 81%	40sec-200sec
Corsini et al [120].	3D model	2D image	rigid	local		reconstruction modality strategy	Multiview with SFM	[127]	cultural heritage	TRE	10.92cm - 0.27°	21600sec
Pintus and Gobbetti [130]	3D model	2D image	rigid	global	features: points	reconstruction modality strategy	Multiview with SFM	[128,129]	cultural heritage	TRE	3.19cm - 0.26°	1140sec-24960sec
Klima et al [131].	3D volume CT	2D x-rays	non rigid	local	intensity:NMI	mapping one modality to another	NL / LM method	private	medical	mTRE	1.23mm	3.19sec-15.77sec
DePose [132]	3D model	2D image	rigid	global		mapping one modality to another	Stochastic / GA	private	general	mTRE SR	0.6cm - 1.0° 75%	1.25sec-1.99sec
EvoPose [133]	3D model	2D image	rigid	global		mapping one modality to another	Stochastic / GA	private	general	mTRE SR	1.28 cm - 2.2 ° 25%	0.68sec-4.11sec
Crombez et al [134].	3D model	2D image	rigid	global	intensity: MI	mapping one modality to another	Stochastic / PSO	private	general	mTRE	6.5cm-0.61°	n/a
Toth et al [135].	3D volume MRI	2D x-rays	rigid	global		reconstruction modality strategy	BnB	private	medical	mTRE	3.87 ± 1.22mm	95.24sec
Wang et al [136].	3D volume	2D x-rays	rigid	global		mapping one modality to another	intensity-based distance optimization	[52,137]	medical	mTRE SR	0.17mm 94.68%	n/a

(*continued on next page*)

Table 2 (*continued*)

Optimization-based Registration Methods												
Method	Modality A	Modality B	Nature of Transform.	Domain of Transform.	Type Of Correspondence	Modality Gap Strategy	Optimization-based Strategy	Dataset	Initial Application	Evaluation Measure	Value of Eval.Measure	Execution time (sec)
Schaffert et al [138].	3D volume	2D x-rays	rigid	global		mapping one modality to another	Multiview with SFM	[139]	medical	mTRE	0.22mm	
[140,141]										SR	98.4%	7.0sec-35.0sec
SoftPosit [142]	3D model	2D image	rigid	local	feature: points, lines	mapping one modality to another	EM-based	private	general	SR	75%	36sec
David et al [143].	3D model	2D image	rigid	local	feature: lines	mapping one modality to another	EM-based	private	general	SR	70%	72sec-100sec
Mastin et al.	3D model (LIDAR point cloud)	2D image (aerial photograph)	rigid	local	intensity: joint entropy	mapping one modality to another	NL / Downhill Simplex	private	urban navigation	SR	98.5%	6.50sec-15.0sec
Parmehr et al [76].	3D model (LIDAR point cloud)	2D image (aerial photograph)	rigid	local	intensity: WNMI	mapping one modality to another	intensity-based distance optimization	private	urban navigation	SR	92%	n/a
Enqvist et al [144].	3D model	2D image	rigid	global	features: points	mapping one modality to another	BnB	[145]	general	SR	96%	2sec-4sec
Brown et al.	3D model	2D image	rigid	global	features:	mapping one modality	BnB	[146,147]	general	SR	25%	500sec-1000sec
[148,149]					points, lines	to another						
GOPAC [150]	3D model	2D image	rigid	global		mapping one modality to another	BnB	DATA61/2D3D [83]	general	TRE SR	2.30m - 2.08° 82%	477sec
BlindPnP [119]	3D model	2D image	rigid	local	feature: points	mapping one modality to another	EM-based	private	general	CR	65%	20sec-100sec
Sanchez et al [151].	3D model	2D image	non rigid	local	feature: points	mapping one modality to another	EM-based	private	general	CR	90%	600sec-1500sec
SoftOI [152]	3D model	2D image	rigid	local	feature: points	mapping one modality to another	EM-based	private	general	CR	75%	10sec-60sec
Corsini et al [153].	3D model	2D image	rigid	local	intensity:MI	mapping to a common space	NL / Powell's method	private	cultural heritage	CR	80%	4sec
Palma et al [154].	3D model	2D image	rigid	local	intensity:MI	mapping to a common space	NL / Powell's method	private	cultural heritage	CR	70%	n/a
Yang et al [155].	3D model	2D image	rigid	global		mapping one modality to another	Stochastic / GA	private	general	CR	97%	20sec-39sec
Marques et al.	3D model	2D image	rigid		feature: points	mapping one modality to another	NL / Linear Regression	private	general	FS	25%	n/a
Enqvist et al [156].	3D model	2D image	rigid	global	features: points	mapping one modality to another	BnB	[157]	general	FS	20%	5sec-15sec

(*continued on next page*)

Table 2 (*continued*)

Optimization-based Registration Methods											
Method	Modality A	Modality B	Nature of Transform.	Domain of Transform.	Type Of Correspondence	Modality Gap Strategy	Optimization-based Strategy	Dataset	Initial Application	Evaluation Measure	Value of Eval.Measure	Execution time (sec)
Kisaki et al [158].	3D volume CT	3D volume MRI	rigid	local	intensity:NMI	mapping one modality to another	NL / LM method	private	medical	MI	0.294	n/a
Talbi et al [159].	3D volume MRI	3D volume CT	rigid	global		learning multimodal similarity measure	Stochastic / HPSO	private	medical	MI	0.6349	n/a
Talbi et al [159].	3D volume MRI	3D volume SPECT	rigid	global		learning multimodal similarity measure	Stochastic / HPSO	private	medical	MI	0.6789	n/a
Talbi et al [159].	3D volume MRI	3D volume PET	rigid	global		learning multimodal similarity measure	Stochastic / HPSO	private	medical	MI	0.6431	n/a
Khoo and Kapoor [160]	3D model	2D image	rigid	global		mapping one modality to another	NL / Convex	[37], private medical datasets	medical	RMSD	6.9mm	n/a
Ayatollahi et al [161].	3D volume MRI	3D volume CT	rigid	global	intensity: MNMI	learning multimodal similarity measure	Stochastic/ HPSO	[162]	medical	RMSD	44%	n/a
Zhao et al [75].	3D range scans	2D image (aerial photograph)	rigid	local	intensity: CMI	mapping one modality to another	intensity-based distance optimization	KITTI [80]	urban navigation	projection error	14%	n/a
RANSAC [67]	3D model	2D image	rigid	local	feature: points	mapping one modality to another	Stochastic	private	general	n/a	n/a	3600sec-36000sec
Beveridge et al [163].	3D model	2D image	rigid	local	feature: lines	mapping one modality to another	EM-based	private	urban navigation	n/a	n/a	n/a
David et al [164].	3D model	2D image	rigid	local	feature: lines	mapping one modality to another	EM-based	private	general	n/a	n/a	100sec
SoftSI [165]	3D model	2D image	rigid	local	feature: points	mapping one modality to another	EM-based	private	general	n/a	n/a	0.6sec-10.01sec
Pan et al [166].	3D Volume(CT/MRI)	2D x-rays	rigid	global		mapping one modality to another	BnB	private	medical	n/a	n/a	4.12sec-12.09sec
Zhao et al [167].	4D video	3D point cloud		local	features: points	reconstruction modality strategy	Multiview with SFM	private	general	n/a	n/a	n/a

fidence. The authors also introduced priors on the camera pose, for example the camera is always above the ground and pointing towards the object. The BlindPnP algorithm outperforms SoftPosit when large amounts of clutter, occlusions and repetitive patterns exist. However, it is susceptible to local optima, requires a pose prior and cannot guarantee global optimality.

Sánchez-Riera et al. proposed a solution [151] inspired by Moreno-Noguer's method for rigid object pose estimation and extended it to non-rigid objects. The method uses weak priors on pose and shape, that have been learned from training data, and models them as Gaussian Mixture Models. These priors can define a region in the image where the algorithm searches for the potential 2D candidates that may be assigned to each 3D point. Using a Kalman filter strategy (as also done by BlindPnP) this search region is progressively shrunk while the estimation of the pose and shape are refined.

The SoftSI algorithm [165] is based on minimizing a global objective function, like SoftPosit, but is based on the combination of two singular value decomposition (SVD)-based shape description theorems, and the PnP algorithm proposed in their paper (SI). Due to the use of the SI algorithm, the method avoids pose ambiguity and quickly eliminates bad initial values, according to the standard deviation of the translation vector at the first iterations. The method is fast and robust to noise, but assumes no occlusion or clutter.

6.1.2. Non-Linear (NL) optimization

Several non-linear optimizers have been applied to the registration problem, such as Powell's method, downhill simplex and the LevenbergMarquardt algorithm.

Corsini et al [153] took inspiration from medical imaging and extended the use of MI to a generic image registration case, in particular to align a 3D model to a given image for Cultural Heritage applications. The main idea is to use different renderings of the 3D model and then align them with a grey-scale version of the input image. The similarity measure that the method uses is mutual information (MI), where the camera parameters are iteratively optimized using **Powell's method** [172] by maximizing the correlation between a real image and different attributes of illumination of the 3D model (i.e. ambient occlusion, specularity, normal field). The approach is robust and fast, but the global minimum of the registration may be different from the best solution. An improvement on [153] was proposed by Palma et al. in [154] for aligning 2D real images with a rendering of a 3D model. The method computes the gradient map of the 3D rendering and the gradient map of the image and, within an iterative optimization algorithm, it tries to maximize their MI until registration is achieved. The method increases the performance and the quality of the original technique.

Mastin et al. [173] introduced the use of MI for registering urban scenes of LiDAR 3D point clouds and aerial imagery. In each iteration, the algorithm renders 3D points that are projected onto the image plane and then uses the **downhill simplex optimization** scheme [174] for maximizing a mutual information metric. The authors proposed three metrics for measuring mutual information between LiDAR and optical imagery in urban scenes, with the most promising being the one that measures the joint entropy among optical image luminance, LiDAR depth information and LiDAR probability of detection values.

In the field of medical model reconstruction, [131] proposed a new automatic image registration method between 3D CT and 2D X-rays. The registration is formulated as a non-linear least squares problem, and is then solved with the **Levenberg-Marquardt (LM) optimization algorithm**. Kisaki et al. [158] performed registration in 3D CT and MRI volumes by applying a global matching method based on Levenberg-Marquardt. The method consists of two steps, a coarse registration based on the proposed similarity criterion named ratio image uniformity (RIU); RIU measures the deviation and a fine registration based on the maximization of normalized mutual information (NMI).

The above methods have modelled the similarity measure as a convex function and then utilize optimization algorithms to find the optimum. Khoo and Kapoor [160] proposed a methodology to convert a **non-convex** function into a convex one in order to obtain global optimality when the correspondences are unknown. Their framework formulates the 2D/3D registration problem as a mixed-integer nonlinear programming problem and relaxes it to a convex semi-definite problem that can be solved efficiently by the interior-point method. The algorithm solved simultaneously the pose and correspondence problems. However, only the rotation is recovered and the method achieved superior results only when there is no noise, which is an unrealistic assumption for most applications. Marques et al [175] viewed the problem as an instance of correspondence permutation, which they solved by a convex relaxation procedure. Their method considers the noiseless observation model and shows that if the permutation matrix maps a sufficiently large number of positions to themselves, then the solution matrix can be recovered. However, the algorithm assumes that no outliers are present, which is unreasonable in most scenarios.

6.1.3. Stochastic registration

Another approach similar to hypothesize-and-test considers all possible correspondences, and then searches the parameter space to find the best solution. Different to the EM-based logic, in each iteration a hypothesis correspondence set is generated and tested; the heuristic algorithms generate most likely correspondences and then try to find the optimal solution within the search space. As exhaustive search is infeasible [176], most strategies search the parameter space more efficiently; genetic algorithms [155], differential evolution algorithms [132] and pose clustering are examples. When prior pose information is provided, they are more robust to occlusions, clutter [177] and repetitive patterns [119]. Stochastic optimization methods produce solutions closer to the global optimum and can be applied efficiently in cases with noise.

A traditional approach to 2D/3D registration is the hypothesis-and-test **RANSAC** algorithm [67]. RANSAC is a re-sampling technique that randomly selects a small set of 2D/3D correspondences, estimates the transformation parameters and verifies the transformation against the rest of the features. If the original and the transformed image features are sufficiently similar, the pose is accepted, otherwise a new correspondence set is hypothesized and the process is repeated. As pointed out by Fischler and Bolles [67], RANSAC uses the smallest data set possible and proceeds to enlarge this set with consistent data points. RANSAC inspired a wide variety of registration methods, mainly in deep-learning field for multimodal registration.

Genetic (or Evolutionary) Algorithms (GA) [178] are a class of widely used parallel search methods that solve complicated global optimization problems, so they are also deployed to correspondence-free 2D-3D registration. GAs simulate the natural evolution process in which the stronger individuals are most likely to survive in a competitive environment. They maintain a population of possible solutions (called individuals) and in each iteration an evolutionary procedure is performed until some criteria are satisfied. In the iterative evolutionary procedure, each individual is assigned a measure of quality and those with the best scores are selected for reproduction in order to generate a new population. Generation after generation, the solutions approach the optimum. Genetic Algorithms are simple, effective and do not need a good initial alignment in order to guarantee a result of good quality, but searching over the pose space is generally expensive.

Rossi et al. proposed an evolutionary based procedure called EvoPose [133]. The authors formulated the pose estimation prob-

lem as an optimization problem and solved it with a Genetic Algorithm, enhanced with heuristic rules in order to improve convergence. EvoPose constructs an objective function of reprojection errors according to the perspective projection model, and in each iteration the population with the minimum mean distance between the model and the image is selected to be evolved. The algorithm converges to a good pose solution after some generations. EvoPose has low computational cost and its performance is comparable to the SoftPosit method [142].

Inspired by EvoPose [133], Xia et al. proposed a Differential Evolution based solution for the model-to-image registration problem without any correspondence information. The method is called DePose [132] and enhances the evolutionary algorithms with a new efficient scheme called "DE/bests/I". The candidate solution is evolved only when the offspring outperforms its parent, so the survival probability of good pose offspring is increased. DePose was compared to EvoPose and outperformed it in accuracy and robustness. Although, both methods improve the convergence rate, they tend to be slow and converge to false solutions due to local minima, especially when missing or false image points exist.

Yang et al. used the Genetic Algorithm methodology for determining the initial pose of 3D objects from 2D images [155]. The authors state that a good initial guess is necessary in order for the global optimum to be reached and for the objective function not to fall into local optima. This is because when the initial guess is selected randomly, the relationships between each guess are neglected, so an appropriate initial correspondence may not be selected in a long time if there are many local optima. Also, a correspondence may be randomly selected even if a similar one has already been selected and discarded, which leads to extra iterations. In this method, the initial pose is calculated based on GA and then an iterative method is used to solve the registration by minimizing a global objective function. The algorithm first generates a set of random initial guesses and then, for each of these candidate solutions, it computes the assignment matrix and the perspective projection error. The solution with the best result is selected for evolution until convergence. Compared with the traditional random start initialization methods, this technique has higher convergence rate and lower number of iterations.

Particle Swarm Optimization (PSO) is a relatively recent population-based evolutionary computation technique for solving optimization problems, which is inspired by the swarming or collaborative behavior of biological populations [179]. PSO algorithms share many similarities with GAs; they are both population-based search methods and search for the optimal solution by updating generations. However, GAs exploit the competitive characteristics of biological evolution in terms of survival of the fittest, while PSO techniques do not use evolution operators such as crossover and mutation. The PSO strategy emulates the swarm behavior of insects when they search for food in a collaborative manner. Each member in the swarm is referred to as a particle and represents a potential solution. Each particle flies through the search space in an adaptable way (velocity) that is dynamically altered by its own experience and other members' flying experience. So, starting from a diffuse randomly generated population, each particle tends to improve itself by imitating traits of its successful peers. PSO it is an iterative technique, where in each iteration a particle moves by the addition of a velocity vector, which is a function of the best position (position with the lowest objective function value) found by this particle and the best position found so far among all particles. Compared to GAs, PSO techniques seem to perform better and converge to an optimal solution within fewer iterations. However, the PSO computational time increases more rapidly than GAs due to the communication between the particles after each generation. Moreover, the PSO algorithms tend to get trapped into local optima in case of multimodality due to the significant nonlinear intensity differences between multimodal images.

Crombez [134] proposed a robust multimodal 2D/3D registration method that takes advantage of both geometrical and dense visual features instead of trying to develop a new similarity measure. The method uses a PSO approach, where a swarm of virtual cameras moves inside the 3D model and tries to reach a desired pose represented by the 2D image. At each iteration, the virtual cameras move in the direction of the camera with the highest similarity score (based on dense visual features) but their movement is also influenced by the best particle in their nearest neighborhood. The particle velocities updated in this way are expected to iteratively move the swarm towards the best solution.

Wachowiak et al. [121] used the PSO strategy to register single slices of 3D volumes to whole 3D volumes of medical images. They proposed a hybrid particle swarm technique with the addition of GA concepts such as crossover and mutation. The method outperformed the evolutionary strategies that was compared to, both in terms of accuracy and efficiency. However, user guidance is needed in order to position the images in approximately the right vicinity.

Chen and Lin [124] stated that the conventional PSO is efficient for 2D/2D multimodal registration but when transferred to three dimensions cannot find he global optimum efficiently; they thus proposed a hybrid method by integrating two methods from the GAs into the standard PSO algorithm [123,125,180]. The hybrid particle swarm optimization (HPSO) method incorporates subpopulation and crossover from GAs into the conventional PSO. The particles are not standalone, but are divided into a number of subpopulations. Each sub-population has its own best optimum and the PSO process is performed for each sub-population. The optima of each sub-population are sorted and the sub-populations with the top two optima are selected as parents for crossover. The HPSO was used for registering MRI and CT volumes showing better results that classical GA and PSO algorithms.

A similar method was proposed by Ayatollahi et al. at [161] but they introduced two new similarity metrics, Modified Normalized Mutual Information (MNMI) and Logarithmic Normalized Mutual Information (LNMI). Experiments showed that MNMI had better results for multimodal registration than LNMI or MI. Moreover, hybrid registration can be automatic, more accurate, and faster than either of its registration components used separately. However, the results were inaccurate in the presence of large shear distortion between images.

Schwab et al. [116] presented four variants of the PSO algorithm for registering 3D CT and MRI volumes. The first version was the initial standard PSO algorithm [181], the second version was a modification of PSO where the inertia weight monotonically decreases during the iterations, the third and fourth versions utilize external intervention in order to improve the initial orientation. The test results showed that the classical PSO reach their limits for the multimodal 3D registration, but when influence of the initial orientation was introduced the results improved.

Another hybrid scheme of PSO algorithms was introduced by Talbi and Batouche [159]. Different from the above methods that mixed PSO algorithms with GA, this technique integrates PSO with Differential Evolution (DE) operator for registering MRI images with a variety of medical modalities (CT, PET, SPECT). The proposed algorithm follows the classical PSO iterative scheme but the DE operator is applied only to the best particle obtained in each iteration for alternate generations.

6.1.4. Branch-and-Bound (BnB)-based registration

Several optimization-based registration methods use the Branch-and-Bound (BnB) framework due to its theoretical optimality guarantees. Assuming that the correct alignment belongs to a

known volume of the search space, first all correspondences and the transformation space are generated. The search space is recursively subdivided into smaller subsets and is reduced according to lower bounds of the registration error in order to be used for pruning. In the end, the only remaining branch will include the aligned solution. The method depends on how tight the bounds are and how quickly they can be computed. The BnB algorithm forms the transformation space as a decision tree where each node is a possible correspondence and then searches it recursively, bounding the objective function at each stage and discarding parts of the transformation in which the solution does not exist. At the end, the remaining transformation space is tightly bounded and includes the globally optimal solution.

An early algorithm, similar to BnB, was proposed by Jurie [182] for 2D/3D alignment with a linear approximation of perspective projection. First, an initial volume of pose space is guessed and all of the correspondences compatible with this volume are considered. Then the method recursively reduces the pose volume until only a single pose remains. The Gaussian error model is used to calculate the score of each sub-volume (named as box) and in each iteration the sub-volumes (boxes) of pose space are pruned. Thus, boxes of pose space are not pruned by counting the number of correspondences, but based on the probability of having an object model in the image within the range of poses defined by the box. Due to the use of the Gaussian error model, the approach is not robust to outliers.

Enqvist et al. [144,156] formulated the registration problem as a graph vertex cover problem and provided an optimal solution. The algorithm makes use of the observation that any two point correspondences generate a 3D surface of the possible camera positions. The main approach is to compute pairwise constraints between pairs of potential correspondences and employ BnB search over the possible camera positions. The method creates a graph of all possible pairs of correspondences and the optimal solution is found by determining the largest set of pairwise consistent correspondences. Finally, the transformation is computed for the found correspondences.

A method that guarantees the global optimality of the registration in case of both points and lines within indoor scenes has been proposed by Brown et al. [148,149]. The method applies a BnB framework in order perform 2D/3D registration without any correspondence knowledge. In order to increase the efficiency, a nested BnB structure was utilized. An outer BnB searches over the rotation space and, for each rotation branch another BnB algorithm is used for searching the camera position. While the approach is not susceptible to local minima, it requires the inlier fraction to be specified in order to trim outliers, which is rarely known in advance.

Similar to Brown's approach [148], a BnB framework was proposed by Campbell et al. in [150], but they introduced new bounds which are proven to be tighter than those used in Brown's formulation. The authors proposed a globally-optimal inlier maximization framework which maximizes the cardinality of the set of features that are within a set inlier threshold from a projected 3D feature. The authors pointed out that the global optimum of a trimmed objective function may not occur at the true pose, particularly when an incorrect objective function is used. So, the main advantage of the method is that no trimming is necessary, so the estimation of the proportion of inliers is not necessary. Both [149] and [150] formulate the 2D/3D registration problem as a camera pose estimation problem, in which the 3D points are fixed and the optimal camera orientation and position are sought so that the image of the 3D points captured by the camera matches the 2D point set. This formulation, however, has as drawback that in order to cover the whole relative angle space between the 3D points and the camera, the camera position needs to move all around the 3D points, and thus the range of transformation parameters that needs to be searched gets very large.

