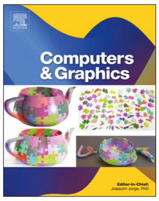
Computers & Graphics 106 (2022) 259–266

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.

© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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
0097-8493/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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	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:

$$\mathbf{T} = \left[\frac{\mathbf{R} \mid \mathbf{t}}{0 \mid 1} \right] \tag{1}$$

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%

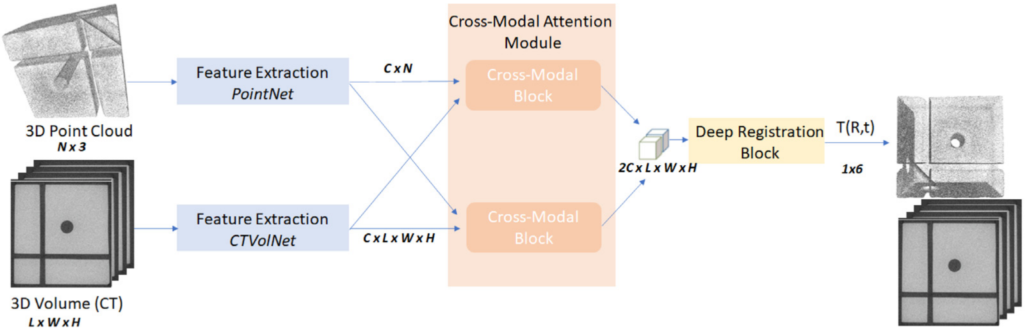


Fig. 1. Overview of the proposed *cross-modal* 3D registration framework. The 3D cross-modal registration network consists of three stages. 1. Each input modality (Point Cloud and 3D CT Volume) is fed into an independent feature extractor network that is suitable for that modality. 2. The captured features 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.

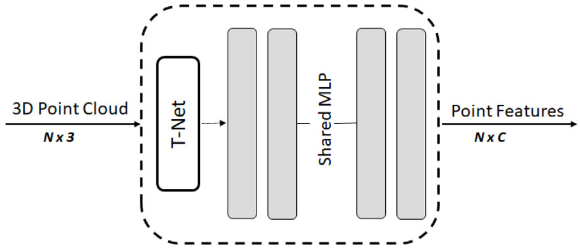


Fig. 2. The adopted PointNet [47] architecture used to extract point cloud features. For each point $\mathbf{P} = \{p_i \mid i = 1, \dots, N\}$ of the point cloud, the network computes $\mathbf{C}$ features.

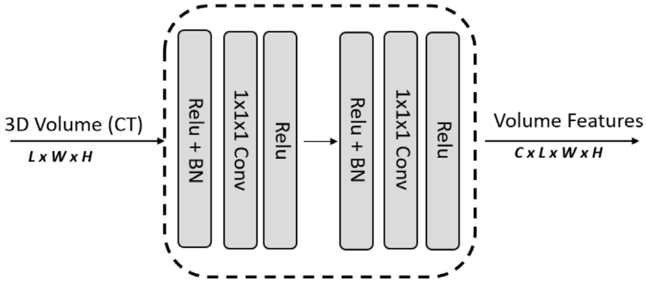


Fig. 3. The CNN architecture used to extract 3D volume features. Given the input volume $\mathbf{V} = \{v_{lwh} \in \mathbb{Z} \mid l = 1, \dots, L, w = 1, \dots, W, h = 1, \dots, H\}$, the network computes the $\mathbf{F_V} \in \mathbb{R}^{LWHxC}$ feature map.

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 $\mathbf{F_P} = \{fp_i \mid i = 1, \dots, N\}$ and $\mathbf{F_V} = \{fv_{lwh} \mid l = 1, \dots, L, w = 1, \dots, W, h = 1, \dots, H\}$; $\mathbf{F_P}$ 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 $\mathbf{CM}$:

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

The cross-modal feature map $\mathbf{CM}$ 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:

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

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

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.

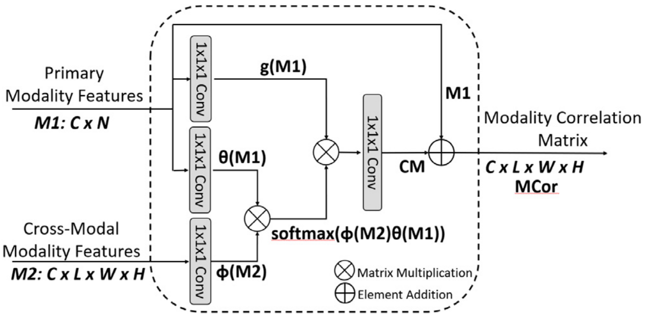


Fig. 4. The detailed architecture of the proposed cross-modal attention module.