The idea of the nested BnB structure in order to accelerate the optimization was also utilized for medical registration of MRI and X-rays in [135]. The method generates a 3D model from MRI images and another one by reconstruction from the X-ray images. The two meshes are then registered by using a globally optimal iterative closest points (Go-ICP) method [183]. The method encapsulates two BnB algorithms and the standard ICP in a globally optimized registration technique. The outer BnB algorithm operates on the rotation space and the inner one on the translation space. The ICP algorithm is called when the upper bound is below the current best estimate.

Liu et al. [126] introduced a 2D/3D registration method based on a globally optimal rotation search algorithm utilizing the Branch-and-Bound (BnB) optimization scheme with four new proposed upper bounds in order to make the search of BnB more effective. The problem is formulated in a similar way to a camera pose estimation problem [149,150], but instead of searching for the optimal camera orientation and position with fixed 3D points, the 2D points and the camera's coordinate system are fixed instead. The pose of the 3D points is then searched for as the rigid transformation that best aligns their projections with the 2D points. The method uses as objective function the cardinality of the inlier set of the 2D projection plane and tries to maximize it with a BnB strategy. Moreover, a synchronized searching schema in translation space is proposed; the translation space is divided into a series of blocks, smaller than the covering region of the search algorithm and a rotation search is run at the center of each block in a synchronized way. A search is terminated and the corresponding block is omitted when its upper bound is smaller than the universal best value of the objective function.

Recently, Pan et al. [166] extended the method of [126] into a multi-view setting to make the registration more feasible in real world applications [52,137,139,184]. The method makes full use of different views to accelerate the searching process and reduces the required iterations. The search space is divided into subspaces and each view shares the same branches, but the upper and lower bounds are different. Each view follows the classic BnB pipeline to update its current best upper bound. When one of the views faces the case that the upper bound is lower than the current best, the corresponding branch is pruned. With the introduction of multiple views instead of only one, the accuracy is improved, and the iterations are reduced. However, no experiments have been conducted on real world applications.

6.1.5. Multiview registration using SfM

Multiview geometry can be applied for registering multiple 2D images with a 3D model. The approach is generally divided into three steps, Structure from motion (SfM), rough registration and fine registration. In the first stage, SfM is utilized in order to reconstruct a 3D point cloud from the 2D images. The problem is then simplified to 3D/3D registration, in which the 3D point cloud produced from the first stage and the initial model have different scales, reference frames, and resolutions. Due to the sparseness and noise of the point clouds produced via SfM, the resulting alignment of the second step may be rather approximate, so a final stage is needed to refine the solution. SfM approaches show high registration accuracy and robustness, but are computationally expensive and demand a large collection of images for the SfM reconstruction.

In 2013, Corsini et al. [120] proposed an automatic 2D/3D registration pipeline, which can handle scale changes between datasets. Instead of aligning each single image with the 3D geometry, the method starts with a group of images as an input, taking advantage also of the relations between the images. At the first stage, the

images are used to compute a sparse point cloud by using Structure from Motion (SfM). Afterwards, this point cloud is aligned to the 3D object with a modified version of the 4 Point Congruent Set (4PCS) algorithm [185]. The 4PCS extension accounts for models with different scales and unknown amount of overlapping regions. The transformation that aligns the sparse point cloud (that resulted from the 2D images) to the dense 3D object is applied to the extrinsic parameters of the cameras. In the final stage, a global refinement method is applied based on Mutual Information (MI), which improves the accuracy of the final 2D/3D alignment. The main advantage of this framework is that there is no need for user intervention, no prior knowledge is necessary and there are no requirements regarding the geometry and the visual features involved. However, the initial step of reconstructing the sparse point cloud can be time-consuming in some cases.

The method of Pintus and Gobbetti [130] is another fully automatic framework for image-to-geometry alignment that uses a GPU-based global affine 3D point set stochastic registration approach. The method consists of three steps. In the first step, an SfM algorithm is applied to the collection of images to construct a sparse 3D model; this is achieved by matching features across the images, merging all camera poses in a common reference frame and estimating the intrinsic parameters of the cameras. The second step aligns the sparse 3D model generated from the SfM by utilizing a stochastic global registration method for point clouds [186]. An extra local refinement step is then performed in order to compute correspondences in the newly aligned point clouds. The method utilizes the approximate GPU-accelerated method of [187]. In the final step, a Specialized Sparse Bundle Adjustment (SBA) calculates the final registration in a non-rigid deformable manner, constraining the features detected in the images to lie on the 3D model. Compared to Corsini et al. [120], this strategy does not require heavy pre-processing for altering the sparse 3D point cloud into a dense model. This is due to the global registration method used that recovers the globally optimal scale, rotation and translation alignment parameters.

A similar approach was proposed by Zhao et al. [167] for aligning a video sequence with a 3D point cloud obtained from a 3D sensor (i.e. LiDAR). First, the camera pose is estimated and secondly, 3D structure is reconstructed from the video sequence via a SfM/stereo algorithm. Then, the ICP algorithm is applied to register the input point cloud with the reconstructed one. This method has some limitations, like the computationally expensive process of generating 3D point clouds from video. Also, due to the use of ICP, the initial poses of the point clouds should be close in order to find a good solution while the alignment may fail in scenes with discontinuities.

A depth-aware 2D/3D registration technique is proposed in [136] that utilizes a Point-to-Plane (PPC) model introduced in [188]. The method measures the local misalignment between the projection of a 3D volume and a 2D image (X-ray), followed by the computation of the 3D rigid transformation using the PPC model required to align them. In the initialization step, the method computes a set of 3D feature points from the 3D volume, which are then used to identify the salient structures to be further registered. Then, in each iteration, first a set of contour generator points are selected, as a subset of the initially computed points, and projected onto the image plane, with their depths and 3D gradient preserved (depth aware gradient projections (DGP)). Afterwards, the local misalignment is measured between the DGP and the X-ray image. The goal is to minimize the visual misalignment between the DGP and the actual contour points from the 2D X-ray image. This iterative scheme is accurate in single-view scenarios and robust against outliers but only when they are a minority.

In [141] and [138] the authors extended the [136] method to multi-view registration. In [141], the method performs single-view registration for all views, selects the most promising results and refines the out-of-plane parameters using the other view(s). Alternatively, in [138], a variant of [141] has been proposed, which first computes the transformation sequentially for each view and then each iteration alternates between the different views. The result is then selected as the iteration which leads to the best alignment.

6.2. Learning-based registration

Recently, machine learning approaches have been increasingly applied to multimodal registration, instead of the conventional optimization-based techniques, in order to overcome the challenges of prolonged running time and the risk of falling into local minima.

Two common strategies exist, the first one is to estimate a similarity metric via deep learning techniques and the other is to predict the transformation parameters directly with deep learning. The former approach utilizes deep learning methods so as to learn a similarity metric from training data and then feed it in a traditional registration framework. The latter uses deep learning networks to predict without iteration the transformation parameters, so a deep neural network acts like a regressor to find the transformation that aligns the datasets. This can be further classified, according to the training process, into reinforcement learning, supervised and unsupervised.

Table 3 provides an overview of multimodal 3D registration methods according to the above categorization.

6.2.1. Learning of similarity metric

As a first attempt to use deep learning (DL) in registration, researchers used neural networks to learn similarity metrics between the data to be registered from a large set of paired labeled ground-truths. The estimated similarity measure between modalities is then used within a typical iterative optimization registration method. The strategy followed is to seek a similarity metric that best suits the multimodal datasets, thus taking into consideration the differences in intensity per case study. The similarity metric is then provided to an iterative optimization registration framework in order to determine the transformation parameters [212,213] in a conventional way, without the use of neural networks. Combining deep learning with conventional registration, these methods achieved better performance and accuracy than conventional, iterative, intensity-based registration techniques, especially in the multimodal case, where it is difficult to find a general similarity metric that can be successfully deployed in different modalities.

Lee et al. [197] presented a supervised technique to learn a similarity function based on features extracted from the neighborhoods around the voxels of interest. The problem of learning a similarity metric was formulated as binary classification, where the goal is to discriminate between aligned and misaligned patches. Support vector machine (SVM) regression was employed to learn the similarity metric and then used within a standard rigid registration algorithm. Experiments have been performed on CT-MRI and PET-MRI image volumes showing accuracy and robustness.

Chou et al. [200] presented a 2D/3D deformable registration method that rapidly detects an objects 3D rigid motion or deformation from a 2D projection image or a small set of them. The method computes the residual between the DRR and X-ray images as a feature and trains linear regressors to estimate the transformation parameters to reduce the residual. The method consists of two stages: registration pre-processing by shape space and registration via regression. The method is based on producing limited-dimension parameterization of geometric transformations based on the regions 3D images. A Riemannian metric is learned for each deformation parameter and is used in the kernel regression for

Table 3
Overview of Learning-based Registration Methods, grouped by evaluation measure and dataset used.

Learning-based Registration Methods												
Method	Modality A	Modality B	Nature of Transform.	Domain of Transform.	ML Strategy	Method	Dataset	Initial Application	Evaluation Measure	Value of Eval. Measure	Training time	Execution time (sec)
Haskins et al. [189]	3D MRI	3D TRUS	rigid	global	DL of Similarity Metric	Supervised	private	medical	TRE	3.82mm ± 1.63	n/a	n/a
Zheng et al. [190]	3DCT	2D X-rays	rigid	global	PTR-Reinforcement learning		private	medical	TRE FR	5.65mm 11.20%	n/a	n/a
Ma Kai et al. [191]	3D CT	2.5D image	rigid	global	PTR-Reinforcement learning		private	medical	TRE		4days	0.06sec-1.60sec
Miao et al. [192]	3D volume	2D X-rays	rigid	global	PTR-Reinforcement learning		private	medical	TRE	1.76mm	17hours	0.6sec- 2.5sec
Hu et al [193].	3D MRI	3D TRUS	non rigid	global	PTR-Supervised	GAN	private	medical	TRE Dice	6.3 mm 0.82	n/a	0.25sec
Yan et al. [194]	3D MRI	3D TRUS	rigid	global	PTR-Supervised	GAN	private	medical	TRE	3.48mm	8hours	n/a
Salehi et al. [195]	3D MRI	2D slice of MRI	non rigid	global	PTR-Supervised	CNN	private	medical	TRE	12.32mm	n/a	0.30sec
Sedghi et al. [196]	3D MRI	3D US	rigid	global	DL of Similarity Metric		IXI [106]	medical	TRE	1.43mm ± 0.64	n/a	n/a
Lee et al. [197]	3D CT	3D MRI	rigid	local	DL of Similarity Metric	Supervised	RIRE [104]	medical	TRE	1.40mm	n/a	n/a
Lee et al. [197]	3D PET	3D MRI	rigid	local	DL of Similarity Metric	Supervised	RIRE [104]	medical	TRE	2.52mm	n/a	n/a
Hu et al. [198]	3D MRI	3D TRUS	non rigid	global	PTR-Supervised	CNN	SmartTarget [112]	medical	TRE Dice	4.2 mm 0.88	n/a	0.25sec
Hu et al [199].	3D MRI	3D TRUS	non rigid	global	PTR-Supervised	CNN	SmartTarget [112]	medical	TRE Dice	4.8 mm 0.82	n/a	0.25se
Chou et al [200].	3D CBCT	2D image	rigid	global	DL of Similarity Metric	Supervised	private	medical	mTRE	0.34mm Ø 0.24	linear	2.61sec
Wright et al. [201]	3D MRI	3D US	rigid	global	DL of Similarity Metric		private	medical	mTRE	1.8 mm,7.9°	n/a	n/a
Cao et al. [202]	3D MRI	3D CT	non rigid	global	PTR-Reinforcement learning	CNN	private	medical	mTRE Dice	1.23mm ± 0.43 0.905	40hours	15sec
Pei et al. [203]	3D CBCT	2D X-rays	non rigid	global	PTR-Supervised	CNN	private	medical	mTRE	0.41mm ± 0.12	n/a	
*POINT*² [118]	3D CT/CBCT	2D X-rays	rigid	global	PTR-Supervised		private	medical	mTRE FR	5.67mm 2.7%	n/a	2.50sec
Fan et al. [204]	3D MRI	3D CT	rigid	global	PTR-UnSupervised	GAN	private	medical	mTRE Dice	1.57mm ± 0.44 0.86		n/a
DSAC [205]	3D scene	2D image	rigid	global	PTR-Reinforcement learning	CNN	7-Scenes [86]	general	mTRE SR	4.1cm, 1.1° 58.5%	n/a	0.1sec
PoseNet [206]	3D scene	2D image	rigid	global	PTR-Supervised		7-Scenes [86]	general	mTRE	2.31m, 2.69°	1hour	0.005sec
Melekhov et al. [207]	3D scene	2D image	rigid	global	PTR-Supervised	CNN	7-Scenes [86]	general	mTRE	0.24mm, 10.24	n/a	n/a
Kendall et al. [91]	3D scene	2D image	rigid	global	PTR-Supervised		7-Scenes [86], Cambridge Landmarks [90]	general	mTRE	1.49m	4hours-1day	0.2sec
Sun et al. [208]	3D MRI	3D US	non rigid	global	PTR-UnSupervised	CNN	RESECT [114]	medical	mTRE	3.91mm	2.66sec	1.21sec
Shotton et al. [86]	3D scene	2.5D image	rigid	global	PTR-Supervised		7-Scenes [86]	general	SR	92.6%	10min	0.5sec
Miao et al. [209]	3D model	2D X-rays	rigid	global	PTR-Supervised	CNN	VIPS [109]	medical	mTREproj	0.282mm	n/a	0.08sec
Miao et al. [47]	3D CT	2D X-rays	rigid	global	PTR-Supervised	CNN	VIPS [109]		mTREproj	0.106 mm	non trivial	0.1sec
Yu et al. [210]	3D CT	3D PET	non rigid	global	PTR-UnSupervised	CNN	private	medical	NCC MI	0.567 ± 0.038 2.340 ± 0.349	n/a	2.60sec
DenseRegNet [211]	3D CT	3D PET	non rigid	global	PTR-UnSupervised	DenseNet	private	medical	NCC	0.633 ± 0.068	n/a	n/a

registering. The method operates via iterative, multi-scale regression, where the regression matrices are learned in a way specific to the 3D image(s) for the specific patient. The method only applies to affine deformations and low-rank approximations of non-linear deformations.

Sedghi et al. [196] utilized special data augmentation techniques called dithering and symmetrizing to train a CNN to learn a similarity metric from roughly aligned data. The framework was used for registering unimodal 3D MRI images but also experiments were performed for aligning MRI with US volumes.

Haskins et al. [189] proposed to use CNN to learn a similarity metric for multimodal rigid registration of MRI and transrectal (TRUS) volumes. The determination of the similarity is formulated as a deep CNN-based problem, so the designed CNN with a skip connection outputs an estimate of the target registration error (TRE), which is used to assess the quality of the registration. Then, the alignment is performed with a traditional optimization framework, that uses an evolutionary algorithm to explore the solution space. A multi-pass approach is used in order to address the issue that the learnt metric could be non-convex and non-smooth.

Different from the above strategies, Wright et al. [201] proposed a Long Short-Term Memory (LSTM) spatial co-transformer network to iteratively align MRI and US volumes group-wise to a common space. The recurrent spatial co-transformer consists of three components, initially an image wraper, then the parameter prediction network and finally the parameter composer, which updates the transformation estimates. The method is robust and successful, even on initially randomly aligned objects.

6.2.2. Predictive transformation registration (PTR)

This registration framework uses deep neural networks as a regressor so as to directly predict the transformation parameters according to a loss function. The methods can be either iterative, such as Reinforcement Learning techniques that train the agent iteratively with award or penalty, or one-off, such as Supervised and Unsupervised neural network frameworks.

Reinforcement Learning-based registration

Reinforcement learning methods utilize a trained agent to perform the registration in a manner similar to an expert. This type of machine learning technique enables the agent to learn from its actions and experiences and is focused on predicting the best actions to be followed in an environment for each state. A typical framing of reinforcement learning includes an agent with some internal states, transition probabilities, and a reward/penalty rate [214]. The agent learns iteratively to interact with the environment so as to produce the final transformation, which maximizes the similarity of the two datasets. At each iteration, the agent chooses the best action, which is the one with the highest probability to get reward from its application in the environment. In terms of registration, the deep reinforcement learning agent can be applied to rigid/non-rigid transformations, where the states are finite and the agent can converge to an optimal solution where the similarity measure is maximized. In contrast to the deep learning of similarity metric techniques, where deep learning is used to identify the measure to be provided in the conventional registration method, this approach uses a given similarity metric (i.e. MI or CC) to directly predict the transformation parameters.

Liao et al. [30] were the first to use reinforcement learning-based registration to perform alignment of 3D CT volumes. Ma et al. [191], extended their work via a Q-learning framework that automatically learns to extract optimal feature representation in order to reduce the appearance discrepancy between different modalities. The data modalities that are used are the 2.5D depth images and 3D CT/MRI volume data. Initially, for speed up reasons, the method reformulates the 3D volume to a 2D image through a projection process and thus the registration problem is simplified to 2D image registration. The method is derived from Q-learning [215] that automatically extracts compact features, but uses the dueling network architecture of [216] with some modifications so as to minimize the effect of intensity distribution discrepancy across different modalities. This approach outperforms registration methods based on ICP, landmarks, deep Q-networks and dueling network, but a huge amount of state-action histories have to be saved during training.

DSAC [205] algorithm is a combination of the RANSCAC algorithm [67] with the reinforcement learning approach. DSAC learns both the scoring function and the transformation predictions within the RANSAC framework. The method replaces the deterministic RANSAC hypothesis with a smooth, differential objective function. The system is broadly applicable, ranging from small objects to entire scenes. However, this method is designed to mimic RANSAC rather than outperform it.

Instead of training a single agent, [192] proposed a multi-agent system with the auto attention mechanism to register a 3D volume and 2D X-ray images. The 2D/3D registration is formulated as a Markov Decision Process (MDP) [30,217] and multiple agents are used to solve it. Each individual agent is trained with dilated fully convolutional network (FCN) to observe a local region of the image. Finally, the registration is driven based on the proposals from multiple agents. While the method achieves a high robustness and outperforms approaches that use the state-of-the-art similarity metric of [218], registration accuracy remains challenging.

Zheng et al. trained a CNN model under a pairwise domain adaptation (PDA) technique [190] to improve the performance generalization of the CNN model, to limit the training data needed and to cope with the discrepancy between synthetic training data and real testing data. The adaptation module can be trained using a few pairs of real and synthetic data and learn effective representations for multimodal registration. The method showed flexibility and can be adopted in a variety of applications (though clinical oriented) especially when only little training data is available.

Cao et al. [202] developed a deep learning method for multimodal 3D image registration by transforming the problem into unimodal registration tasks. Instead of using ground truth samples, the method uses unimodal image similarity to supervise the multimodal deformable registration of CT and MRI volumes. Specifically, prior to network training, the multimodal registration is simplified to unimodal by using a pre-aligned CT and MRI dataset, in which each pair of CT and MRI is registered as paired data. Thus, an MRI has a pre-aligned CT and a CT has a pre-aligned MRI. Moreover, the method utilizes dual supervision, where the similarity guidance is delivered from not only the MRI modality, but also the CT modality, so they can both train the network effectively. Although the framework outperforms traditional registration methods in particular applications, it is limited to bi-modal images.

Supervised transformation prediction

Both strategies mentioned in the previous subsections (learning the similarity metric and reinforcement learning) are iterative making them computationally expensive. In contrast, supervised registration methods train deep neural networks (DNNs) to predict the transformation parameters in one-shot. In supervised learning, ground-truth data with known transformation parameters is required for the training process. The larger the amount of such data and the more representative it is, the better the accuracy and precision of the registration result.

Shotton et al. [86] made a first attempt to use machine learning techniques in 2D/3D registration without known correspondences. They introduced the concept of scene coordinates for camera localization and a random forest regressor to predict initial 2D/3D correspondences from image appearance. The method uses depth

images to create scene coordinate labels which map each pixel from the camera coordinates to the global scene coordinates. This is then used to train a regression forest in order to regress these labels and finally localize the camera. The limitation on using only RGB-D images makes it unsuitable for outdoor scenes.

PoseNet of Kendall et al. [206] trains a CNN to directly regress the 6D pose of a scene from an RGB image. The scene is a scene obtained by Structure-from-Motion (SfM). To train their model, they automatically generated training labels from a video registered to the scene using SfM and combined with transfer learning from recognition to registration for increased efficiency and accuracy. Although PoseNet overcomes many limitations of the traditional approaches, its performance still lacks behind traditional feature-based approaches where local features perform well.

Later the authors extended PoseNet [206] by learning the weight between the camera translation and rotation loss and incorporating the reprojection loss [91]. Thus, PoseNet became scene-geometry aware by minimizing the reprojection error of 3D points in multiple images.

Another improvement of PoseNet has been proposed by Melekhov et al. [207] with the training of an hourglass network of ResNet34 architecture. Their method used skip connections between the encoder and decoder, to directly regress the camera pose.

Pei et al. [203] presented a CNN regression based method for the non rigid registration between 2D X-rays and 3D volumes, by integrating a mixed residual CNN and an iterative refinement scheme. The regression is performed directly on image slices, without feature extraction. Instead, of the one-shot registration estimation, an iterative feedback scheme is used, where the deformation parameters are iteratively fine tuned. The proposed method achieves reliable and efficient online non rigid registration.

A CNN regression approach, named Pose Estimation via Hierarchical Learning (PEHL), was proposed by Miao et al. [47,209] to directly predict the registration transformation parameters, reaching a large capture range and high accuracy in real time. Different from optimization-based methods, which iteratively optimize the transformation parameters, Miao et al. were the first to use deep learning to predict the rigid transformation matrix that aligns a 3D model to 2D X-rays. Initially, an automatic feature extraction step calculates a Digitally Reconstructed Radiograph (DRR) from the 3D CT image. The CNN regressors are then trained to predict the transformation of 2D/3D X-ray attenuation maps and 2D X-ray images. The ground truth data used were synthesized by transforming already aligned data. Hierarchical regression was proposed in which the six transformation parameters (2 translational, 1 scaling and 3 rotation angles) are partitioned into three groups. In this way, the complex regression task is divided into multiple simpler sub-tasks that can be learned independently. This method has significantly higher regression success rates than the traditional optimization-based methods, like MI, CC and gradient correlation.

Salehi et al. [195] proposed a deep residual regression network and a bi-invariant geodesic distance based loss function to perform 2D/3D rigid registration. A CNN is used to predict both rotation and translation using extracted image features. The regression method learns the relation between slice pose and 3D image according to the appearance of the 2D slice. The method uses both mean squared error (MSE) and the geodesic distance as loss function. The addition of geodesic distance improved the performance of the registration method.

Yan et al. [194] proposed an adversarial image registration of MRI and TRUS, inspired by the GAN framework. The method trains two deep networks simultaneously, one for transformation parameter estimation and the other for the discriminator component, which evaluates the quality of the alignment. The paired training data is manually registered by experts and are used as ground-truth. The trained discriminator provides an adversarial loss for regulation and a discriminator score for alignment evaluation, thus the discriminator serves as a certainty evaluator during testing.

Hu et al. [198,199] labeled corresponding structures for training the network for registering MRI and TRUS volumes. The framework requires the anatomical labels and full image voxel intensities as training data so that the end-to-end registration network only requires a pair of MRI and TRUS images without any labels. Later, in [193] they directly regressed the multimodal deformable registration via a weakly supervised anatomical label driven GAN. An adversarial approach is used to constrain CNN training for 3D image registration. During training the registration network simultaneously maximizes the similarity between anatomical labels, and minimizes an adversarial generator loss that measures divergence between the predicted and simulated deformation. However, the registration performance of framework [193] was inferior to [198].

Recently, Liao et al. [118] proposed to address multi-view 2D/3D rigid registration via a Point-of-Interest (POI) Network for Tracking and Triangulation (POINT2). POINT2 directly aligns the 3D CT data with the 2D X-ray by using DNNs to establish a point to point correspondence between multiple views of them, and then performs a shape alignment between the matched points to estimate the 3D CT pose. For 3D correspondence, a triangulation layer projects the tracked POIs in the X-ray images of multiple views back into 3D. While this method achieves an improved performance, it requires a large training set and is only applicable to multi-view registration.

Unsupervised transformation prediction

The lack of large datasets with known transformations to be used as a training data, motivated the development of unsupervised registration methods [219]. In unsupervised registration, DNNs are trained without ground-truth data to construct regression models in order to predict the transformation parameters. The methods use data augmentation techniques to overcome the absence of large ground-truths. Moreover, conventional similarity metrics are used as the loss function of the network. However, defining the proper loss function for a network without ground-truth transformations is not trivial, especially in the case of multimodal registration where defining a similarity metric suitable for different modalities is challenging. Thus, methods using unsupervised learning are still limited.

Sun and Zang [208] proposed an unsupervised method for 3D MRI/US registration with a 3D CNN. The framework is composed of three components, a feature extractor, a deformation field generator and a spatial sampler. Initially, for feature extraction, two fully convolutional neural networks are used to extract higher level representative features from MRI and US images respectively. Then, the features are fed into the deformation field generator, where a deformation field is generated and finally, a spatial sampler is used to apply the deformation field to a regular spatial grid. The network is trained using a similarity metric that incorporates both image intensity and gradient, thus it allows accurate and fast registration.

Yu et al. [210] proposed an unsupervised deep learning method for automatic image registration between 3D PET and CT images. The framework consists of two modules, a low-resolution displacement vector field (LR-DVF) estimator and a 3D spatial transformer and resampler. The LR-DVF estimator uses a 3D deep convolutional network (ConvNet) to directly estimate the voxel-wise displacement (3D vector field) between PET and CT images, and the spatial transformer and resampler warps the PET images to match the anatomical structures in the CT images by using the estimated 3D vector field. The method improves the deep learning network DIRNet of de Vos et al. [220], but the use of Normalized Cross Correla-

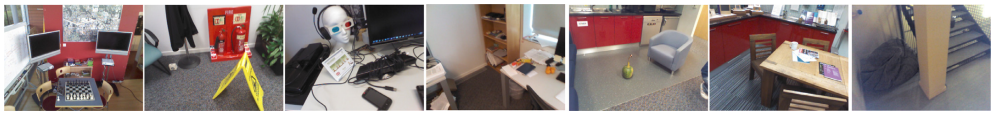


Fig. 8. 7-Scenes dataset sample images from left to right: Chess, Fire, Heads, Office, Pumpkin, Red Kitchen and Stairs.

tion (NCC) as a similarity metric results in over-deforming the PET images.

Kang et al. [211] improved the work of [210] in terms of network structure, loss function and evaluation measures. The method utilizes a 'DenseNet'-based architecture as the displacement vector field (DVF) regressor, for predicting 3D displacement fields. Then, a spatial transformer for warping 3D images is used to obtain the registration result. Moreover, a two-level similarity measure is proposed to optimize the training process, Normalized Cross Correlation (NCC) is used to measure the similarity of voxels at the global level and Maximum Mean Discrepancy (MMD) measures the similarity of data distributions at the higher dimensional level. As for evaluation measures, two anatomical measures are used along with NCC to evaluate the registration results.

Fan et al. [204] proposed an adversarial similarity network to automatically predict the deformation in one-pass, without using any arbitrary similarity metric. The network, which is inspired by generative adversarial networks (GAN), is trained in an adversarial and unsupervised way and does not need ground-truth. A registration network and a discrimination network are connected with a deformable transformation layer. The registration network takes two input 3D images and outputs similarly sized predicted deformations. The registration network is trained with the feedback from the discrimination network, which is designed to judge whether a pair of images are sufficiently aligned. The discrimination network is trained from the registration network's output. The framework is applicable to both unimodal and multimodal registration. Specifically, for multimodal registration, positive image alignments are pre-defined by using paired CT and MRI images. The method effectively registers multimodal images and the use of adversarial loss increases performance.

7. Experimental evaluation of 2D/3D registration methods

Although many authors provide evaluation of their methods, only few of these experiments and results allow a direct comparison against the state-of-the-art. The main reasons are that most of the algorithms are only evaluated on private datasets, they are assessed using different measures and their source code is not publicly available.

In order to provide a useful comparison, we have tested methods with publicly available source code on the same dataset. The only methods with publicly available source code are [67,86,91,126,142,195,199,205,206] [199]. and [195] are medically oriented methods that register 3D MRI volumes with 3D TRUS and 2D slices of MRI respectively. These methods could not be compared with the rest of the methods to align 3D models or scenes with 2D images or points, so experiments have been performed only on the seven remaining methods. Even these methods were not exactly aligning the same modalities. More specifically, [91,205,206] register 3D scenes and 2D images, [86] registers 3D scenes and 2.5D images, while [67,126,142] register 3D point clouds and 2D points. Thus, the main challenge was to identify a publicly available dataset that could be used for our tests. The dataset that fitted best was the 7-Scenes dataset [85,86], sample frames of which are shown in Fig. 8.

Table 4
Information about the scenes and the data of the 7-Scenes dataset.

Scene	Spatial	# Frames	
	Extent (m)	Train	Test
Chess	3 × 2 × 1m	4000	2000
Fire	2.5 × 1 × 1m	2000	2000
Heads	2 × 0.5 × 1m	1000	1000
Office	2.5 × 2 × 1.5m	6000	4000
Pumpkin	2.5 × 2 × 1m	4000	2000
Red Kitchen	4 × 3 × 1.5m	7000	5000
Stairs	2.5 × 2 × 1.5m	2000	1000

Shotton et al. in [86] also propose a method for aligning a 3D scene with a 2.5D image, with experiments on the 7-Scenes dataset that they also provide. Apart from this, DSAC, [205], PoseNet [206] and [91] also register 3D scenes but with 2D images (not 2.5D), thus the 7-Scenes dataset can also be used by ignoring the depth information. The authors of these three methods have also used the 7-Scenes dataset themselves for evaluating their results. However, SoftPOSIT [142], RANSAC [67] and [126] are registration methods between a 3D point cloud and 2D points. In order to test those methods on 7-Scenes, we had to alter the modalities of the dataset from 3D scene and 2D image into 3D point cloud and 2D points. We converted the 3D models from the so called TSDF volume [87] into 3D point clouds with the technique presented in [221] while the 2D points were detected from the PNG images using the Harris Detector [222].

The 7-Scenes dataset consists of RGB-D images (RGB images in PNG format and depth files) of 7 indoor environments and a 3D model (TSDF volume) of each scene. Each scene contains multiple sequences of RGB-D images that represent independent camera paths. Each image frame is annotated with its 6D camera pose, that defines the ground truth for our experiments. The data of each scene are partitioned into testing or training subsets, with RGB-D image numbers varying from 1k to 7k (Table 4). However, the dataset does not include an explicit image set for validation. Testing took place on a random selection of 10% of the images of one sequence per scene.

The results of the 2D/3D registration experiments are summarized in Tables 5 and 6. The results were evaluated by comparing the final registration errors, expressed as translation and rotation error (Table 5) and mean target registration error mTRE (Table 6), see Eq. 2. The registration results of RANSAC [67], SoftPOSIT [142] and [126] should be seen with caution as these methods were developed for slightly different data. In order for future multimodal registration methods to be more fairly compared, the creation of a publicly available dataset with more modalities and specified ground truth is necessary.

As an additional measure, Shotton et al. proposed the Success Rate (SR), defined as the percentage of test frames for which the registration is considered 'correct' [86]. In particular, for the 7-Scenes dataset, a registered pose is considered 'correct' if it has no more than 5cm translational error and 5°angular error. Not all methods reach the bound as defined by Shotton, so we consider it unfair to provide a comparison on this measure. Table 7, gives

Table 5
Summary of the experimental results of the 2D/3D registration methods. Mean registration error of translation and rotation are given in meters and degrees respectively.

Scene	Registration Error of Methods						
	RANSAC [67]	Shotton et al [86].	PoseNet [206]	Kendal et al [91].	DSAC [205]	SoftPOSIT [142]	Liu et al [126].
Chess	0.042m, 1.4°	0.022m, 1.0°	0.32m, 4.06°	0.13m, 4.48°	0.042m, 1.1°	9.43m, 1.10°	0.95m, 0.02°
Fire	0.371m, 2.1°	0.051m, 2.4°	0.47m, 7.33°	0.27m, 11.3°	0.067m, 3.1°	2.46m, 1.57°	0.72m, 1.09°
Heads	0.098m, 3.1°	0.125m, 5.1°	0.29m, 6.00°	0.17m, 13.0°	0.125m, 4.1°	5.85m, 1.72°	0.90m, 4.71°
Office	0.089m, 1.6°	0.046m, 1.4°	0.48m, 3.84°	0.19m, 5.55°	0.098m, 2.7°	4.26m, 1.26°	1.17m, 1.47°
Pumpkin	0.045m, 1.7°	0.065m, 3.7°	0.47m, 4.21°	0.26m, 4.75°	0.040m, 1.5°	9.94m, 1.35°	1.14m, 1.29°
Red Kitchen	0.087m, 2.4°	0.072m, 2.1°	0.59m, 4.32°	0.23m, 5.35°	0.078m, 2.6°	20.7m, 1.29°	0.64m, 1.18°
Stairs	0.65m, 3.2°	0.149m, 2.6°	0.47m, 6.93°	0.35m, 12.4°	0.493m, 3.1°	9.02m, 1.53°	1.00m, 1.48°

Table 6
Summary of experimental results of 2D/3D registration methods, using mTRE (in meters).

Scene	mTRE of Methods						
	RANSAC [67]	Shotton et al [86].	PoseNet [206]	Kendal et al [91].	DSAC [205]	SoftPOSIT [142]	Liu et al [126].
Chess	0.03m	0.032m	0.45m	0.24m	0.04m	6.68m	2.94m
Fire	0.4m	0.045m	0.34m	0.45m	0.07m	4.26m	1.07m
Heads	0.12m	0.210m	0.52m	0.29m	0.14m	4.60m	1.09m
Office	0.07m	0.121m	0.67m	0.17m	0.19m	3.99m	3.56m
Pumpkin	0.03m	0.256m	0.49m	0.36m	0.03m	9.80	3.29m
Red Kitchen	0.09m	0.06m	0.61m	0.25m	0.06m	20.96m	5.55m
Stairs	0.75m	0.161m	0.58m	0.46m	0.04m	10.58m	3.17m

Table 7
Summary of experimental results of 2D/3D registration methods, using the SR measure.

Scene	SR of Methods		
	RANSAC [67]	Shotton et al [86].	DSAC [205]
Chess	96.8%	92.6%	97.4%
Fire	71.8%	82.9%	71.6%
Heads	66.7%	49.4%	67.0%
Office	57.6%	74.9%	59.4%
Pumpkin	59.0%	73.7%	58.3%
Red Kitchen	40.1%	71.8%	42.7%
Stairs	12.8%	27.8%	13.4%

the SR measures as they have been stated in the related papers [86,205].

Although the execution time is very important, the experiments were performed in a non-optimized environment, thus execution time results are not reported.

8. Discussion

3D registration has been an active research field since the 1980s; multimodal 3D registration gained popularity in the past decade, while in the last few years it has been really active.

Some useful conclusions can be extracted from Tables 2 and 3. To begin with, 63% of the presented methods belong to the optimization-based category which leaves the learning-based registration category with 37% of the methods (see Fig. 9). Even though optimization-based techniques are well studied, several problems remain unresolved. First, the iterative nature of such algorithms leads to high computational complexity and thus these algorithms cannot be used in real-time applications like medical imaging. Second, most optimization-based techniques are dependent on the initial pose of the data to be aligned. If the initial position of the data to be registered is not proper, the resulted registration is not accurate. Research is focused on trying to gain better registration results by adjusting traditional optimization algorithms for the multimodal case [149,166] or by proposing new similarity metrics [136] that show better results on the chosen modalities. The trend in the number of methods published each year shows a consistent interest in conventional techniques; thus this area appears to still have prospects. Further investigation in this area should focus on improving the robustness of the methods and decrease computational cost.

Learning-based methods are more recent, with a strong trend in the last 5 years in this category. This trend is supported by the fact that learning-based techniques achieve, in general, better results in terms of registration errors and computational time. We believe that learning-based methods have become particularly attractive in multimodal registration, because it is quite challenging to write code that defines correspondences across different modalities. Another factor that may have hastened the introduction of learning-based methods in multimodal registration, is recent breakthroughs that allowed deep learning networks to consume 3D meshes or 3D point clouds, such as Geometric Deep Learning [223].

In Fig. 10 more statistics of registration methods using deep-learning are illustrated. The supervised methodology is most commonly used. The main reason for this could be that supervised methods perform registration non-iteratively and are thus faster. Supervised registration methods are practically real time, thus it is easier to utilize them in applications such as computer-aided surgery and image-guided therapy. Methods that employ the deep-learning of a similarity measure are also increasing in number since the first DL techniques appeared in 2013. This kind of strategy uses deep learning to identify the similarity measure that is then passed to a traditional optimization-based method. They are thus easier to be understood and implemented. Particularly in multimodal registration, these techniques can be trained to identify structural differences between modalities and result in better registration accuracy. However, they also inherit the computational burden of iterative approaches. Both the aforementioned approaches, are dependent on large datasets of annotated ground truth for their training phase. This is the reason why reinforcement learning and the unsupervised category are gaining popularity in the last 3 years. Unsupervised methods avoid the large amount of annotated data needed for the training process and the associated computational cost for training. Although the unsupervised methodology appears to become a new trend in multimodal registration, it also has its challenges. Unsupervised methods use similarity measure(s) as loss function to guide the learning process.

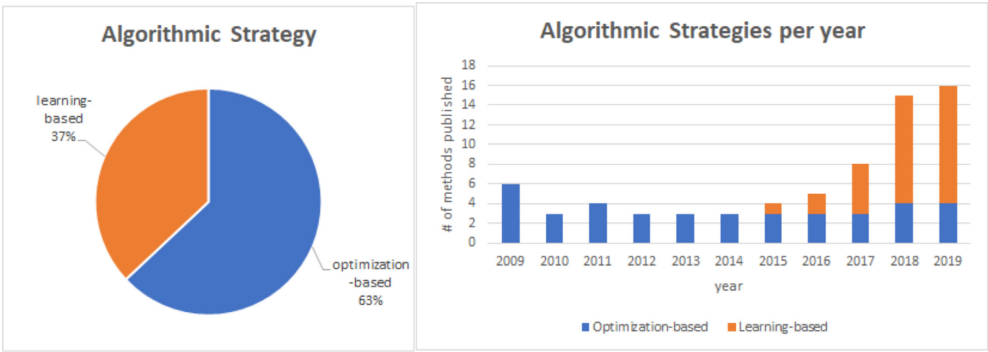


Fig. 9. Overview of the number of publications in multimodal 3D registration based on their algorithmic strategy .

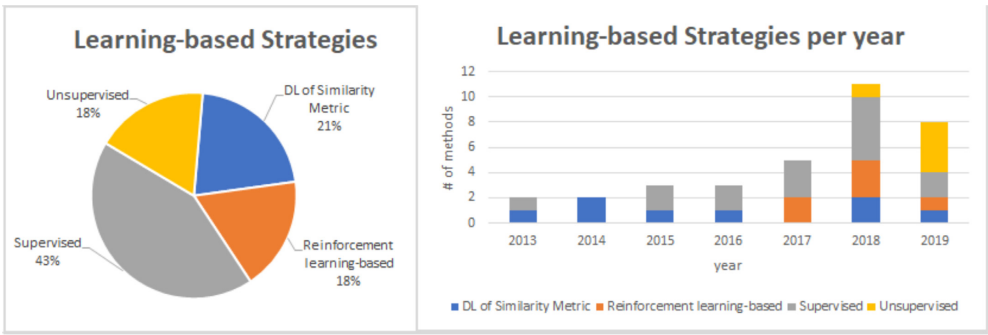


Fig. 10. Overview of the number of proposed learning-based methods for multimodal 3D registration .

However, the multimodal case is more complicated and the traditional similarity measures are not applicable and inefficient; novel similarity measures are expected to be introduced in the future.

Regarding the datasets upon which experiments were conducted by the presented techniques, it should be highlighted that 53% are private while 47% are publicly available (see Fig. 11). The lack of large-scale open datasets is the most frequent challenge of 3D registration. From Fig. 11, it is obvious that there is no single dataset that is most commonly used for testing and benchmarking analysis. The majority of state-of-the-art methodologies use their own small-size proprietary datasets for experiments. The use of different datasets, makes comparison between the different approaches hard. Also, the use of small datasets for evaluation, results in less significant and unreliable findings. Moreover, due to the lack of a unified dataset consisting of multiple modalities, it is not possible to test if the state-of-the-art techniques can be extended to work efficiently with other modalities. Multimodal registration encompasses a variety of modalities, with the same or different dimensions. Most of the techniques focus on aligning two modalities and their evaluation datasets contain only these modalities. From Table 1, it can be seen that there are a few datasets with 3D models and 2D images that are used for testing 2D/3D registration techniques. The rest of the datasets are medically oriented, consisting also of two modalities in most cases. Having algorithms tested on the same benchmark dataset(s) provides direct and reliable comparisons. Furthermore, having a benchmark with multiple modalities would ease the testing of the registration techniques across different modalities. Thus, a public benchmark with gold standard annotations would allow new approaches to be fairly tested against the state-of-the-art. So, it appears that there is a strong need for the creation of better benchmark multimodal datasets.

Various evaluation measures have been used for measuring the accuracy of registration results (Fig. 12) with the TRE, mTRE and SR being the top three in terms of popularity. The variety in evaluation measures challenges fair comparisons even further, especially when combined with the above mentioned variety in evaluation datasets. Since there are significant differences between modalities (e.g. appearance, scale, dimension), it is difficult to define a single measure that could apply to different modality combinations. Future techniques are expected to adopt the aforementioned measures (TRE, mTRE and SR) along with well-defined ground truth registration databases in order to be easily comparable against the state-of-the-art.

The efficiency of registration is also an important attribute for comparing the techniques, in addition to registration accuracy. Unfortunately, most researchers focus on accuracy results and do not report the computational cost and complexity of their approaches in detail. Moreover, computational time can only provide a rough estimate of performance because there is high dependency on the hardware used, which is quite different among researchers, as well as on the server load at the time of the experiments. In addition,

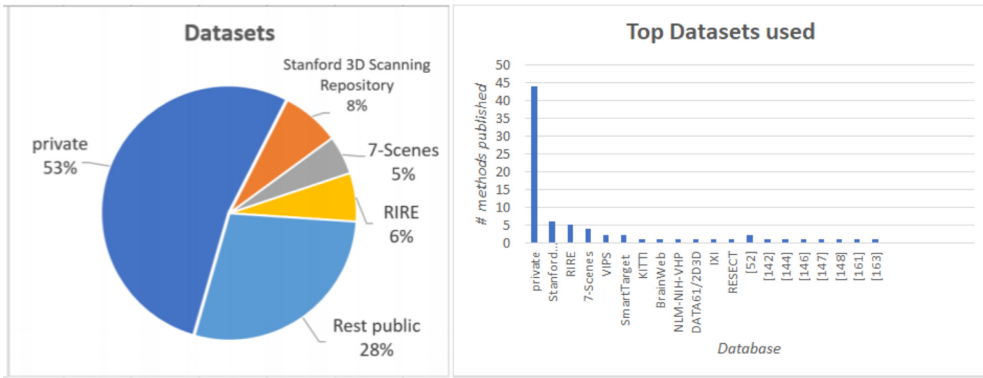


Fig. 11. Overview of the datasets used to implement/test the presented techniques.