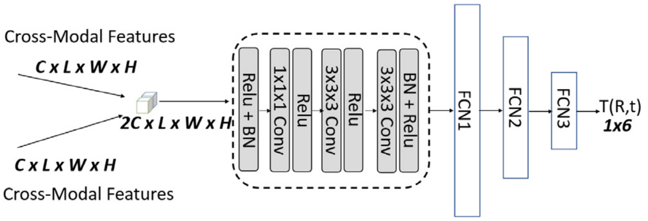


Fig. 5. The detailed architecture of the deep registration module.

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

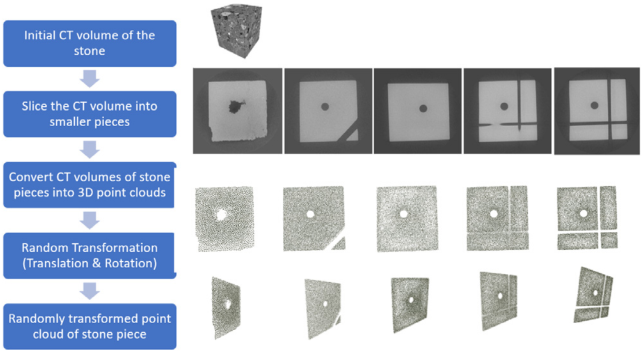

Initial CT volume of the stone
Slice the CT volume into smaller pieces
Convert CT volumes of stone pieces into 3D point clouds
Random Transformation (Translation & Rotation)
Randomly transformed point cloud of stone piece

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 sub-volumes 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 × 90 × 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].

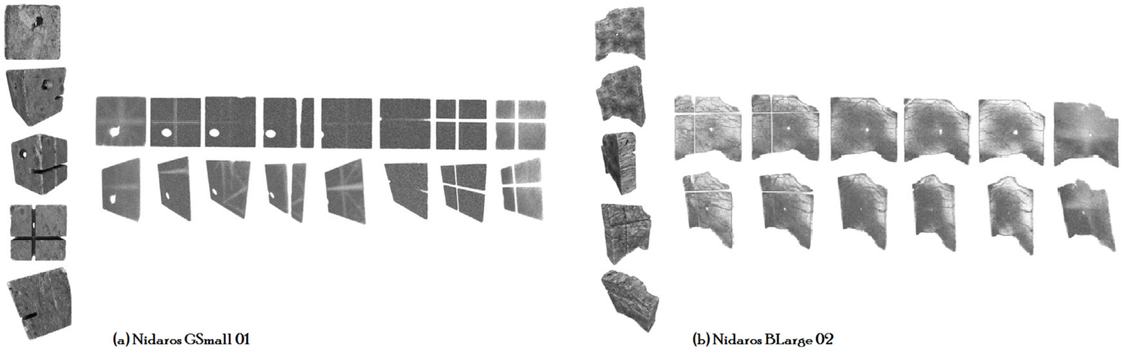
In an attempt to measure the improvement in alignment of the compared methods, we also calculated the percentage change (PC) in TRE as [62]:

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

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].


(a) Nidaros GSmall 01

(b) Nidaros BLarge 02

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.

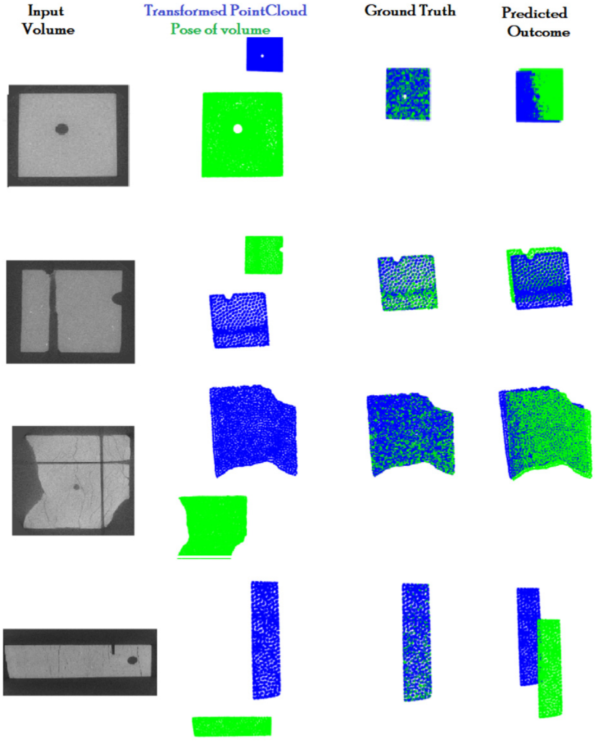


Fig. 8. Example multimodal registration outcomes for the proposed method.

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.