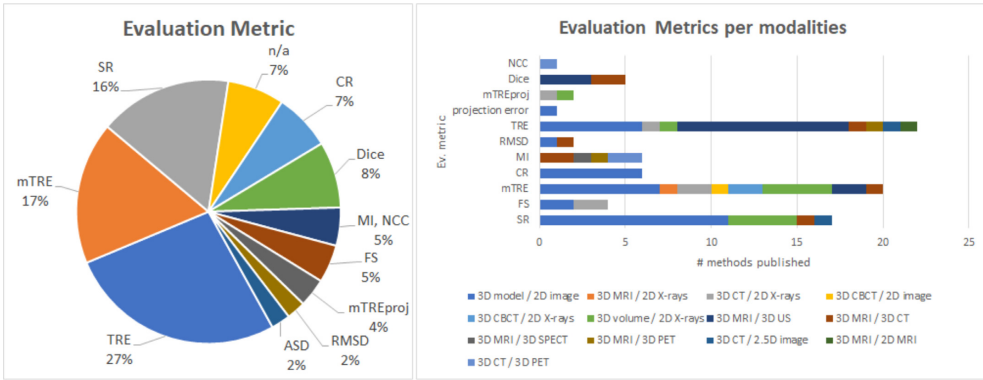


Fig. 12. Overview of evaluation measures used in the presented multimodal 3D registration methods.

the comparison of computational time is not fair because the experiments have been executed on different datasets with different modalities, scale and complexity. This leads once again to the conclusion that the creation of a large scale benchmark database, along with the corresponding ground truth, would be a very positive addition to this thriving field.

In terms of implementation hardware, most of the latest methods utilize GPUs in order to speed up the registration process. GPUs are highly parallel computing engines, which can execute multiple threads in parallel. Although, GPUs offer a good acceleration vehicle, not all algorithmic parts of multimodal registration can be implemented on the GPU. Hybrid CPU-GPU implementations appear to achieve the best performance, so a common implementation strategy of recent years is to use the CPU for execution of optimization algorithms and the GPU to calculate similarity measures in parallel.

The majority of the methods are implemented in C++ or Python and a small portion in Matlab. Matlab is suitable for API prototyping and proof-of-concept, but it is rather slow, which makes it inappropriate for integration with third party software tools. C++ and Python are widely applicable and suitable for real-time applications. Most deep-learning methods chose Python because it provides many open frameworks, especially for DL. TensorFlow, Pytorch and Caffe are the most popular packages because they provide efficient implementations for deep-learning techniques; it is expected that they will continue to be used for registration in future research.

Finally, with respect to the originating applications, the medical one seems by far the biggest group with 50% of the methods, followed by the general category with 30% (see Fig. 13). Naturally, in the medical field, there are many body scanning modalities that need to be registered in order to acquire an integrated view of the body. As shown in the right hand chart of Fig. 13, registration of 3D models to 2D images is the most common case across applications. This is due to the general nature of these modalities, that can be applied in many fields. Moreover, the vast variety of sensors (i.e. digital cameras, 3D laser scanners, Kinect-like RGB-D sensors) produce 3D models (point clouds, meshes). Other than that, there is no single modality that is most commonly used for registration across applications; however, many methods have focused on modalities like MRI, CT and X-rays. These modalities are medically oriented, so most of the methods focus on registration of a specific body organ and do not easily generalize. Taking into consideration the modalities of the publicly available datasets and the number of

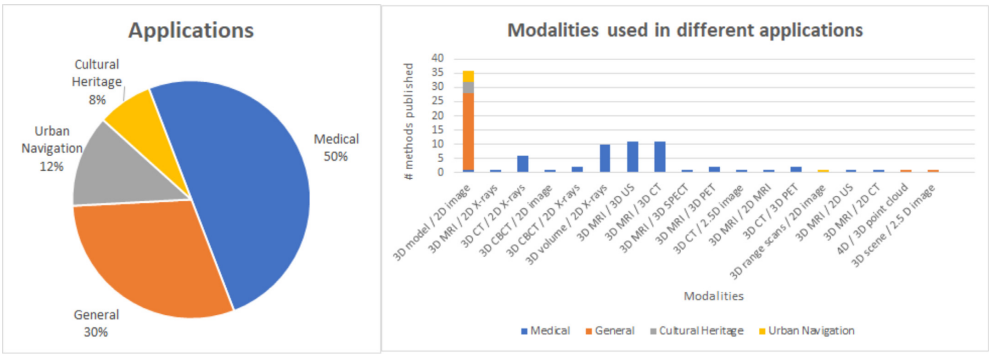


Fig. 13. Pie Charts of applications and modalities registered per application.

subjects that each one contains (Table 1) it can be said that most of such public datasets contain only a small number of subjects in one or two different modalities. The medical field could offer the opportunity of building a dataset with multiple modalities and objects, but there may be challenges related to privacy. The most recent multimodal datasets, IXI [106] and SmartTarget [111], consist of a large number of subjects (600 and 129 respectively). However, even such an amount of data is not sufficient for training and testing of deep-learning registration methods. Also, datasets with Cultural Heritage objects are not large enough, because this kind of object faces many challenges, e.g. too fragile or too large for scanning. The limited availability of large-scale datasets is expected to lead to more methods focusing on transfer learning for registering multimodal data in the near future.

Given the importance of the medical area and available funding, we expect it to remain strong in multimodal registration research. Another significant source of multimodal registration methods has been Cultural Heritage and, given the fact that there are many European projects and open calls in this field [224,225], we expect it to remain strong.

9. Conclusions

Multimodal registration has significantly grown within the last decade. It is a core procedure in multiple applications, like medical imaging, cultural heritage and autonomous navigation. As each modality has its own unique characteristics and each application its own requirements, it is challenging to develop a general registration framework that applies to all modalities and uses.

In this paper, the problem of 3D multimodal registration has been explicitly defined, and the most representative, classical and up-to-date algorithms have been surveyed. The methods were classified according to their nature and strategy followed. The two main categories presented are optimization-based and learning-based, each of which is further sub-categorized. The approaches in each category mostly share the same algorithmic philosophy, principles, advantages and drawbacks. Using such a classification, several aspects of multimodal registration were examined and useful insights regarding future trends were extracted.

Declaration of Competing Interest

The authors declare that they do not have any financial or non-financial conflict of interests

CRediT authorship contribution statement

E. Saiti: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Software, Writing - original draft. **T. Theoharis:** Conceptualization, Funding acquisition, Project administration, Methodology, Resources, Supervision, Validation, Visualization, Writing - review & editing.

References

[1] Fonseca MJ, Ferreira A, Jorge JA. Towards 3D modeling using sketches and retrieval. In: Eurographics Workshop on Sketch-Based Interfaces and Modeling 2004. Citeseer; 2004. p. 127.
[2] Kim P, Chen J, Cho YK. SLAM-Driven robotic mapping and registration of 3D point clouds. Autom Constr 2018;89:38–48.
[3] Weinmann M, Leitloff J, Hoegner L, Jutzi B, Stilla U, Hinz S. Thermal 3D mapping for object detection in dynamic scenes. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2014;2(1):53.
[4] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2013. p. 2100–6.
[5] Mellado N, Dellepiane M, Scopigno R. Relative scale estimation and 3D registration of multi-modal geometry using growing least squares. IEEE Trans Vis Comput Graph 2015;22(9):2160–73.
[6] Chang W, Zwicker M. Global registration of dynamic range scans for articulated model reconstruction. ACM Transactions on Graphics (TOG) 2011;30(3):1–15.
[7] Zollhöfer M, Stotko P, Görlitz A, Theobalt C, Nießner M, Klein R, et al. State of the art on 3D reconstruction with RGB-D cameras. In: Computer graphics forum, 37. Wiley Online Library; 2018. p. 625–52.
[8] Russell BC, Sivic J, Ponce J, Dessales H. Automatic alignment of paintings and photographs depicting a 3D scene. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE; 2011. p. 545–52.
[9] Aubry M, Maturana D, Efros AA, Russell BC, Sivic J. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of cad models. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 3762–9.
[10] Chane CS, Mansouri A, Marzani FS, Boochs F. Integration of 3D and multispectral data for cultural heritage applications: survey and perspectives. Image Vis Comput 2013;31(1):91–102.
[11] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. IEEE Trans Med Imaging 2013;32(7):1153–90.
[12] Birkfellner W, Figl M, Furtado H, Renner A, Hatamikia S, Hummel J. Multi–modality imaging: a software fusion and image-guided therapy perspective. Front Phys 2018;6:66.
[13] Kim H, Evans A, Blat J, Hilton A. Multimodal visual data registration for web-based visualization in media production. IEEE Trans Circuits Syst Video Technol 2016;28(4):863–77.
[14] Bartoli G. Image registration techniques: a comprehensive survey. Visual Information Processing and Protection Group 2007:1–54.
[15] Salvi J, Matabosch C, Fofi D, Forest J. A review of recent range image registration methods with accuracy evaluation. Image Vis Comput 2007;25(5):578–96.
[16] Bellekens B, Spruyt V, Berkvens R, Penne R, Weyn M. A benchmark survey of rigid 3D point cloud registration algorithms, 8; 2015. p. 118–27.

[17] Maiseli B, Gu Y, Gao H. Recent developments and trends in point set registration methods. J Vis Commun Image Represent 2017;46:95–106.
[18] Tam GK, Cheng Z-Q, Lai Y-K, Langbein FC, Liu Y, Marshall D, et al. Registration of 3D point clouds and meshes: asurvey from rigid to nonrigid. IEEE Trans Vis Comput Graph 2012;19(7):1199–217.
[19] Díez Y, Roure F, Lladó X, Salvi J. A qualitative review on 3D coarse registration methods. ACM Computing Surveys (CSUR) 2015;47(3):1–36.
[20] Ferrante E, Paragios N. Slice-to-volume medical image registration: asurvey. Med Image Anal 2017;39:101–23.
[21] Andrade N, Faria FA, Cappabianco FAM. A practical review on medical image registration: from rigid to deep learning based approaches. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE; 2018. p. 463–70.
[22] Viergever M.A., Maintz J.A., Klein S., Murphy K., Staring M., Pluim J.P.. A survey of medical image registration–under review. 2016.
[23] Mani V, Arivazhagan S. Survey of medical image registration. Journal of Biomedical Engineering and Technology 2013a;1(2):8–25.
[24] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[25] Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. arXiv preprint arXiv:190302026 2019a.
[26] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: areview. arXiv preprint arXiv:191212318 2019.
[27] Boveiri HR, Khayami R, Javidan R, MehdiZadeh AR. Medical image registration using deep neural networks: a comprehensive review. arXiv preprint arXiv:200203401 2020.
[28] Kotsas PD, Dodd T. A review of methods for 2D/3D registration. World Acad Sci Eng Technol 2011;59:606–9.
[29] Bosché F. Plane-based registration of construction laser scans with 3D/4D building models. Adv Eng Inf 2012;26(1):90–102.
[30] Liao R, Miao S, de Tournemire P, Grbic S, Kamen A, Mansi T, et al. An artificial agent for robust image registration. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017. p. 4168–75.
[31] Rusinkiewicz S, Levoy M. Efficient variants of the icp algorithm. In: Proceedings Third International Conference on 3-D Digital Imaging and Modeling. IEEE; 2001. p. 145–52.
[32] Yang J, Li H, Campbell D, Jia Y. Go-ICP: a globally optimal solution to 3D ICP point-set registration. IEEE Trans Pattern Anal Mach Intell 2015a;38(11):2241–54.
[33] Besl PJ, McKay ND. Method for registration of 3-D shapes. In: Sensor fusion IV: control paradigms and data structures, 1611. International Society for Optics and Photonics; 1992. p. 586–606.
[34] Huang X, Fan L, Wu Q, Zhang J, Yuan C. Fast registration for cross-source point clouds by using weak regional affinity and pixel-wise refinement. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2019. p. 1552–7.
[35] Yoshimura R, Date H, Kanai S, Honma R, Oda K, Ikeda T. Automatic registration of MLS point clouds and SfM meshes of urban area. Geo-spatial Information Science 2016;19(3):171–81.
[36] Chee E, Wu Z. Airnet: self-supervised affine registration for 3D medical images using neural networks. arXiv preprint arXiv:181002583 2018.
[37] Levoy M., Gerth J., Curless B., Pull K.. The Stanford 3D scanning repository. http://graphics.stanford.edu/data/3Dscanrep/; Accessed on April 2020.
[38] Pujol-Miro A, Ruiz-Hidalgo J, Casas JR. Registration of images to unorganized 3D point clouds using contour cues. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE; 2017. p. 81–5.
[39] Eck M, DeRose T, Duchamp T, Hoppe H, Lounsbery M, Stuetzle W. Multiresolution analysis of arbitrary meshes. In: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques; 1995. p. 173–82.
[40] Oliveira MM, Brauwers M. Real-time refraction through deformable objects. In: Proceedings of the 2007 symposium on Interactive 3D graphics and games; 2007. p. 89–96.
[41] Markelj P, Tomaževič D, Likar B, Pernuš F. A review of 3D/2D registration methods for image-guided interventions. Med Image Anal 2012;16(3):642–61.
[42] Lepetit V, Moreno-Noguer F, Fua P. Epnp: an accurate o(n) solution to the pnp problem. Int J Comput Vis 2009;81(2):155.
[43] Lu XX. A review of solutions for perspective-n-point problem in camera pose estimation. In: Journal of Physics: Conference Series, 1087. IOP Publishing; 2018. p. 052009.
[44] El-Gamal FE-Z A, Elmogy M, Atwan A. Current trends in medical image registration and fusion. Egyptian Informatics Journal 2016;17(1):99–124.
[45] Yu W, Tannast M, Zheng G. Non-rigid free-form 2D–3D registration using a b-spline-based statistical deformation model. Pattern Recognit 2017;63:689–99.
[46] Van de Kraats EB, Penney GP, Tomaževič D, van Walsum T, Niessen WJ. Standardized evaluation of 2D-3D registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2004. p. 574–81.
[47] Miao S, Wang ZJ, Liao R. A CNN regression approach for real-time 2D/3D registration. IEEE Trans Med Imaging 2016a;35(5):1352–63.
[48] Bhatnagar G, Wu QJ, Liu Z. A new contrast based multimodal medical image fusion framework. Neurocomputing 2015;157:143–52.
[49] James AP, Dasarathy BV. Medical image fusion: a survey of the state of the art. Information fusion 2014;19:4–19.
[50] Alam F, Rahman SU. Challenges and solutions in multimodal medical image subregion detection and registration. J Med Imaging Radiat Sci 2019;50(1):24–30.
[51] Vrubel A, Bellon OR, Silva L. A 3D reconstruction pipeline for digital preservation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 2687–94.
[52] Pernus F, et al. 3D-2D Registration of cerebral angiograms: a method and evaluation on clinical images. IEEE Trans Med Imaging 2013;32(8):1550–63.
[53] Bruno F, Bruno S, De Sensi G, Luchi M-L, Mancuso S, Muzzupappa M. From 3D reconstruction to virtual reality: acomplete methodology for digital archaeological exhibition. J Cult Herit 2010;11(1):42–9.
[54] El-Hakim S, Gonzo L, Voltolini F, Girardi S, Rizzi A, Remondino F, et al. Detailed 3D modelling of castles. International journal of architectural computing 2007;5(2):199–220.
[55] Guislain M, Digne J, Chaine R, Monnier G. Fine scale image registration in large-scale urban lidar point sets. Comput Vision Image Understanding 2017;157:90–102.
[56] Wolcott RW, Eustice RM. Visual localization within lidar maps for automated urban driving. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE; 2014. p. 176–83.
[57] Yao L, Wu H, Li Y, Meng B, Qian J, Liu C, et al. Registration of vehicle-borne point clouds and panoramic images based on sensor constellations. Sensors 2017;17(4):837.
[58] Abayowa BO, Yilmaz A, Hardie RC. Automatic registration of optical aerial imagery to a lidar point cloud for generation of city models. ISPRS J Photogramm Remote Sens 2015;106:68–81.
[59] Taneja A, Ballan L, Pollefeys M. Geometric change detection in urban environments using images. IEEE Trans Pattern Anal Mach Intell 2015;37(11):2193–206.
[60] Daras P, Axenopoulos A. A 3D shape retrieval framework supporting multimodal queries. Int J Comput Vis 2010;89(2–3):229–47.
[61] Daras P, Manolopoulou S, Axenopoulos A. Search and retrieval of rich media objects supporting multiple multimodal queries. IEEE Trans Multimedia 2011;14(3):734–46.
[62] Kim H, Pabst S, Sneddon J, Waine T, Clifford J, Hilton A. Multi-modal big-data management for film production. In: 2015 IEEE International Conference on Image Processing (ICIP). IEEE; 2015. p. 4833–7.
[63] Huttenlocher DP, Ullman S. Recognizing solid objects by alignment with an image. Int J Comput Vis 1990;5(2):195–212.
[64] Ponce J., Hebert M., Schmid C., Zisserman A.. Towards category-level object recognition. 2006.
[65] Olson CF. A general method for geometric feature matching and model extraction. Int J Comput Vis 2001;45(1):39–54.
[66] Pascoe G, Maddern W, Stewart AD, Newman P. Farlap: Fast robust localisation using appearance priors. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2015. p. 6366–73.
[67] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 1981;24(6):381–95.
[68] Kneip L, Yi Z, Li H. SDICP: Semi-dense tracking based on iterative closest points.. In: Bmvc; 2015. 100–1
[69] Middel A, Scheler I, Hagen H. Visualization of large and unstructured data sets–applications in geospatial planning, modeling and engineering; 2011.
[70] Mani V, Arivazhagan DS. Survey of medical image registration. Journal of Biomedical Engineering and Technology 2013b;1(2):8–25.
[71] Elsen PV, Pol E-J, Viergever M. Medical image matching: a review with classification. IEEE Eng in Medicine and Biology Magazine 1993;12(1):26–39.
[72] Girija J, Murthy GK, Reddy PC. 4D medical image registration: A survey. In: 2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2017. p. 539–47.
[73] Wells III WM, Viola P, Atsumi H, Nakajima S, Kikinis R. Multi-modal volume registration by maximization of mutual information. Med Image Anal 1996;1(1):35–51.
[74] Viola P, Wells III WM. Alignment by maximization of mutual information. Int J Comput Vis 1997;24(2):137–54.
[75] Zhao Y, Wang Y, Tsai Y. 2D-image to 3D-range registration in urban environments via scene categorization and combination of similarity measurements. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2016. p. 1866–72.
[76] Parmehr EG, Zhang C, Fraser CS. Automatic registration of multi-source data using mutual information. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2012;7:301–8.
[77] Parmehr EG, Fraser CS, Zhang C, Leach J. Automatic registration of optical imagery with 3D lidar data using statistical similarity. ISPRS J Photogramm Remote Sens 2014;88:28–40.
[78] Sottile M, Dellepiane M, Cignoni P, Scopigno R. Mutual correspondences: An hybrid method for image-to-geometry registration.. In: Eurographics Italian chapter conference; 2010. p. 81–8.
[79] The KITTI vision benchmark. http://www.cvlibs.net/datasets/kitti/raw_data.php; Accessed on April 2020.
[80] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2012. p. 3354–61.
[81] PNG format. https://en.wikipedia.org/wiki/Portable_Network_Graphics; Accessed on April 2020.

[82] Data61/2D3D dataset. https://research.csiro.au/data61/automap-datasets-and-code/; Accessed on April 2020.
[83] Namin ST, Najafi M, Salzmann M, Petersson L. A multi-modal graphical model for scene analysis. In: 2015 IEEE Winter Conference on Applications of Computer Vision. IEEE; 2015. p. 1006–13.
[84] LAR format. https://knowledge.autodesk.com/support/autocad-map-3d/learn-explore/caas/CloudHelp/cloudhelp/2019/ENU/MAP3D-Use/files/GUID-7C7DD8A7-B561-45B0-A803-852E0A667F3C-htm.html; Accessed on April 2020.
[85] RGB-D 7-scenes dataset. https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/; Accessed on April 2020.
[86] Shotton J, Glocker B, Zach C, Izadi S, Criminisi A, Fitzgibbon A. Scene coordinate regression forests for camera relocalization in RGB-D images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 2930–7.
[87] Curless B, Levoy M. A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques; 1996. p. 303–12.
[88] Izadi S, Kim D, Hilliges O, Molyneaux D, Newcombe R, Kohli P, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology; 2011. p. 559–68.
[89] Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, et al. KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality. IEEE; 2011. p. 127–36.
[90] Cambridge landmarks. https://www.mi.eng.cam.ac.uk/projects/relocalisation/; Accessed on April 2020.
[91] Kendall A, Cipolla R. Geometric loss functions for camera pose regression with deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 5974–83.
[92] NVM format. http://ccwu.me/vsfm/doc.html; Accessed on April 2020.
[93] Xyz-rgb. https://www.xyzrgb.com/; Accessed on April 2020.
[94] Ply - polygon file format. http://paulbourke.net/dataformats/ply/; Accessed on April 2020.
[95] Gardner A, Tchou C, Hawkins T, Debevec P. Linear light source reflectometry. ACM Transactions on Graphics (TOG) 2003;22(3):749–58.
[96] Krishnamurthy V, Levoy M. Fitting smooth surfaces to dense polygon meshes. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques; 1996. p. 313–24.
[97] Turk G, Levoy M. Zippered polygon meshes from range images. In: Proceedings of the 21st annual conference on Computer graphics and interactive techniques; 1994. p. 311–18.
[98] brainweb dataset. https://brainweb.bic.mni.mcgill.ca/; Accessed on April 2020.
[99] Cocosco CA, Kollokian V, Kwan RK-S, Pike GB, Evans AC. Brainweb: Online interface to a 3D MRI simulated brain database. In: NeuroImage. Citeseer; 1997.
[100] MINC standard. http://www.bic.mni.mcgill.ca/software/MDP/HTML/MINC_prog_guide.html/the-minc-format.html; Accessed on April 2020.
[101] NLM-NIH visible human project. https://www.nlm.nih.gov/research/visible/visible_human.html; Accessed on April 2020.
[102] Ackerman MJ. Visible human project®: from data to knowledge. Yearb Med Inform 2002;11(01):115–17.
[103] RIRE dataset. https://www.insight-journal.org/rire/index.php; Accessed on April 2020.
[104] West J, Fitzpatrick JM, Wang MY, Dawant BM, Maurer Jr CR, Kessler RM, et al. Comparison and evaluation of retrospective intermodality brain image registration techniques. J Comput Assist Tomogr 1997;21(4):554–68.
[105] DICOM standard. https://www.dicomstandard.org/; Accessed on April 2020.
[106] IXI: Information extraction from images.https://brain-development.org/ixi-dataset/; Accessed on April 2020.
[107] Preprocessed ixi dataset. https://github.com/OpenXAIProject/Preprocessed_IXI_Dataset; Accessed on April 2020.
[108] NIFTI format. https://nifti.nimh.nih.gov/; Accessed on April 2020.
[109] Vetter S, Mühlhäuser I, Recum Jv, Grützner P-A, Franke J. Validation of a virtual implant planning system (VIPS) in distal radius fractures. In: Orthopaedic Proceedings, 96. The British Editorial Society of Bone & Joint Surgery; 2014. 50–50
[110] CAD standards. https://en.wikipedia.org/wiki/CAD_standards; Accessed on April 2020.
[111] SmartTarget dataset. https://www.europeanurology.com/article/S0302-2838(18)30592-X/addons; Accessed on April 2020.
[112] Donaldson I, Hamid S, Barratt D, Hu Y, Rodell R, Villarini B, et al. MP33-20 The smarttarget biopsy trial: a prospective paired blinded trial with randomisation to compare visual-estimation and image-fusion targeted prostate biopsies. J Urol 2017;197(4):e425.
[113] RESECT dataset. https://archive.norstore.no/pages/public/searchResult.jsf; Accessed on April 2020.
[114] Xiao Y, Fortin M, Unsgård G, Rivaz H, Reinertsen I. Retrospective evaluation of cerebral tumors (RESECT): a clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Med Phys 2017;44(7):3875–82.
[115] Fitzpatrick JM, West JB. The distribution of target registration error in rigid-body point-based registration. IEEE Trans Med Imaging 2001;20(9):917–927.
[116] Schwab L, Schmitt M, Wanka R. Multimodal medical image registration using particle swarm optimization with influence of the data's initial orientation. In: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE; 2015. p. 1–8.
[117] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
[118] Liao H, Lin W-A, Zhang J, Zhang J, Luo J, Zhou SK. Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 12638–47.
[119] Moreno-Noguer F, Lepetit V, Fua P. Pose priors for simultaneously solving alignment and correspondence. In: European Conference on Computer Vision. Springer; 2008. p. 405–18.
[120] Corsini M, Dellepiane M, Ganovelli F, Gherardi R, Fusiello A, Scopigno R. Fully automatic registration of image sets on approximate geometry. Int J Comput Vis 2013;102(1–3):91–111.
[121] Wachowiak MP, Smolíková R, Zheng Y, Zurada JM, Elmaghraby AS. An approach to multimodal biomedical image registration utilizing particle swarm optimization. IEEE Trans Evol Comput 2004;8(3):289–301.
[122] Kwan R-S, Evans AC, Pike GB. MRI Simulation-based evaluation of image-processing and classification methods. IEEE Trans Med Imaging 1999;18(11):1085–97.
[123] Chen Y-W, Mimori A, Lin C-L. Hybrid particle swarm optimization for 3-D image registration. In: 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE; 2009. p. 1753–6.
[124] Chen Y-W, Lin C-L, Mimori A. Multimodal medical image registration using particle swarm optimization. In: 2008 Eighth International Conference on Intelligent Systems Design and Applications, 3. IEEE; 2008. p. 127–31.
[125] Lin C-L, Mimori A, Chen Y-W. Hybrid particle swarm optimization and its application to multimodal 3D medical image registration. Comput Intell Neurosci 2012;2012.
[126] Liu Y, Dong Y, Song Z, Wang M. 2D-3D Point set registration based on global rotation search. IEEE Trans Image Process 2018;28(5):2599–613.
[127] Strecha C, Von Hansen W, Van Gool L, Fua P, Thoennessen U. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. ieee; 2008. p. 1–8.
[128] Bettio F, Gobbetti E, Merella E, Pintus R. Improving the digitization of shape and color of 3D artworks in a cluttered environment. In: 2013 Digital Heritage International Congress (DigitalHeritage), 1. IEEE; 2013. p. 23–30.
[129] Marton F, Rodriguez MB, Bettio F, Agus M, Villanueva AJ, Gobbetti E. Isocam: interactive visual exploration of massive cultural heritage models on large projection setups. Journal on Computing and Cultural Heritage (JOCCH) 2014;7(2):1–24.
[130] Pintus R, Gobbetti E. A fast and robust framework for semiautomatic and automatic registration of photographs to 3D geometry. Journal on Computing and Cultural Heritage (JOCCH) 2015;7(4):1–23.
[131] Klima O, Kleparnik P, Spanel M, Zemcik P. Intensity-based femoral atlas 2D/3D registration using Levenberg-Marquardt optimisation. In: Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging, 9788. International Society for Optics and Photonics; 2016. p. 97880F.
[132] Xia J, Xu X, Xiong J. Simultaneous pose and correspondence determination using differential evolution. In: 2012 8th International Conference on Natural Computation. IEEE; 2012. p. 703–7.
[133] Rossi C, Abderrahim M, Diaz JC. EvoPose: a model-based pose estimation algorithm with correspondences determination. In: IEEE International Conference Mechatronics and Automation, 2005, 3. IEEE; 2005. p. 1551–6.
[134] Crombez N, Seulin R, Morel O, Fofi D, Demonceaux C. Multimodal 2D image to 3D model registration via a mutual alignment of sparse and dense visual features. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2018. p. 6316–22.
[135] Toth D, Panayiotou M, Brost A, Behar JM, Rinaldi CA, Rhode KS, et al. 3D/2D Registration with superabundant vessel reconstruction for cardiac resynchronization therapy. Med Image Anal 2017;42:160–72.
[136] Wang J, Schaffert R, Borsdorf A, Heigl B, Huang X, Hornegger J, et al. Dynamic 2-D/3-D rigid registration framework using point-to-plane correspondence model. IEEE Trans Med Imaging 2017;36(9):1939–54.
[137] Madan H, Pernuš F, Likar B, Špiclin Ž. A framework for automatic creation of gold-standard rigid 3D–2D registration datasets. Int J Comput Assist Radiol Surg 2017;12(2):263–75.
[138] Schaffert R, Wang J, Fischer P, Maier A, Borsdorf A. Robust multi-view 2-D/3-D registration using point-to-plane correspondence model. IEEE Trans Med Imaging 2019;39(1):161–74.
[139] Tomaževič D, Likar B, Pernuš F. æGold standardg data for evaluation and comparison of 3D/2D registration methods. Computer aided surgery 2004;9(4):137–44.
[140] Schaffert R, Wang J, Fischer P, Borsdorf A, Maier A. Metric-driven learning of correspondence weighting for 2-D/3-D image registration. In: German Conference on Pattern Recognition. Springer; 2018. p. 140–52.
[141] Schaffert R, Wang J, Fischer P, Borsdorf A, Maier A. Multi-view depth-aware rigid 2-D/3-D registration. In: 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). IEEE; 2017. p. 1–4.
[142] David P, Dementhon D, Duraiswami R, Samet H. Softposit: simultaneous pose and correspondence determination. Int J Comput Vis 2004;59(3):259–284.

[143] David P, DeMenthon D. Object recognition in high clutter images using line features. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2. IEEE; 2005. p. 1581–8.
[144] Enqvist O, Kahl F. Robust optimal pose estimation. In: European conference on computer vision. Springer; 2008. p. 141–53.
[145] Snavely N, Seitz SM, Szeliski R. Photo tourism: exploring photo collections in 3D. In: ACM Siggraph 2006 Papers; 2006. p. 835–46.
[146] Klaudiny M., Tejera M., Malleson C., Guillemaut J., Hilton A.. Scene digital cinema datasets. http://epubs.surrey.ac.uk/807665/; 2014.
[147] Kim H.. Impart multi-modal dataset. http://epubs.surrey.ac.uk/807707/; 2015.
[148] Brown M, Windridge D, Guillemaut J-Y. Globally optimal 2D-3D registration from points or lines without correspondences. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 2111–19.
[149] Brown M, Windridge D, Guillemaut J-Y. A family of globally optimal branch-and-bound algorithms for 2D–3D correspondence-free registration. Pattern Recognit 2019;93:36–54.
[150] Campbell D, Petersson L, Kneip L, Li H. Globally-optimal inlier set maximisation for simultaneous camera pose and feature correspondence. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 1–10.
[151] Sánchez-Riera J, Östlund J, Fua P, Moreno-Noguer F. Simultaneous pose, correspondence and non-rigid shape. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE; 2010. p. 1189–96.
[152] Dong H, Sun C, Zhang B, Wang P. Simultaneous pose and correspondence determination combining softassign and orthogonal iteration. IEEE Access 2019;7:137720–30.
[153] Corsini M, Dellepiane M, Ponchio F, Scopigno R. Image-to-geometry registration: a mutual information method exploiting illumination-related geometric properties. In: Computer Graphics Forum, 28. Wiley Online Library; 2009. p. 1755–64.
[154] Palma G, Corsini M, Dellepiane M, Scopigno R. Improving 2D-3D registration by mutual information using gradient maps.. In: Eurographics Italian Chapter Conference; 2010. p. 89–94.
[155] Yang H, Wang F, Li Z, Dong H. Simultaneous pose and correspondence estimation based on genetic algorithm. Int J Distrib Sens Netw 2015b;11(11):828241.
[156] Enqvist O, Josephson K, Kahl F. Optimal correspondences from pairwise constraints. In: 2009 IEEE 12th international conference on computer vision. IEEE; 2009. p. 1295–302.
[157] Kushal A, Ponce J. Modeling 3D objects from stereo views and recognizing them in photographs. In: European Conference on Computer Vision. Springer; 2006. p. 563–74.
[158] Kisaki M, Yamamura Y, Kim H, Tan JK, Ishikawa S, Yamamoto A. High speed image registration of head CT and MR images based on Levenberg-Marquardt algorithms. In: 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS). IEEE; 2014. p. 1481–5.
[159] Talbi H, Batouche M. Hybrid particle swarm with differential evolution for multimodal image registration. In: 2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT'04., 3. IEEE; 2004. p. 1567–72.
[160] Khoo Y, Kapoor A. Non-iterative rigid 2D/3D point-set registration using semidefinite programming. IEEE Trans Image Process 2016;25(7):2956–70.
[161] Ayatollahi F, Shokouhi SB, Ayatollahi A. A new hybrid particle swarm optimization for multimodal brain image registration. J Biomed Sci Eng 2012;5(4).
[162] Johnson K., Becker J.. The whole brain atlas. http://www.med.harvard.edu/AANLIB/home.html; 2008.
[163] Beveridge JR, Riseman EM. Optimal geometric model matching under full 3D perspective. Comput Vision Image Understanding 1995;61(3):351–64.
[164] David P, DeMenthon D, Duraiswami R, Samet H. Simultaneous pose and correspondence determination using line features. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., 2. IEEE; 2003. II–II
[165] Zhou H, Zhang T, Lu W. Vision-based pose estimation from points with unknown correspondences. IEEE Trans Image Process 2014;23(8):3468–77.
[166] Pan J, Min Z, Zhang A, Ma H, Meng MQ-H. Multi-view global 2D-3D registration based on Branch and Bound algorithm. In: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE; 2019. p. 3082–7.
[167] Zhao W, Nister D, Hsu S. Alignment of continuous video onto 3D point clouds. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1305–18.
[168] Christmas WJ, Kittler J, Petrou M. Structural matching in computer vision using probabilistic relaxation. IEEE Trans Pattern Anal Mach Intell 1995;17(8):749–64.
[169] Gold S, Rangarajan A, Lu C-P, Pappu S, Mjolsness E. New algorithms for 2D and 3D point matching: pose estimation and correspondence. Pattern Recognit 1998;31(8):1019–31.
[170] Dementhon DF, Davis LS. Model-based object pose in 25 lines of code. Int J Comput Vis 1995;15(1–2):123–41.
[171] Lu C-P, Hager GD, Mjolsness E. Fast and globally convergent pose estimation from video images. IEEE Trans Pattern Anal Mach Intell 2000;22(6):610–22.
[172] Powell MJ. The NEWUOA software for unconstrained optimization without derivatives. In: Large-scale nonlinear optimization. Springer; 2006. p. 255–97.
[173] Mastin A, Kepner J, Fisher J. Automatic registration of lidar and optical images of urban scenes. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. p. 2639–46.
[174] Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965;7(4):308–13.
[175] Marques M, Stošić M, Costeira J. Subspace matching: Unique solution to point matching with geometric constraints. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE; 2009. p. 1288–94.
[176] Bhat KS, Heikkilä J. Line matching and pose estimation for unconstrained model-to-image alignment. In: 2014 2nd International Conference on 3D Vision, 1. IEEE; 2014. p. 155–62.
[177] Olson CF. Efficient pose clustering using a randomized algorithm. Int J Comput Vis 1997;23(2):131–47.
[178] Goldberg D. Genetic algorithms in search, optimization, and machine learning, addison-wesley, reading, ma, 1989. NN Schraudolph and J 1989;3(1).
[179] Kennedy J. Swarm intelligence. In: Handbook of nature-inspired and innovative computing. Springer; 2006. p. 187–219.
[180] Chen Y-W, Mimori A. Hybrid particle swarm optimization for medical image registration. In: 2009 Fifth International Conference on Natural Computation, 6. IEEE; 2009. p. 26–30.
[181] Bratton D, Kennedy J. Defining a standard for particle swarm optimization. In: 2007 IEEE swarm intelligence symposium. IEEE; 2007. p. 120–7.
[182] Jurie F. Solution of the simultaneous pose and correspondence problem using gaussian error model. Comput Vision Image Understanding 1999;73(3):357–73.
[183] Yang J, Li H, Jia Y. Go-icp: Solving 3D registration efficiently and globally optimally. In: Proceedings of the IEEE International Conference on Computer Vision; 2013. p. 1457–64.
[184] Korez R, Ibragimov B, Likar B, Pernuš F, Vrtovec T. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. IEEE Trans Med Imaging 2015;34(8):1649–62.
[185] Aiger D, Mitra NJ, Cohen-Or D. 4-Points congruent sets for robust pairwise surface registration. In: ACM SIGGRAPH 2008 papers; 2008. p. 1–10.
[186] Papazov C, Burschka D. Stochastic global optimization for robust point set registration. Comput Vision Image Understanding 2011;115(12):1598–609.
[187] Cayton L. A nearest neighbor data structure for graphics hardware.. In: ADMS@ VLDB; 2010. p. 9–14.
[188] Wang J, Borsdorf A, Heigl B, Köhler T, Hornegger J. Gradient-based differential approach for 3-d motion compensation in interventional 2-D/3-D image fusion. In: 2014 2nd International Conference on 3D Vision, 1. IEEE; 2014. p. 293–300.
[189] Haskins G, Kruecker J, Kruger U, Xu S, Pinto PA, Wood BJ, et al. Learning deep similarity metric for 3D MR–TRUS image registration. Int J Comput Assist Radiol Surg 2019b;14(3):417–25.
[190] Zheng J, Miao S, Wang ZJ, Liao R. Pairwise domain adaptation module for CNN-based 2-D/3-D registration. J Med Imaging 2018;5(2):021204.
[191] Ma K, Wang J, Singh V, Tamersoy B, Chang Y-J, Wimmer A, et al. Multimodal image registration with deep context reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017. p. 240–8.
[192] Miao S, Piat S, Fischer P, Tuysuzoglu A, Mewes P, Mansi T, et al. Dilated fcn for multi-agent 2D/3D medical image registration. In: Thirty-Second AAAI Conference on Artificial Intelligence; 2018. p. 4694–701.
[193] Hu Y, Gibson E, Ghavami N, Bonmati E, Moore CM, Emberton M, et al. Adversarial deformation regularization for training image registration neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2018a. p. 774–82.
[194] Yan P, Xu S, Rastinehad AR, Wood BJ. Adversarial image registration with application for MR and TRUS image fusion. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2018. p. 197–204.
[195] Salehi SSM, Khan S, Erdogmus D, Gholipour A. Real-time deep registration with geodesic loss. arXiv preprint arXiv:180305982 2018.
[196] Sedghi A, Luo J, Mehrtash A, Pieper S, Tempany CM, Kapur T, et al. Semi-supervised deep metrics for image registration. arXiv preprint arXiv:180401565 2018.
[197] Lee D, Hofmann M, Steinke F, Altun Y, Cahill ND, Scholkopf B. Learning similarity measure for multi-modal 3D image registration. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 186–93.
[198] Hu Y, Modat M, Gibson E, Ghavami N, Bonmati E, Moore CM, et al. Label–driven weakly-supervised learning for multimodal deformable image registration. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE; 2018b. p. 1070–4.
[199] Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 2018c;49:1–13.
[200] Chou C-R, Frederick B, Mageras G, Chang S, Pizer S. 2D/3D Image registration using regression learning. Comput Vision Image Understanding 2013;117(9):1095–106.
[201] Wright R, Khanal B, Gomez A, Skelton E, Matthew J, Hajnal JV, et al. LSTM Spatial co-transformer networks for registration of 3D fetal US and MR brain images. In: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis. Springer; 2018. p. 149–59.
[202] Cao X, Yang J, Wang L, Xue Z, Wang Q, Shen D. Deep learning based inter–modality image registration supervised by intra-modality similarity. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2018. p. 55–63.
[203] Pei Y, Zhang Y, Qin H, Ma G, Guo Y, Xu T, et al. Non-rigid craniofacial 2D-3D registration using CNN-based regression. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2017. p. 117–25.

[204] Fan J, Cao X, Wang Q, Yap P-T, Shen D. Adversarial learning for mono-or multi-modal registration. Med Image Anal 2019;58:101545.
[205] Brachmann E, Krull A, Nowozin S, Shotton J, Michel F, Gumhold S, et al. DSAC-differentiable RANSAC for camera localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 6684–92.
[206] Kendall A, Grimes M, Cipolla R. Posenet: A convolutional network for real–time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 2938–46.
[207] Melekhov I, Ylioinas J, Kannala J, Rahtu E. Image-based localization using hourglass networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 879–86.
[208] Sun L, Zhang S. Deformable MRI-ultrasound registration using 3D convolutional neural network. In: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation. Springer; 2018. p. 152–8.
[209] Miao S, Wang ZJ, Zheng Y, Liao R. Real-time 2D/3D registration via CNN regression. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE; 2016b. p. 1430–4.
[210] Yu H, Zhou X, Jiang H, Kang H, Wang Z, Hara T, et al. Learning 3D non–rigid deformation based on an unsupervised deep learning for PET/CT image registration. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, 10953. International Society for Optics and Photonics; 2019. p. 109531X.
[211] Kang H, Jiang H, Zhou X, Yu H, Hara T, Fujita H, et al. An optimized registration method based on distribution similarity and DVF smoothness for 3D PET and CT images. IEEE Access 2019.
[212] Simonovsky M, Gutiérrez-Becker B, Mateus D, Navab N, Komodakis N. A deep metric for multimodal registration. In: International conference on medical image computing and computer-assisted intervention. Springer; 2016. p. 10–18.
[213] Cheng X, Zhang L, Zheng Y. Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2018;6(3):248–52.
[214] Sutton RS, Barto AG, et al. Introduction to reinforcement learning, 135. MIT press Cambridge; 1998.
[215] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[216] Wang Z, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N. Dueling network architectures for deep reinforcement learning. arXiv preprint arXiv:151106581 2015.
[217] Bellman R. On the theory of dynamic programming. Proc Natl Acad Sci USA 1952;38(8):716.
[218] De Silva T, Uneri A, Ketcha M, Reaungamornrat S, Kleinszig G, Vogt S, et al. 3D–2D Image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. Physics in Medicine & Biology 2016;61(8):3009.
[219] de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I. A deep learning framework for unsupervised affine and deformable image registration. Med Image Anal 2019;52:128–43.
[220] de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-end unsupervised deformable image registration with a convolutional neural network. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; 2017. p. 204–12.
[221] Zeng A, Song S, Nießner M, Fisher M, Xiao J, Funkhouser T. 3Dmatch: Learning local geometric descriptors from RGB-D reconstructions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1802–11.
[222] Harris CG, Stephens M, et al. A combined corner and edge detector.. In: Alvey vision conference, 15. Citeseer; 1988. p. 10–5244.
[223] Geometric deep learning. http://geometricdeeplearning.com/; Accessed on April 2020.
[224] CHANGE project. https://change-itn.eu/; Accessed on April 2020.
[225] PRESIOUS project. http://www.presious.eu/; Accessed on April 2020.

Chapter 9

Paper D - Multimodal registration across 3D point clouds and CT-volumes

Authors

Evdokia Saiti, and Theoharis Theoharis.

Published in

Computer & Graphics, Volume 106, Pages 259-266, Elsevier, 2022

Computers & Graphics 106 (2022) 259–266

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on 3DOR 2022

Multimodal registration across 3D point clouds and CT-volumes

E. Saiti *, T. Theoharis

Norwegian University of Science and Technology (NTNU), Department of Computer and Information Science, Norway

ARTICLE INFO

Article history:
Received 28 March 2022
Received in revised form 23 May 2022
Accepted 23 June 2022
Available online 25 June 2022

Keywords:
Multimodal
3D registration
Data fusion
3D volume
3D point cloud
Alignment

ABSTRACT

Multimodal registration is a challenging problem in visual computing, commonly faced during medical image-guided interventions, data fusion and 3D object retrieval. The main challenge of multimodal registration is finding accurate correspondence between modalities, since different modalities do not exhibit the same characteristics. This paper explores how the coherence of different modalities can be utilized for the challenging task of 3D multimodal registration. A novel deep learning multimodal registration framework is proposed by introducing a siamese deep learning architecture, especially designed for aligning and fusing modalities of different structural and physical principles. The cross-modal attention blocks lead the network to establish correspondences between features of different modalities. The proposed framework focuses on the alignment of 3D point clouds and the micro-CT 3D volumes of the same object. A multimodal dataset consisting of real micro-CT scans and their synthetically generated 3D models (point clouds) is presented and utilized for evaluating our methodology.

1. Introduction

The exploitation of multimodal data has benefited many visual computing applications by increasing the performance of operations such as 3D object recognition [1], classification [2], 3D shape retrieval [3,4] and data fusion [5,6]. Applications include medical imaging [7], cultural heritage [8–10] and autonomous driving [11,12].

Registration is the process of aligning different sets of spatial data by determining the proper geometrical transformation [13] between them. Multimodal registration is a special case, where the data to be aligned are of different modalities (e.g. capture techniques or sensors) but represent the same object. These data can be 2D images, 2.5D data (image + depth), 3D images acquired by tomographic modalities like CT, MR or PET, 3D point clouds or 3D meshes. Most multimodal registration research has arisen in the medical imaging field, but cultural heritage (CH) and other areas can equally benefit from the visual combination of multiple modalities in order to produce an accurate and useful representation of, e.g., CH assets [9].

Cultural heritage documentation aims at a multimodal record of CH objects that enables a range of operations, such as inspection, virtual reconstruction of fragmented artefacts and fabrication processes [14–18]. An accurate model of an object's surface and inner structure can also contribute to preservation and monitoring, by detecting any structural damages and deformations in structure or cracks, blistering or erosion. The detailed representation of both the interior and external surface can be used as a foundation for future change monitoring of the object. Alterations can be accurately recorded, quantified and tracked through the years [18]. While our specific motivation and data have arisen from the CH field, the applications of the proposed method are not limited to CH.

Geometry acquired from 3D surface scanners is a core aspect of a digital model, but is limited due to the fact that only data from the surface are acquired and the inner structure of the object cannot be documented. The penetrative capabilities of CT scanning allow the digitization of the interior of an object without having to perform physically invasive actions [18]. By combining 3D surface models and CT imaging techniques, it is possible to produce more precise 3D representations of an object, consisting of an accurate geometric model of the surface along with a detailed representation of its internal structure [19–21].

Multimodal registration is a long standing research area with many challenges. Finding an accurate, robust and fast multimodal alignment[1] is still very challenging, since different modalities come from different acquisition systems, having different representations and properties. In particular, the core difficulty of aligning CT volumes and point clouds comes from the significant difference in physical characteristics and representation which

* Corresponding author.
E-mail addresses: evdokia.saiti@ntnu.no (E. Saiti), theotheo@ntnu.no (T. Theoharis).

[1] We shall use the terms *alignment* and *registration* as synonyms.

https://doi.org/10.1016/j.cag.2022.06.012

Table 1
Registration results of the proposed method on the *'3DPCD-CT'* dataset when random rotations and translations are performed on the initial sub pieces. The metrics evaluated are target registration error (TRE), and Recall with threshold 6.00. The initial TRE of the transformations was 15.34.

Method	TRE	$\text{Recall}_a(\%)$	Mean Exec. time (s)
Proposed	**5.15**	**62**	**0.12**
Excluded module			
3D PointCloud FE	12.42	12	0.12
3D Volume FE	11.37	14	0.05
Cross-modal attention	13.32	11	0.03

manifest themselves, for example, in the lack of a general rule for the comparison and evaluation of the final alignment. The most common practice for aligning such modalities is the conversion of one modality into the other, or of both modalities into a third common one, and their alignment using unimodal techniques. Conversion however results in extra computational cost and loss of structural information. This is the gap that we attempt to address in the current paper.

We propose a deep network architecture capable of registering two different modalities, without transforming either of them before feeding them to the network which performs the registration process. The proposed PCD2VOL method aligns 3D surface data with 3D CT volume data. To the best of our knowledge, this is the first time that a deep learning network is trained to register such modalities. The main contributions of this paper can be summarized as follows:

- The problem of multimodal 3D registration of CT volumes and 3D point clouds is formally defined and a framework for such registration is proposed. *Publicly available upon publication.*
- To the best of our knowledge, it is the first deep learning network that combines regular CNNs suited for data with a standard grid structure and geometric deep learning suited for unstructured data.
- The proposed network employs a siamese architecture for a novel attention mechanism for effective multimodality fusion.
- A multimodal dataset for evaluating algorithms for aligning CT volumes and 3D point clouds. *Publicly available upon publication.*

The remainder of this paper is organized as follows: In Section 2 related works are discussed while in Section 3 the problem of 3D multimodal registration of CT volumes and Point clouds is defined. In Section 4 the proposed methodology for 3D multimodal registration is introduced. The proposed evaluation benchmark and experimental results on multimodal alignment are presented in Section 5. The paper is concluded in Section 6.

2. Related work

Multimodal datasets are increasingly being created and exploited. There has also been growing research on the registration of 3D data obtained from different acquisition sensors or data of different structure. Approaches have been proposed for integrating different data modalities so as to produce complete models. However, according to the specific application, the modalities and the approach vary considerably. Medical imaging [22], remote sensing [23] and cultural heritage documentation [6] have emerged as the most fruitful application areas for 3D multimodal registration. A comprehensive review of 3D multimodal registration methodologies across application domains can be found in [24].

3D multimodal registration has been extensively researched in the medical domain, due to the variety of medical modalities that need to be fused. Medically oriented registration methods focus on specific modality pairs, clinical task or body organs. Detailed surveys on medical multimodal registration can be found in [25–27].

Registration methodologies can be broadly classified based on the type of correspondence between the data (parts, structure or context of each dataset). They may be feature-based or intensity-based. In feature-based registration, features (such as interest points, contours or lines) are first extracted from each dataset and are subsequently used to determine the proper correspondence and alignment. Intensity-based methodologies attempt to identify context similarity between the datasets based on the correlation between pixel/voxel intensities [28]. Both techniques have been successfully employed for aligning data from different modalities by identifying salient structures [29] or statistical dependency of the intensities [30–32] across the different modalities. Alternatively, methods exist that try to simplify the multimodal registration problem to unimodal by reconstructing or mapping one modality onto the other [33,34].

Over the last few years, there is a clear predominance in the use of deep learning techniques for registration [35–38]. However, most of these methods involve the same modality, the specific combination of 2D images/3D model, or are somehow restricted in application to the medical field due to the assumptions made. There is virtually no research in 3D multimodal registration outside the medical field where the modalities are differentiated in both structure and physical principles.

Our work is motivated by the idea of using attention mechanisms for multimodal registration. An attention mechanism enables a model to focus on important information for a task; thus it has been applied widely to various computer vision problems, including image classification [39], object detection [40], image generation [41] and image captioning [42]. Recently this technique has also been used for multimodal registration. [43] fused RGB images and point clouds by learning feature interactions between the modalities with a cross-modal attention scheme while [44] developed a self-attention mechanism specifically for aligning 3D medical volumes of MRI and TRUS modalities.

Our problem is generic in that it concerns the alignment of 3D modalities that are complementary since they jointly describe the interior and the surface of a 3D object. The proposed network exploits cross attention for the challenging task of aligning 3D modalities of different geometric data structures. The proposed framework is a combination of CNN, geometric deep learning for feature extraction and a siamese architecture of cross modal attention network, trained to identify correspondences and fuse regular input data formats (like 3D voxels) and irregular 3D geometric data (like 3D point clouds). To the best of our knowledge, this is the first time that registration of such different modalities, without projecting one modality onto the other, is explored.

3. Problem statement

Given a set of 3D points $\mathbf{P} = \{p_i \in \mathbb{R}^3 |\ i = 1, 2, \ldots, N\}$ and a 3D CT Volume $\mathbf{V} = \{v_{lwh} \in \mathbb{Z} \mid l = 1, \ldots, L, w = 1, \ldots, W, h = 1, \ldots, H\}$, the aim is to find the unknown rigid transformation $\mathbf{T}$, so as to align the two input modalities as well as possible.

The registration result is a rigid transformation matrix $\mathbf{T}(\mathbf{R}, \mathbf{t})$, where $\mathbf{T} \in SE(3)$. It consists of two components; a rotation submatrix $\mathbf{R} \in SO(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$. The rigid transformation $\mathbf{T}$ can then be represented by the following homogeneous 4×4 matrix:

$$\mathbf{T} = \left[\begin{array}{c|c} \mathbf{R} & \mathbf{t} \\ \hline 0 & 1 \end{array}\right] \tag{1}$$

Table 2
Performance comparison between multimodal registration methods.

Method	Data modalities		Modalities structure		Data	Runtime	Initial	TRE	Percent
	M1	M2	S1	S2	conversion	(s)	TRE		change
[45]	MRI	US	3D volume	3D volume	No	20	6.76	2.12	68%
[44]	MRI	TRUS	3D volume	3D volume	No	0.003	8.00	3.63	54%
[46]	RGB	Depth Map	2D image	2D image	No	n/a	35.46	6.93	80%
[22]	MRI	CT	3D volume	3D volume	No	320.4	13.49	7.12	47%
[29]	RGB	Point cloud	2D image	3D model	Yes	9000	n/a	30.19	n/a
Proposed	CT	Point cloud	3D volume	3D model	No	0.12	15.34	5.15	62%

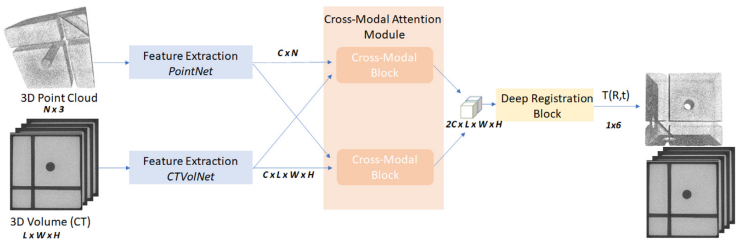


Fig. 1. Overview of the proposed *cross-modal* 3D registration framework. The 3D cross-modal registration network consists of three stages. 1. Each input modality (Point Cloud and 3D CT Volume) is fed into an independent feature extractor network that is suitable for that modality. 2. The captured features are fed to a siamese architecture of cross-modal attention blocks. 3. The registration block fuses the cross-modal features into the final registration parameters.

3D Point Clouds and 3D CT Volumes have different geometrical and physical characteristics. Hence, identifying a distance measure for alignment is challenging. Parameters like the centroid or the bounding box (orientation and location) could approximately measure if two instances of these modalities are aligned. It is inherently difficult to come up with a traditional algorithm which could find correspondences across these modalities. Both modalities represent the same object, therefore common features exist to guide the registration. In our methodology and experiments we take advantage of a ground truth in order to train a neural network and evaluate our results.

4. Method overview

The proposed framework, as illustrated in Fig. 1 consists of three main components. Initially, the 3D point cloud and the 3D CT volume are fed into two modality-specific feature extraction network blocks to identify regional and geometric features of each modality independently. Then, the modality-based features are passed to a siamese architecture of cross-modal attention blocks, in order to capture local features and their global correspondence across the modalities. Finally, the deep registration block processes the fused feature representation to extract the registration parameters. The details of each component are discussed in the following subsections.

4.1. Feature extraction

Each input modality is initially passed to the respective feature extraction network. The feature extraction of the 3D point cloud modality, adopts a variant of PointNet [47]. PointNet has been chosen for this task due to its efficiency in capturing critical geometric features of point clouds. The architecture is shown in Fig. 2.

The 3D CT Volume is passed through CTVolNet, a CNN-based architecture to efficiently represent the CT volume. Based on [48], two sets of convolutional and max-pooling layers are used to capture regional features, shown in Fig. 3.

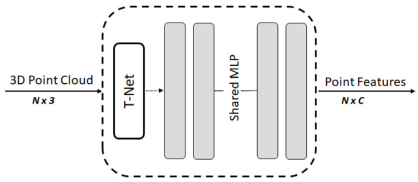


Fig. 2. The adopted PointNet [47] architecture used to extract point cloud features. For each point $\mathbf{P} = \{p_i \mid i = 1, \ldots, N\}$ of the point cloud, the network computes $\mathbf{C}$ features.

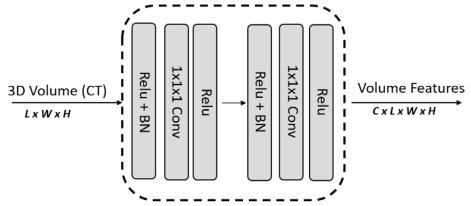


Fig. 3. The CNN architecture used to extract 3D volume features. Given the input volume $\mathbf{V} = \{v_{lwh} \in \mathbb{Z} \mid l = 1, \ldots, L, w = 1, \ldots, W, h = 1, \ldots, H\}$, the network computes the $\mathbf{F_V} \in \mathbb{R}^{LWHC}$ feature map.

4.2. Cross-modal attention siamese architecture

The proposed cross-modal attention block identifies local features and jointly determines the spatial correspondence between the input modalities. The cross-modal module utilizes the modal correlations and adaptively adjusts the modality features for an accurate fusion result. After the representations for each modality have been extracted, the cross-modal attention block captures the distinct parts of one modality given the context features of the other modality as proposed in [49,50]. Rather than considering

features of each modality equally, the proposed cross modal attention block estimates a bidirectional relationship between the input modalities. The cross-modal attention block highlights the important information for one modality related to the other and achieves a inter-modality relationship.

The two input modality feature maps are denoted as $\mathbf{F_P} = \{fp_i \mid i = 1, \ldots, N\}$ and $\mathbf{F_V} = \{fv_{lwh} \mid l = 1, \ldots, L, w = 1, \ldots, W, h = 1, \ldots, H\}$; $\mathbf{F_P}$ and $\mathbf{F_V}$ are the point cloud feature map and the CT volume feature map respectively. The modality feature maps are sent to a siamese architecture of cross-modal attention blocks; each modality feature map will be sent as primary modality to one cross-modal attention block and as cross-modal modality to the second block (see Fig. 1).

Without loss of generality, we will present the cross-modal attention block independently of the input modality context. The block receives a primary input $\mathbf{M_1} \in \mathbb{R}^{CxN}$ and a cross-modal input $\mathbf{M_2} \in \mathbb{R}^{CxLWH}$. C denotes the number of features that have been identified in the previous steps (we use $C = 32$ in our experiments), N and LWH indicate the size of each 3D feature map. The cross-modal attention block computes a new feature map $\mathbf{M_{Cor}}$ that shows the modality correlation, as the sum of the initial primary feature map $\mathbf{M_1}$ and the cross-modal feature map **CM**:

$$\mathbf{M_{Cor}} = \mathbf{CM} + \mathbf{M_1} \tag{2}$$

The cross-modal feature map **CM** shows the corresponding relationship between a position i of the primary input $\mathbf{M_1}$ and all positions j of the cross-modal input $\mathbf{M_2}$ and is computed following [44,51] as an extended non-local operation:

$$\mathbf{CM_i} = \frac{1}{\mathcal{F}} \sum_{j \in \mathbf{M_2}} f(\mathbf{M_2}, \mathbf{M_1}) g(\mathbf{M_1}) \tag{3}$$

Function $f(\mathbf{M_2}, \mathbf{M_1})$ computes the relationship between the feature in the ith position of the first modality and all features j of the second modality. Function g computes a representation of the first modality at position j:

$$f(\mathbf{M_2}, \mathbf{M_1}) = e^{\phi^T(\mathbf{M_{2i}})\theta(\mathbf{M_{1i}})} \tag{4}$$

$$g(\mathbf{M_{1i}}) = \mathbf{W_g M_{1i}} \tag{5}$$

θ, ϕ are also linear embeddings:

$$\theta(\mathbf{M_{1i}}) = \mathbf{W}_\theta \mathbf{M_{1i}} \quad and \quad \phi(\mathbf{M_{2j}}) = \mathbf{W}_\phi \mathbf{M_{2j}} \tag{6}$$

where $\mathbf{W_g}$, $\mathbf{W}_\theta$ and $\mathbf{W}_\phi$ are the weight matrices to be learned during training. $\mathcal{F}$ is a normalization factor of the final result and can be calculated as:

$$\mathcal{F} = \sum_{j \in M_2} f(M_2, M_1). \tag{7}$$

Therefore, $\mathbf{CM_i}$ is calculated as:

$$\mathbf{CM_i} = \frac{e^{\phi^T(\mathbf{M_{2i}})\theta(\mathbf{M_{1i}})}}{\sum_{j \in \mathbf{M_2}} e^{\phi^T(\mathbf{M_{2j}})\theta(\mathbf{M_{1i}})}} \tag{8}$$

which can be estimated by a softmax computation for i along j:

$$\mathbf{CM_i} = softmax_i\left(\phi^T(\mathbf{M_2})\theta(\mathbf{M_1})\right) g(\mathbf{M_1}) \tag{9}$$

This cross-modal attention module plays a vital role when the features to be fused are from different modalities. It preserves the information from each individual modality and makes them complementary to each other so as to eliminate the modality gap. The module's output $\mathbf{M_{Cor}}$ summarizes the features on all locations of the first modality weighted by their correlations with the cross-modal features on the specific location. By using a Siamese network of cross-modal attention blocks, the network investigates the relationships of each modality as both a primary and a cross-modality input and identifies their respective correlations. Fig. 4 shows details of the cross-modal attention block.

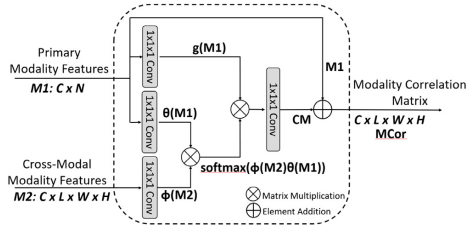


Fig. 4. The detailed architecture of the proposed cross-modal attention module.

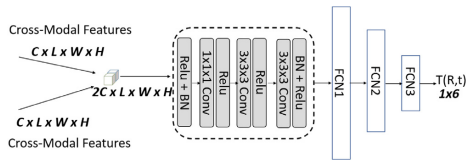


Fig. 5. The detailed architecture of the deep registration module.

4.3. Deep registration block

After computing the spatial correspondences between the input point cloud and volume, the registration block fuses the two sets of feature maps and computes the registration parameters. The deep registration block's architecture is shown in Fig. 5.

The network is supervised by calculating the RMSE (Registration Mean Square Error) between the predicted and the ground truth transformation as the loss function. The loss function of the Deep Registration Module is then back-propagated through all three components and allows the adjustment of the network parameters and the minimization of the error.

5. Evaluation

5.1. Dataset

The proposed fully supervised deep learning method is dependent on sufficient training data with ground truth. The biggest challenge was the lack of a publicly available dataset with ground truth for aligning 3D models from the source modalities of 3D point clouds and 3D micro-CT volumes. The dataset of the PRESIOUS project [52–54], is publicly available and contains 3D models of the modalities of interest. It consists of 17 stone slabs, captured in several modalities across accelerated erosion cycles; the modalities involved are 3D geometry scans (point clouds and 3D meshes), micro-CT volumes, 3D microscopy and petrography. A total of 38 pairs of 3D geometry scans and micro-CT volumes of stone slabs exist.

The use of the PRESIOUS dataset presented a number of challenges. First, the amount of data are limited and insufficient for training our deep network. Moreover, the 3D geometry scans and micro-CT captures were performed independently, without the use of any external reference points; thus the data from the two

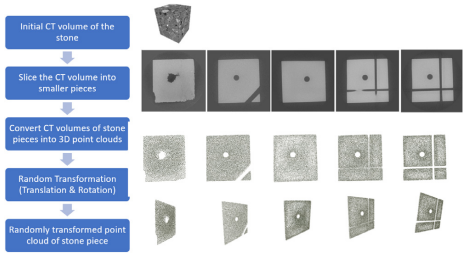


Fig. 6. The process of creating the 3DPCD-CT dataset.

modalities do not possess the necessary ground truth for training our supervised network.

We thus followed a different path in order to expand and augment the cultural heritage dataset of PRESIOUS stones for benchmarking and training multimodal 3D registration algorithms. The process for the creation of the *'3DPCD-CT'* dataset is outlined in Fig. 6. Starting with the micro-CT data of the PRESIOUS stone slabs, we sliced each slab resulting in a larger dataset of subvolumes and then synthetically generated the 3D surface geometry of each piece. Since, the generated 3D point clouds exactly correspond to the respective 3D CT volumes, we consider this as ground truth for training and evaluation purposes.

Every micro-CT volume was divided into a smaller volumes of 50 slices each, providing an average of 35 new smaller volumes. From these smaller volumes, we excluded those with high noise content and no beneficial stone information, resulting in 636 smaller CT volumes, which were then resized to $90 \times 90 \times 50$ voxels each. The corresponding 3D point clouds were then synthetically generated using the marching cubes method of [55]. The outcome consisted of very dense surfaces, so we simplified each model to 13.455 points using the algorithms from [56,57]. The dataset is split into a training set (80% of the dataset) and a test set (20% of the dataset). The training set contains 508 objects and the test set has 128 objects. Each object contains the CT volume, the respective point cloud and the ground truth transformation (see Fig. 7).

5.2. Experimental results

We evaluated our 3D multimodal registration framework on the *'3DPCD-CT'* dataset. Since there is no established performance measure for the registration error between a volume and a geometry surface, we employed the **target registration error (TRE)** [58]. TRE measures the effect of the predicted transformation $\mathbf{T}_{pred}$ against the ground truth transformation $\mathbf{T}_{GT}$ on the initial point cloud $\mathbf{P} = \{p_i \mid i = 1, \ldots, N\}$ based on [59]:

$$\mathrm{TRE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{T}_{pred}p_i - \mathbf{T}_{GT}p_i)\|^2} \tag{10}$$

All tests were run on a PC with an i7-7700K CPU at 4.20 GHz, NVIDIA GeForce GTX 1080 Ti GPU and 32 GB of RAM. In Table 1 we summarize the quantitative registration results on the challenging *'3DPCD-CT'* dataset for multimodal 3D alignment; Fig. 8 illustrates some qualitative results.

An accurate and fair comparison between our method and different literature approaches is not straightforward because we could not identify any previous registration method that *directly* aligns point clouds and CT volumes. We thus used the classic ICP [60] as a baseline, but in order to do so, we pre-processed the CT volumes and converted them into point clouds. We then run the ICP algorithm between these point clouds and the point clouds of the *'3DPCD-CT'* dataset. In general, ICP fails when it comes to large rigid transformation differences. To succeed, ICP needs a good initial transformation estimation (not the case in realistic applications). Thus, in most cases, ICP did not converge. Moreover, ICP and other state of the art registration techniques, requires inputs of the same modality (point clouds in general) necessitating the conversion of one of the inputs in order to address the modality gap. This conversion involves loss of information, which can significantly affect the registration result. In addition, such a conversion can be expensive, especially when large 3D volumes are involved, as in CH applications. For example, in our experiments the conversion of a CT volume into a point cloud representation took approximately 1 h. Conversely, after training, our method requires 0.12 s per registration.

We thus opted for a direct comparison of our method against other multimodal registration methods, even though they may represent different modalities, as this was the nearest we could get to comparing against other methods. Table 2 presents quantitative registration results of the latest state-of-the-art 3D multimodal registration methods. Most of these methods align data of different modalities but of the *same structure*. Of course, the results are only indicative, since each method registers different modalities and the datasets that experiments were conducted on are different and oriented to the specific modalities and task. The table shows the TRE metric as it is considered to be a more generic measure of registration accuracy [58]. In general, TRE is the distance between the corresponding points of the inputs, but due to the fact that the modalities that each method fuses are different, the exact calculation of TRE may differ.

The methods that align different representations of data are [29] and the proposed one (Table 2). [29] aligns 2D images against a 3d model. However this method converts one modality to the other as a first step (the 2D images to a 3D model) and then executes a typical unimodal registration; the conversion involves the penalties of cost [29] and information loss, as also attested by its high TRE. The proposed method directly registers different data modalities and of different structure, which is a more challenging task compared to registering multimodal data of the same structure.

Interestingly the initial TRE, corresponding to the initial pose of the inputs of the compared methods, varies significantly. The results displayed in Table 2 show that the registration error is associated to the difference in initial pose of the inputs.[2] When input modalities start with a pose close to the ideal solution, the initial TRE is lower and so is the registration (final TRE). However, many commonly used registration methods could produce non sufficient results if the modalities are not initialized properly [61].

In an attempt to measure the improvement in alignment of the compared methods, we also calculated the percentage change (PC) in TRE as [62]:

$$\mathrm{PC} = \frac{|TRE - InitTRE|}{InitTRE}100\% \tag{11}$$

Higher values of PC denote a larger improvement on the initial pose. We chose a high initial TRE for the evaluation of our method in order to mimic real, challenging, situations. Taking into consideration the PC of the proposed method and the fact that it operates on modalities of different data structure, the results obtained can be considered as very competitive.

However, there are some cases where our method fails to accurately register the inputs. Such an example is depicted in

[2] Depending on the application and input modalities, an initial pose might be considered as poor if it is within the range of 8 mm and 16 mm [61].

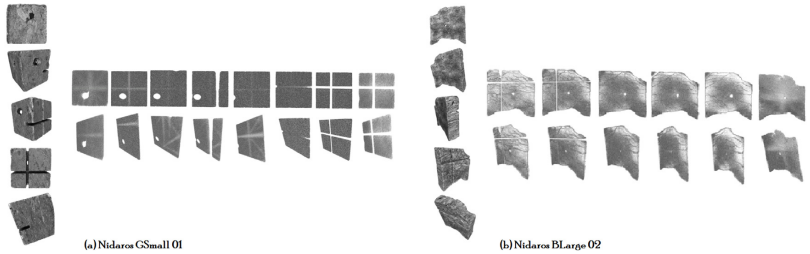


Fig. 7. Example point clouds in the 3DPCD-CT dataset. Two different object cases are shown: a. the Nidaros GSmall 01 stone and b. the Nidaros BLarge 02 stone. For each case it is shown: on the left the whole 3D geometry of the stone and on the right: point clouds of different stone pieces generated from the respective piece of CT-volume.

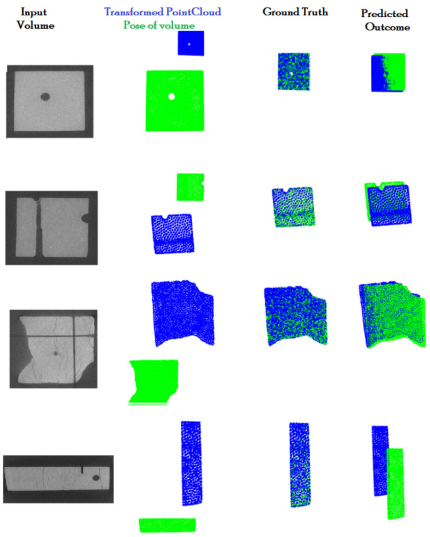


Fig. 8. Example multimodal registration outcomes for the proposed method.

the last row of Fig. 8, where the initial pose of the inputs was considerable, both in terms of rotation and translation; although the method determined the proper rotation it failed to detect the correct translation.

The modified registration Siamese network proposed here is the first registration mechanism that attempts to align two different data modalities not only in terms of data type but *data structure* as well. In this light, the achieved results can be considered as satisfactory as well as promising. For example, the work of [44] which also uses a cross-modal attention block to register MRI and TRUS data, achieves comparable registration results and has competitive computational cost. [44] achieves target registration error between the surfaces of 3.63 and a PC of 54%. However, MRI and TRUS have the same structure (sequences of images), so the network uses the same feature extractor for representing both input volumes. Moreover, method [44] seems to be more efficient in terms of run-time; since this involved absolute execution time based on specific experiments and datasets, we do not think that it represents a conclusive comparison against the proposed. Our method deals with high resolution input data of different structures, thus the search for spatial correspondences through the cross-modal block increases the computational cost.

3D volume modalities (i.e. CT, MRI, TRUS) contain details about the inner structure of the object, like cracks, porosity and voids. Methods like [22,44,45] can detect and use contextual information based on the respective intensities in order to fuse different modalities of 3D volumes. On the other hand, 3D models contain a precise representation of the external surface of the object. A conversion from one modality to the other might result in information loss that will significantly affect the registration result. For example, a 3D model of the surface lacks information of the inner details, so a conversion will not contain any valuable contextual information of the interior and this is likely to affect the registration result. Conversely, a conversion of a 3D volume to a 3D model might add extra computational time without the respective benefit on registration accuracy.

5.3. Ablation study

To demonstrate the contribution of the proposed framework and to validate the effectiveness of each component we executed three different trials of our network by excluding a different module each time.

The results are shown in the lower part of Table 1. It can be seen that removing any of the components has strongly diminutive effects in the registration accuracy; removing the cross-modal attention module results in the worst loss.

6. Conclusions and future work

In this work, we present a direct solution for the challenging task of 3D multimodal registration between 3D volumes and 3D point clouds. A novel deep network that consumes and fuses different 3D modalities (CT-volumes and point clouds) is proposed. These modalities are treated directly (no conversion of one onto the other) to avoid information loss and time penalty. Our network introduces a novel siamese architecture of cross-modal attention blocks that captures and fuses features of two structurally different modalities.

We believe that this approach is an important step forward as it addresses the non-trivial task of aligning modalities of different structural and physical principles, for which it is also extremely challenging to write traditional (non deep learning) code. The method presented can potentially be extended to other computer vision tasks, such as multimodal retrieval and recognition. Moreover, it can be generalized to different modalities due to its adjustable framework. Using alternative feature extraction

methods suitable per modality, the method can be extended to fuse modalities such as 3D meshes, voxel data or medical imaging modalities such as MRI, 3D TRUS etc.

CRediT authorship contribution statement

E. Saiti: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Software, Writing – original draft. **T. Theoharis:** Conceptualization, Funding acquisition, Project administration, Methodology, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 813789. Additionally, the authors would like to thank the NTNU IDUN computing cluster [63] for the provision of additional computing resources and the PRESIOUS project (European Union's Seventh Framework Program for research, technological development, and demonstration under grant no. 600533) for giving access to the datasets used in this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Liu A-A, Guo F-B, Zhou H-Y, Yan C-G, Gao Z, Li X-Y, Li W-H. Domain-adversarial-guided siamese network for unsupervised cross-domain 3-D object retrieval. IEEE Trans Cybern 2022.
[2] Chen Z, Jing L, Liang Y, Tian Y, Li B. Multimodal semi-supervised learning for 3D objects. 2021, arXiv preprint arXiv:2110.11601.
[3] Joaquim MJFAF, Jorge A. Towards 3D modeling using sketches and retrieval. In: Eurographics workshop on sketch-based interfaces and modeling 2004. Citeseer; 2004, p. 127.
[4] Ruan Y, Lee H-H, Zhang K, Chang AX. TriCoLo: TRimodal contrastive loss for fine-grained text to shape retrieval. 2022, arXiv preprint arXiv:2201.07366.
[5] Yin T, Zhou X, Krähenbühl P. Multimodal virtual point 3D detection. Adv Neural Inf Process Syst 2021;34.
[6] Hess M, Petrovic V, Meyer D, Rissolo D, Kuester F. Fusion of multimodal three-dimensional data for comprehensive digital documentation of cultural heritage sites. In: 2015 digital heritage, Vol. 2. IEEE; 2015, p. 595–602.
[7] Hervella AS, Rouco J, Novo J, Ortega M. Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement. Procedia Comput Sci 2018;126:97–104.
[8] Zhan K, Fritsch D, Wagner J. Integration of photogrammetry, computed tomography and endoscopy for gyroscope 3D digitization. Int Arch Photogramm Remote Sens Spat Inf Sci 2021;46:925–31.
[9] Ramos MM, Remondino F. Data fusion in cultural heritage-A review. Int Arch Photogramm Remote Sens Spat Inf Sci 2015;40(5):359.
[10] Mannes D, Schmid F, Frey J, Schmidt-Ott K, Lehmann E. Combined neutron and X-ray imaging for non-invasive investigations of cultural heritage objects. Physics Procedia 2015;69:653–60.
[11] Feng D, Haase-Schütz C, Rosenbaum L, Hertlein H, Glaeser C, Timm F, Wiesbeck W, Dietmayer K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans Intell Transp Syst 2020;22(3):1341–60.
[12] Cui Y, Chen R, Chu W, Chen L, Tian D, Li Y, Cao D. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Trans Intell Transp Syst 2021.
[13] Fitzpatrick JM, Hill DL, Maurer CR, et al. Image registration. Handb Med Imaging 2000;2:447–513.
[14] Adamopoulos E, Rinaudo F. 3D Interpretation and fusion of multidisciplinary data for heritage science: A review. In: 27th CIPA international symposium-documenting the past for a better future, Vol. 42. International Society for Photogrammetry and Remote Sensing; 2019, p. 17–24.
[15] Seales B. The virtues of virtual unrolling. Herculaneum Archaeol: Newsl Friends Herculaneum Soc 2005;3:4–5.
[16] Boust C, Lambert E, Hochart C, Mille B. X-ray tomography and aggregated analysis for bavay treasure bronze statuettes analysis. In: Optics for arts, architecture, and archaeology vii, Vol. 11058. International Society for Optics and Photonics; 2019, p. 110580J.
[17] Scopigno R, Cignoni P, Pietroni N, Callieri M, Dellepiane M. Digital fabrication techniques for cultural heritage: a survey. In: Computer graphics forum, Vol. 36. Wiley Online Library; 2017, p. 6–21.
[18] Payne EM. Imaging techniques in conservation. J Conserv Museum Stud 2013;10(2).
[19] CHANGE EU Project 2019–2023. 2022, https://change-itn.eu/ Accessed on May 2022.
[20] IMPACT4Art Project. 2022, https://www.nicas-research.nl/projects/impact4art/ Accessed on May 2022.
[21] Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: A survey. Inf Fusion 2019;45:153–78.
[22] Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady M, Schnabel JA. MIND: MOdality independent neighbourhood descriptor for multi-modal deformable registration. Med Image Anal 2012;16(7):1423–35.
[23] Ghassemian H. A review of remote sensing image fusion methods. Inf Fusion 2016;32:75–89.
[24] Saiti E, Theoharis T. An application independent review of multimodal 3D registration methods. Comput Graph 2020;91:153–78.
[25] Andrade N, Faria FA, Cappabianco FAM. A practical review on medical image registration: From rigid to deep learning based approaches. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE; 2018, p. 463–70.
[26] Glocker B, Sotiras A, Komodakis N, Paragios N. Deformable medical image registration: setting the state of the art with discrete methods. Annu Rev Biomed Eng 2011;13:219–44.
[27] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: A survey. IEEE Trans Med Imaging 2013;32(7):1153–90.
[28] Ma J, Jiang X, Fan A, Jiang J, Yan J. Image matching from handcrafted to deep features: A survey. Int J Comput Vis 2021;129(1):23–79.
[29] Pintus R, Gobbetti E. A fast and robust framework for semiautomatic and automatic registration of photographs to 3D geometry. J Comput Cult Herit (JOCCH) 2015;7(4):1–23.
[30] Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 1997;16(2):187–98.
[31] Klein S, Van Der Heide UA, Lips IM, Van Vulpen M, Staring M, Pluim JP. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med Phys 2008;35(4):1407–17.
[32] Corsini M, Dellepiane M, Ponchio F, Scopigno R. Image-to-geometry registration: a mutual information method exploiting illumination-related geometric properties. In: Computer graphics forum, Vol. 28. Wiley Online Library; 2009, p. 1755–64.
[33] Moreno-Noguer F, Lepetit V, Fua P. Pose priors for simultaneously solving alignment and correspondence. In: European conference on computer vision. Springer; 2008, p. 405–18.
[34] Liu X, Jiang D, Wang M, Song Z. Image synthesis-based multi-modal image registration framework by using deep fully convolutional networks. Med Biol Eng Comput 2019;57(5):1037–48.
[35] Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. Mach Vis Appl 2020;31(1):1–18.
[36] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. Phys Med Biol 2020;65(20):20TR01.
[37] Boveiri HR, Khayami R, Javidan R, Mehdizadeh A. Medical image registration using deep neural networks: A comprehensive review. Comput Electr Eng 2020;87:106767.
[38] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6(1):1–48.
[39] Yu Z, Yu J, Xiang C, Zhao Z, Tian Q, Tao D. Rethinking diversified and discriminative proposal generation for visual grounding. 2018, arXiv preprint arXiv:1805.03508.
[40] Wu Y, Wang S, Song G, Huang Q. Learning fragment self-attention embeddings for image-text matching. In: Proceedings of the 27th ACM international conference on multimedia, 2019, pp. 2088–2096.
[41] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. In: International conference on machine learning. PMLR; 2019, p. 7354–63.
[42] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. PMLR; 2015, p. 2048–57.
[43] Zou L, Huang Z, Wang F, Yang Z, Wang G. CMA: CRoss-modal attention for 6D object pose estimation. Comput Graph 2021;97:139–47.
[44] Song X, Guo H, Xu X, Chao H, Xu S, Turkbey B, Wood BJ, Wang G, Yan P. Cross-modal attention for MRI and ultrasound volume registration. In: International conference on medical image computing and computer-assisted intervention. Springer; 2021, p. 66–75.

[45] Heinrich MP, Jenkinson M, Papież BW, Brady SM, Schnabel JA. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: International conference on medical image computing and computer-assisted intervention. Springer; 2013, p. 187–94.
[46] Arar M, Ginger Y, Danon D, Bermano AH, Cohen-Or D. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13410–13419.
[47] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.
[48] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015, p. 234–41.
[49] Chen H, Ding G, Liu X, Lin Z, Liu J, Han J. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12655–12663.
[50] Maleki D, Tizhoosh H. LILE: Look in-depth before looking elsewhere–a dual attention network using transformers for cross-modal information retrieval in histopathology archives. 2022, arXiv preprint arXiv:2203.01445.
[51] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
[52] PRESIOUS FP7-600533 EU Project -FinalEvaluationReport. 2022, http://presious.eu/file_downloads/PRESIOUS-D5.8-FinalEvaluationReport.pdf Accessed on May 2022.
[53] Theoharis T, Papaioannou G. PRESIOUS 3D cultural heritage fragments. 2013, URL: http://presious.eu/resources/**3d-data-sets.
[54] Theoharis T, Rieke-Zapp D. PRESIOUS 3D cultural heritage differential scans. 2013, URL: http://presious.eu/resources/3d-data-sets.
[55] Lewiner T, Lopes H, Vieira AW, Tavares G. Efficient implementation of marching cubes' cases with topological guarantees. J Graph Tools 2003;8(2):1–15.
[56] Low K-L, Tan T-S. Model simplification using vertex-clustering. In: Proceedings of the 1997 symposium on interactive 3D graphics, 1997, pp. 75–ff.
[57] Yuksel C. Sample elimination for generating Poisson disk sample sets. In: Computer graphics forum, Vol. 34. Wiley Online Library; 2015, p. 25–32.
[58] Maurer CR, Fitzpatrick JM, Wang MY, Galloway RL, Maciunas RJ, Allen GS. Registration of head volume images using implantable fiducial markers. IEEE Trans Med Imaging 1997;16(4):447–62.
[59] Saiti E, Danelakis A, Theoharis T. Cross-time registration of 3D point clouds. Comput Graph 2021;99:139–52.
[60] Besl PJ, McKay ND. Method for registration of 3D shapes. In: Sensor fusion iv: Control paradigms and data structures, Vol. 1611. International Society for Optics and Photonics; 1992, p. 586–606.
[61] Haskins G, Kruecker J, Kruger U, Xu S, Pinto PA, Wood BJ, Yan P. Learning deep similarity metric for 3D MR–TRUS image registration. Int J Comput Assist Radiol Surg 2019;14(3):417–25.
[62] Kaiser L. Adjusting for baseline: change or percentage change? Stat Med 1989;8(10):1183–90.
[63] Själander M, Jahre M, Tufte G, Reissmann N. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. eprint arXiv:1912.05848, 2019.

Chapter 10

Paper E - A pipeline for monitoring the external and inner structure of cultural heritage objects

Authors

Evdokia Saiti, and Theoharis Theoharis.

Under review in

Proceedings of Conference *Archiving*, 2023

Chapter 11

Paper F - Adaption of imaging techniques for monitoring cultural heritage objects

Authors

Siatou Amalia, Athanasia Papanikolaou, and Evdokia Saiti.

Published In

Advanced Nondestructive and Structural Techniques for Diagnosis, Redesign and Health Monitoring for the Preservation of Cultural Heritage: Selected work from the TMM-CH 2021, Cham: Springer International Publishing, 2022

Adaption of Imaging Techniques for Monitoring Cultural Heritage Objects

Amalia Siatou[1,2( )], Athanasia Papanikolaou[3], and Evdokia Saiti[4]

[1] HES-SO, Haute Ecole Arc Conservation-Restauration, Neuchâtel, Switzerland
amalia.siatou@he-arc.ch
[2] Laboratory of Imaging and Artificial Vision, UBFC, Dijon, France
[3] Faculty of Mechatronics, WUT, Warsaw, Poland
Athanasia.Papanikolau@pw.edu.pl
[4] Department of Computer Science, NTNU, Trondheim, Norway
evdokia.saiti@ntnu.no

Abstract. The paper describes the ongoing research on an interdisciplinary approach regarding the technological developments adapted for monitoring CH objects. It covers aspects from data capturing, to data processing and cross-time registration methodologies. The work of three individual projects, that are carried out in the framework of ITN-CHANGE (Horizon 2020, GA 813789) project, are presented. These projects are based on the different backgrounds and expertise of the co authors which, when combined, can cover a wide spectrum of information indispensable for the accurate monitoring of CH objects. The potentiality of 3D Digital Image Correlation (3D DIC) for monitoring in and out of plane displacements as well as advances in Reflectance Transformation Imaging (RTI) for data processing for monitoring specular surfaces, are examined. Computational cross-time and multi-modal registration algorithms are developed for correlating 3D non-registered data over-time. Feasibility studies on mock-ups and simulated data are presented for the validation of the adapted methodologies.

Keywords: Monitoring · Cross-time registration · Reflectance transformation imaging · 3D Digital Image Correlation

1 Introduction

The interaction of Cultural Heritage (CH) objects with the environment can result in changes of their physical properties and their appearance attributes. Extensive research has been carried out on imaging methods for understanding,

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 813789. A. Papanikolaou greatly appreciates the financial support granted by the Scientific Council of the Discipline Automatic, Electronics and Electrical Engineering, WUT, grant agreement No. 504/04542/1143/43.020004.

A. Osman and A. Moropoulou (Eds.): *Advanced Nondestructive and Structural Techniques for Diagnosis, Redesign and Health Monitoring for the Preservation of Cultural Heritage*, SPM 16, pp. 38–47, 2022.
https://doi.org/10.1007/978-3-031-03795-5_6

documenting and monitoring these type of changes. Imaging techniques provide powerful tools for capturing and tracking CH object alterations [1]; however, monitoring is possible only with appropriate data processing through computational methods and data interpretation by experts [2,3]. CHANGE" [4] is a research and innovation program under the auspices of the European Union's Horizon 2020 programme. The project aims at developing imaging techniques and systems for the acquisition, documentation and monitoring of CH objects. The final goal is to combine and correlate expertise of different fields under a common framework. To this sense, this paper draws upon knowledge from the fields of optical metrology, computer engineering and conservation science in an interdisciplinary approach for the development of change detection methodologies. The scientific activities are grouped into three pillars which cover different technological aspects. The first pillar refers to the strategies and systems for capturing and tracking changes on CH objects, whereas the second applies computational methods for studying cross-time changes and the last presents feasibility studies for the validation of protocols developed within the other two pillars. Currently, each pillar advances independently, with the aim to merge with the perspective to merge them as the project evolves. The remainder of the paper is organized as follows: In Sect. 2, *change capture and tracking strategies are discussed* while in Sect. 3 the problem of *data registration* is analyzed. Section 4 covers *data interpretation* and the paper concludes in Sect. 5 *with discussion and future aspects.*

2 Change Capture and Tracking Strategies

The first pillar describes technological tools and methodologies adapted for data acquisition at different time intervals using 3D DIC and RTI.

2.1 Digital Image Correlation (DIC)

DIC is a versatile, full-field, optical metrology technique with applications mainly in mechanical and civil engineering. Typically, 3D DIC is used to study heterogeneous materials under different loading conditions and to accurately calculate the maps of in and out of plane displacements and strains [5]. In plane displacements correspond to deformation in the X and Y axes, and out of plane to Z, providing thus the arbitrary 3D vector of displacements. While 3D DIC a is portable, non-invasive and low-cost technique, with an adjustable field of view, its application to CH objects can be challenging. In particular, to achieve optimum results and sub-pixel accuracy, this technique requires the surface under investigation to have a random texture that meets specific criteria (e.g. randomly positioned speckles with adequate contrast, firmly adherent pattern, etc.) [5].

A common 3D DIC set-up consists of two cameras simultaneously capturing pairs of images of the object under a certain geometrical configuration. Appropriate calibration protocols are applied to enable the triangulation and correlation of the captured data set (Fig. 1). The correlation algorithm works by detecting

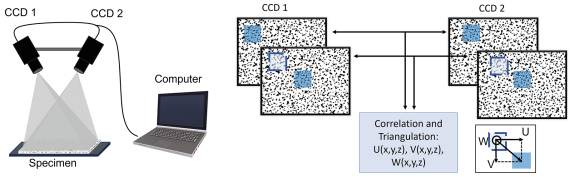


Fig. 1. Representation of (a) 3D DIC configuration and (b) working principle

intensity differences among adjacent groups of pixels (called subsets). Each subset is then localized through the subsequent images that correspond to different deformation stages, with a specific searching step. Both the subset and step values are user-defined parameters that need to be adjusted according to the object under investigation and the experimental configuration (i.e. the optical magnification, density and size of the speckle pattern). For the correlation to be feasible, each subset of the image should contain a unique pattern, thus enabling the calculation of the displacement and strains. The accuracy of the calculations and data processing time depends on the selection of subset and step size. For many CH objects, monitoring of the deformations and strains is considered necessary to understand surface alterations. Here, the 3D DIC technique can be a solution, as it can be used *in situ*, in full field-of-view (FoV) and at selected time intervals. Nevertheless, introducing an artificial speckle pattern on a CH object is not permitted. Thus, a compromise to the calculation accuracy is necessary or alternative approaches might be considered [6]. Some examples of 3D DIC applications to CH objects include displacements measurement on model canvas with random pattern [7–9], and historical parchment [10] provoked by controlled changes in the relative humidity, as well as, mechanical displacements induced on canvas paintings [6,7].

2.2 Reflectance Transformation Imaging (RTI)

RTI is a multi-light technique following a fixed configuration with a camera positioned perpendicular to an object for acquiring a set of images at different light angles (Fig. 2). This multi-angle illumination can provide photometric and geometric documentation of surfaces [11]. It has found application in CH as an easy-to-use, non-invasive, portable technique [12]. There are numerous references on the application of RTI in CH, varying in methodologies and material applications; however, most address enhancing legibility and surface details related to topography, such as examining artists' brush-strokes, or deciphering epigraphs [13,14]. In this section the feasibility of monitoring of objects is investigated. A dome with a fully calibrated light source and motorized camera functions is used to ensure the reproducibility and cross-time registration of the acquired data [15,16]. One of the system's novelties is the ability to extract raw RTI data, providing the possibility to further data processing that goes beyond the simple visualization or image enhancement [16]. In particular, geometric and statistical calculations of the stack of images produced can provide information related to

the surface topography or the per-pixel reflectance response of the surface at different light angles, respectively. This results in visualization through maps depicting the surface features (features maps) that can either enhance or isolate surface information.

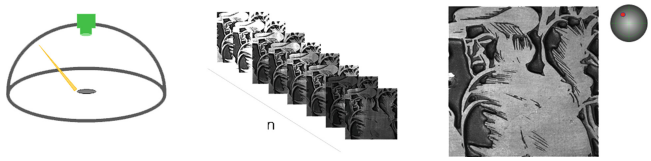


Fig. 2. Simplified representation of RTI from left to right: The dome system (left). Acquisition of a stack of images at "n" light positions (middle). Relightable images at selected light direction (right).

In this paper monitoring the condition of specular surfaces was selected. A global examination was followed, consisting of data acquisitions of artificially aged mock-ups at different time intervals. The methodology consists of data acquisitions in different RTI-modalities and correlation of the results based on visual inspection and comparison to imaging techniques routinely used in the field of CH. To ensure repeatability, the systems used and the acquisition parameters remain stable throughout the experimental process. Interpretation of results relies on CH expertise.

3 Computational Methods for Cross-Time Registration

Following the 3D model acquisition of the CH objects, the data need to be analyzed in order to accurately understand and monitor any change [17]. In general, data captured from different acquisitions can be geometrically and chronologically incoherent. In order to facilitate their study and detect changes, the data need to be registered. Registration aims to find the transformation (rotation and translation) that optimally aligns two or more instances of the same objects at different times (cross-time data), from different viewpoints (multi-view data) or by different modalities (multi-modal data) in order to bring the data in a common reference frame [18].

3.1 Cross-time Registration

Methods that monitor the geometric change of an object over time, try to compare the 3D representation of the same object captured at different time intervals. Considering that modifications may have occurred on the surface of the object (i.e. surface alteration due to weathering or conservation-restoration treatments), shape differences may have encountered between acquisitions, resulting

to a non-trivial correspondence of the object's surface. Moreover, acquisition processes cannot always ensure that cross-time captures will be at the exact same position. Thus, accurate 3D spatial relations between data from different acquisitions may not be directly obtained, which makes the cross-time registration process a challenging task.

Given two 3D point clouds of the same object, but captured at different time frames, the aim of the 3D cross-time registration is to find the unknown rigid transformation so as to align the two point clouds as accurately as possible. Currently, our research is focused on the surface alterations due to weathering, where the examined object is assumed to have been uniformly exposed to weather conditions, both spatially and temporally. A framework for cross-time 3D registration is proposed in [19] that copes with big data using a down-sampling scheme that is appropriate for objects exhibiting uniform change over time. The proposed method generally outperforms the state-of-the-art in both accuracy and efficiency (Fig. 7).

4 Data Interpretation on Selected Case-Studies

The final part is the application of the above described methodologies to CH objects.

4.1 3D DIC Data Interpretation

Examples of 3D DIC application to CH objects with inherent surface texture and patterns, adequate to perform the analysis, are presented. The first case study is an oil painting on canvas that was subjected to deformation by applying mechanical pressure (loading, simply by pushing outwards) on its back surface. The pressure was applied with the intent to create a random and complex deformation distribution, in order to examine the effectiveness of the 3D DIC and its effectiveness on detecting changes using the surface features of the painting and without applying artificial texture patterns. The painting and the results of the 3D DIC study of the loading are presented in Fig. 3.

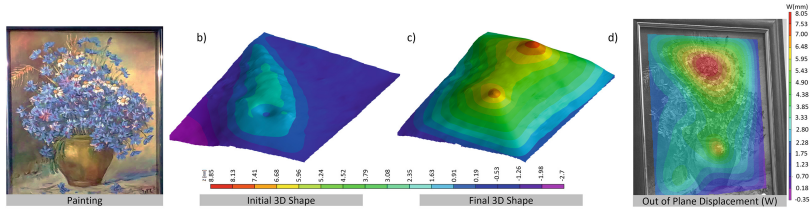


Fig. 3. 3D DIC analysis of a painting without an artificial pattern. The 3D shape maps before (b) and after loading (c), along with the 2D map out of plane displacement map (d) are presented.

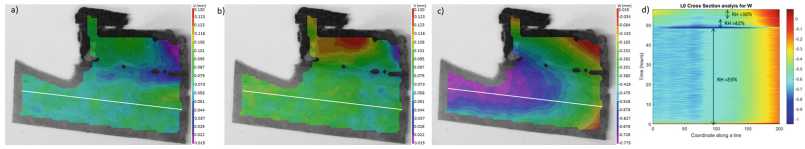


Fig. 4. 3D DIC analysis on a parchment without an applied artificial texture. The in- (U and V in a, and b respectively) and out-of-plane (W in c) displacement maps corresponding to 50% of RH are shown. In 4d the spatio-temporal analysis of W for the line corresponding to 60h duration is presented.

The second case study is the calculation of displacements in a historical parchment subjected to an environment with fluctuating Relative Humidity (RH) (Fig. 4). The in and out of plane displacements, that correspond to 50% RH, are presented along with a selected line which is analysed through time. The spatio-temporal map of the out of plane displacement (W), that corresponds to the selected line, is shown in (Fig. 4 d). During the observation time RH is gradually decreased over a duration of 60 h.

These objects represent two important groups of CH artifacts that are sensitive to environmental fluctuations and mechanical damage. For both examples, 3D DIC has provided valuable information for characterizing the surface displacements caused by different factors, detailed information can be found in [6,10]. These case studies represent short-time measurement sessions (with stable object-sensor configuration), which eliminates the need for cross-time data alignment. They, support the potential of condition monitoring of CH objects, as well as, the capability of monitoring conservation-restoration treatments with 3D DIC.

4.2 RTI Application and Interpretation

RTI was investigated for its feasibility to monitor the formation fingerprints during the tarnishing of silver objects [20]. Data were acquired for extracting information on surface appearance attributes, with the goal to isolate information related to the topography and identify characteristics related to the reflectance response. A fingerprint was placed on pure silver coupons, with an isotropic texture, and were subsequently artificially tarnished at different levels, corresponding to change over time. The surfaces were examined both with techniques routinely used for CH object documentation as well as the proposed RTI methodology [20].

The routine imaging techniques consist of calibrated photography using a light-box and correcting data through a color-checker, documentation of the surface under high magnification using optical microscopy and color measurements with a spectrophotometer.

The RTI experimental set-up involved a dome and acquisitions were preformed using a monochromatic camera; whereas, multi-light illumination was achieved using a single light source (high-power collimated white LED light).

For each acquisition set, 150 light positions were acquired covering an azimuth angle from 0° to 360° and an elevation angle of 5°–60°. All acquisitions were homogeneous, i.e., the lighting positions were spread uniformly around the dome covering an entire hemisphere and thus providing overall angular illumination of the surface. To ensure repeatability, exposure time, acquisition parameters and selected ROI (region of interest) were kept constant for the different tarnish levels. Data processing consisted of calculating the per-pixel mean reflectance response of the stack of images and visualising the results through gray-scale colormaps.

Figure 5 presents an example of the comparison between different imaging techniques for registering information related to monitoring the cross-time surface change of fingerprints on silver. Despite the different scales presented for each technique, the possibility of enhancing or isolating specific information related to the change of the reflectance response of the surface, in the area of the fingerprint, is evident in the mean reflectance response of the RTI data, in a form of gray-scale colormap, even at light levels of tarnish.

Global examination of feature maps, at different levels of silver tarnishing in the presence of fingerprints, has shown promising results in the ability to detect and enhance visualization in comparison to routinely used imaging techniques or usual RTI visualization and surface enhancement. Furthermore, from the CH perspective, the detection of fingerprints at low levels of tarnish, which is difficult to document through regular inspection, provides a tool for non-invasive examination of CH surfaces. However, for quantification of results, further data processing is necessary to better evaluate surface change over time and to apply actual cross-time registration on the examined surfaces.

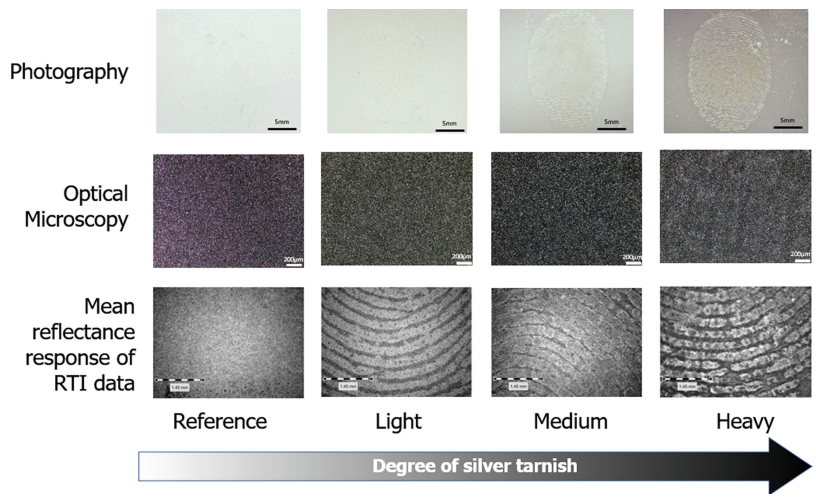


Fig. 5. Monitoring silver tarnishing by RTI features maps and comparison to the routinely used imaging techniques of photography and optical microscopy.

4.3 Cross-time Registration Data Interpretation

The main challenge in cross-time registration is the lack of a publicly available dataset with altered objects with the respective ground truth. In order to overcome this, a dataset of weathered CH objects was synthetically created. Starting from the publicly available dataset [21], first a random rotation and translation were applied to the objects; then two weathering effects were simulated and applied on the transformed objects to create the relative surface alteration. The effects simulated are the dry deposition of crust due to pollution and the recession by acid rain. These effects can result in gain or loss of material on the surface of an object.

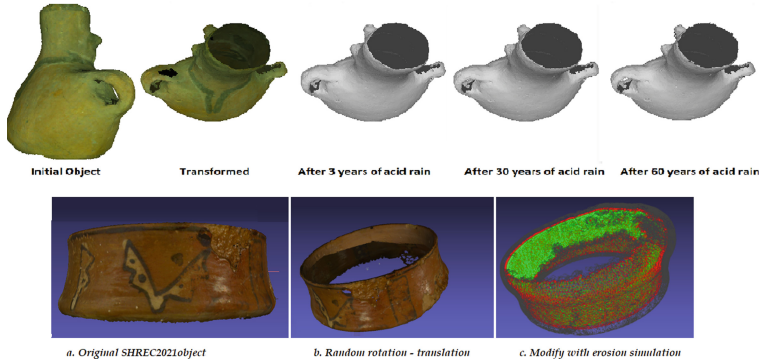


Fig. 6. Up: The steps of the dataset creation for one object. The object is initially transformed and then the erosion simulator runs for 20 epochs of 3 years each. In this example, the initial model is shown degraded due to the effect of acid rain after 3, 30 and 60 years. Down:Simulated dataset for weathering. Original CH object from SHREC2021 dataset (a), along with the transformed instance (b). On (c), different weathered data are depicted. The reference object is depicted in gray color, the object after 30 years of ageing in red and the after 60 years in green.

Since weathering is performed *in situ* and the transformation parameters are known, the ground truth for benchmarking cross-time registration algorithms can be acquired. The process is outlined in Fig. 6. The training part of the dataset is then used to train our deep network to register weathered objects. The proposed method (Fig. 7) first down-samples the reference and weathered point clouds using their principal curvatures. Then, the down-sampled point clouds are segmented and finally, the registration is performed by aligning the component centroids of each segment.

5 Discussion and Future Aspects

In summary, different imaging techniques and methodologies are being examined and adjusted by the authors to meet the requirements for monitoring CH objects

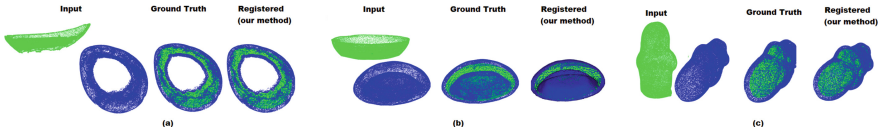


Fig. 7. Examples of the simulated weathered dataset along with the results of the proposed methodology for cross-time registration.

in terms of set-ups, proper registration and data processing. For the accurate interpretation of the results with the developed methodologies, a key factor is proper data assessment in collaboration with CH specialists. As a future step, we consider the combination of the presented methodologies in an end-to-end application, in order to facilitate the process of monitoring CH objects.

Through the interdisciplinary approach of the ITN-CHANGE project different perspectives in terms of technologies and expertise are contributing in the adaptation of new methodologies for monitoring CH objects. The aim of this approach is to connect the information collected from the aforementioned pillars with conservation-restoration strategies related to the long-term preservation of tangible CH with particular focus on monitoring, by capturing and tracking changes. Currently, each pillar is developed independently with the intent to later combine the processes and information under a common framework. To achieve the interdisciplinary goals of the CHANGE project, researchers are in constant dialogue, collaboration and training to familiarize themselves with the different scientific fields so that the needs and requirements of each specialty can be met and understood. The cooperation and teamwork of scientists with different backgrounds is considered a necessary link for the adaptation of technologies from different fields to CH.

References

1. MacDonald, L.: Digital Heritage: Applying Digital Imaging to Cultural Heritage. Routledge, Milton Park (2006)
2. Pintus, R., et al.: Geometric analysis in cultural heritage. In: GCH, pp. 117–133 (2014)
3. Saha, S., Siatou, A., Sitnik, R.: Classification of surface geometry behavior of cultural heritage surfaces based on monitoring change. In: Optics for Arts, Architecture, and Archaeology VIII. vol. 11784. International Society for Optics and Photonics (2021)
4. CHANGE cultural heritage analysis for new generations, European union's horizon 2020 research and innovation programme. https://change-itn.eu/. Accessed Apr 2021
5. Sutton, M.A., Orteu, J.J., Schreier, H.W.: Image correlation for shape, motion and deformation measurements : basic concepts, theory and applications. Springer, Boston (2009). https://hal-mines-albi.archives-ouvertes.fr/hal-01729219, https://doi.org/10.1007/978-0-387-78747-3
6. Papanikolaou, A., Garbat, P., Kujawinska, M.: Colour digital image correlation method for monitoring of cultural heritage objects with natural texture. In: Liang, H., Groves, R. (eds.) Optics for Arts, Architecture, and Archaeology VIII, vol.

11784, pp. 166–177. International Society for Optics and Photonics, SPIE (2021). https://doi.org/10.1117/12.2592549
7. Kujawinska, M., et al.: Digital image correlation method: a versatile tool for engineering and art structures investigations. In: Rodríguez-Vera, R., et al. (eds.) 22nd Congress of the International Commission for Optics: Light for the Development of the World, vol. 8011, pp. 2599–2606. SPIE (2011). https://doi.org/10.1117/12.915566
8. Malesa, M., et al.: Application of digital image correlation for tracking deformations of paintings on canvas. In: Pezzati, L., Salimbeni, R. (eds.) O3A: Optics for Arts, Architecture, and Archaeology III, vol. 8084, pp. 157–164. International Society for Optics and Photonics, SPIE (2011). https://doi.org/10.1117/12.889452
9. Malowany, K., et al.: Application of 3d digital image correlation to track displacements and strains of canvas paintings exposed to relative humidity changes. Appl. Opt. **53**(9), 1739–1749 (2014). https://doi.org/10.1364/AO.53.001739
10. Papanikolaou, A., Dzik-kruszelnicka, D., Saha, S., Kujawinska, M.: 3D digital image correlation system for monitoring of changes induced by RH fluctuations on parchment. In: Proceedings of the IS&T International Symposium on Electronic Imaging: 3D Imaging and Applications, pp 65-1–65-7 (2021). https://doi.org/10.2352/ISSN.2470-1173.2021.18.3DIA-065
11. Castro, Y., et al.: A new method for calibration of the spatial distribution of light positions in free-form RTI acquisitions. In: SPIE Optical Metrology, 2019, Munich, Germany, vol. 11058, p. 38. SPIE, Munich, Germany, June 2019. https://doi.org/10.1117/12.2527504, https://hal-univ-bourgogne.archives-ouvertes.fr/hal-02353517
12. CHI 2021: Cultural heritage imaging. http://culturalheritageimaging.org/Technologies/RTI/ Accessed Jun 2021
13. Mudge, M., et al.: Image-based empirical information acquisition, scientific reliability, and long-term digital preservation for the natural sciences and cultural heritage. Eurographics (Tutorials) **2**(4) (2008)
14. Earl, G., et al.: Reflectance transformation imaging systems for ancient documentary artefacts. Electron. Vis. Arts (EVA 2011) 147–154 (2011)
15. Pitard, G., et al.: Discrete modal decomposition: a new approach for the reflectance modeling and rendering of real surfaces. Mach. Vis. Appl. **28** (2017). https://doi.org/10.1007/s00138-017-0856-0
16. Nurit, M., et al.: HD-RTI: An adaptive multi-light imaging approach for the quality assessment of manufactured surfaces. Comput. Ind. **132** (2021). https://doi.org/10.1016/j.compind.2021.103500
17. Saha, S., Foryś, P., Martusewicz, J., Sitnik, R.: Approach to analysis the surface geometry change in cultural heritage objects. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (eds.) ICISP 2020. LNCS, vol. 12119, pp. 3–13. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51935-3_1
18. Saiti, E., Theoharis, T.: An application independent review of multimodal 3D registration methods. Comput. Graph. **91**, 153–178 (2020)
19. Saiti, E., Danelakis, A., Theoharis, T.: Cross-time registration of 3d point clouds. Comput. Graph. **99**, 139–152 (2021)
20. Siatou, A., et al.: Surface appearance assessment as a tool for characterizing silver tarnishing, November 2020. https://doi.org/10.5281/zenodo.4299912
21. Sipiran, I., et al.: Shrec 2021: retrieval of cultural heritage objects. Comput. Graph. **100**, 1–20 (2021)

Chapter 12

Paper G - Automated 3D registration techniques for applications in cultural heritage monitoring

Authors

Evdokia Saiti, and Theoharis Theoharis.

Under review in

CHANGE project's Final Book: Cultural Heritage Analysis for New GEnerations (CHANGE), 2023

ISBN 978-82-326-7010-9 (printed ver.)
ISBN 978-82-326-5814-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology