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

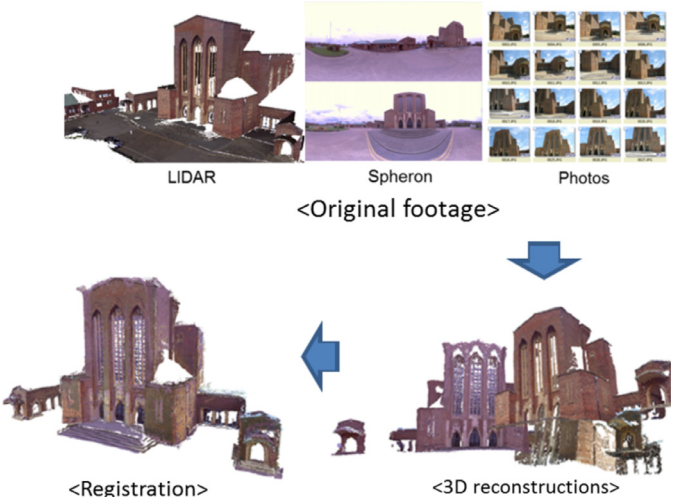

Fig. 1. Multimodal data registration as presented in [13].

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)

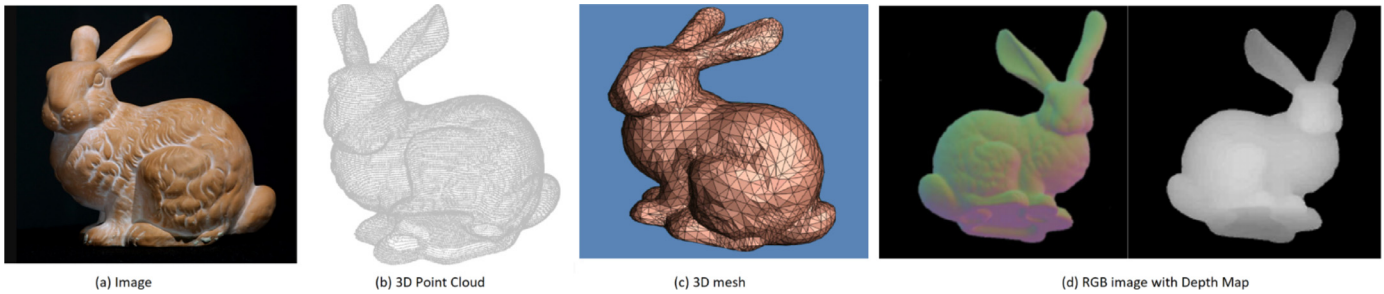

Fig. 2. The Stanford Bunny in Different Modalities as presented in [37–40].

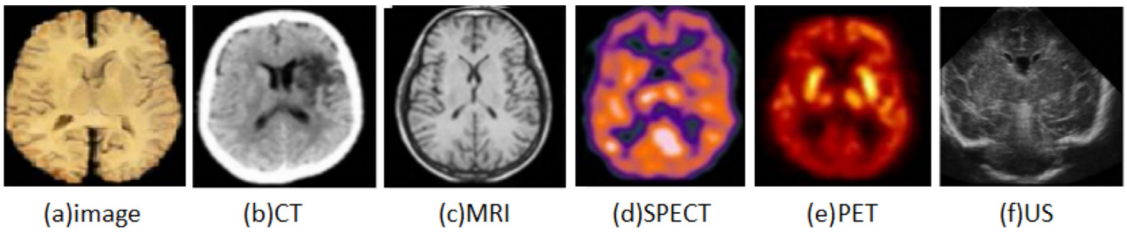

Fig. 3. Different Modality Representations of Brain Anatomy [44].

- 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.

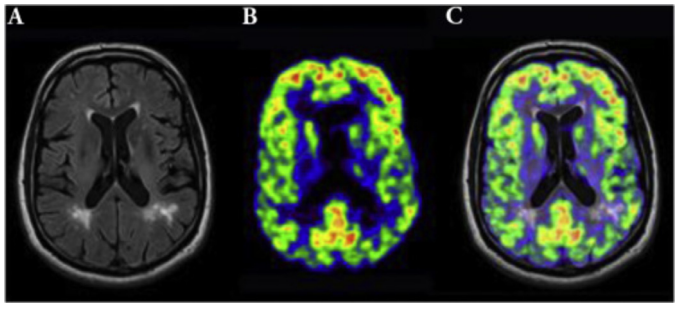

Fig. 4. Medical fusion of MRI and PET modalities. (A) MRI and (B) PET images are registered and fused (C) [50].

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.

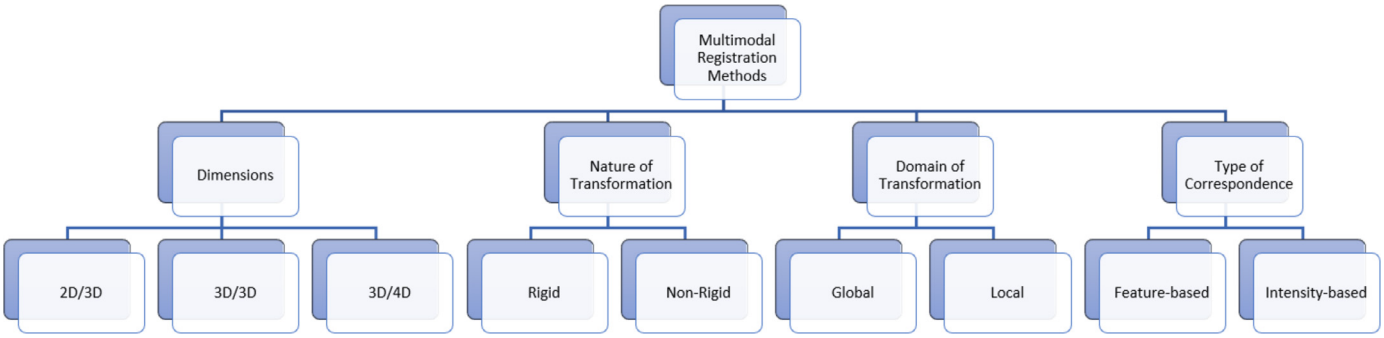

Fig. 5. Attributes of Registration Methods.

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.

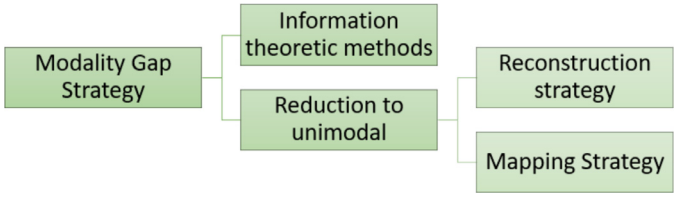

Fig. 6. Modality Gap Strategies.

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-

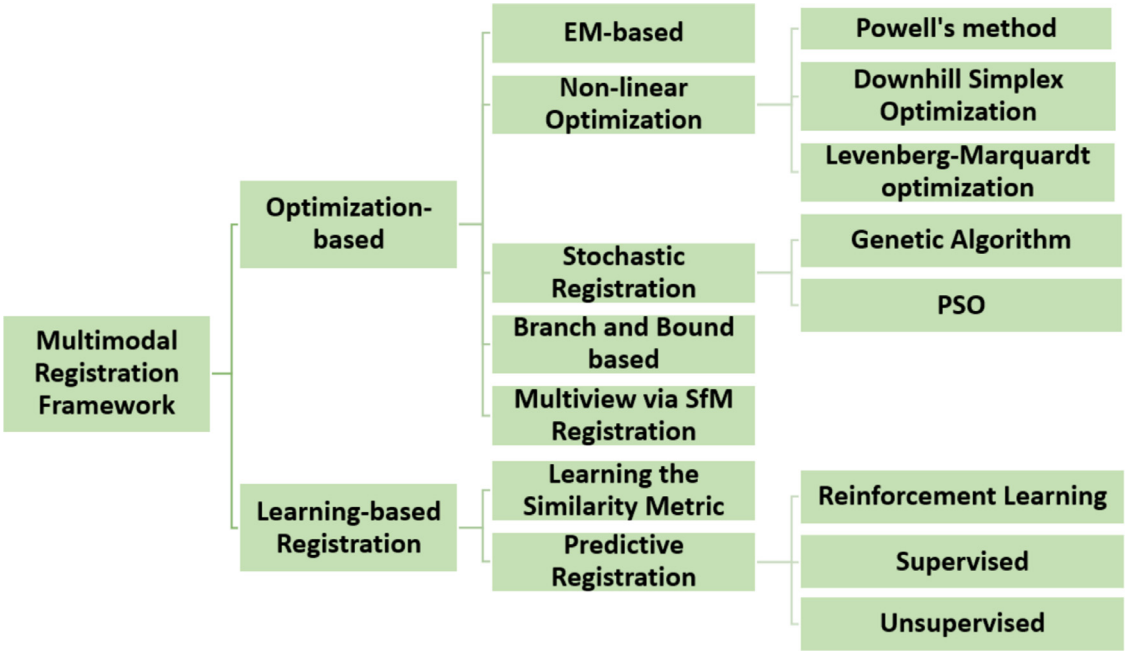

Fig. 7. A classification of presented multimodal registration strategies.

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]. [140,141]	3D volume	2D x-rays	rigid	global		mapping one modality to another	Multiview with SFM	[139]	medical	mTRE SR	0.22mm 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. [148,149]	3D model	2D image	rigid	global	features: points, lines	mapping one modality to another	BnB	[146,147]	general	SR	25%	500sec-1000sec
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	5.65mm	n/a	n/a
									FR	11.20%		
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	6.3 mm	n/a	0.25sec
									Dice	0.82		
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	4.2 mm	n/a	0.25sec
									Dice	0.88		
Hu et al [199].	3D MRI	3D TRUS	non rigid	global	PTR-Supervised	CNN	SmartTarget [112]	medical	TRE	4.8 mm	n/a	0.25se
									Dice	0.82		
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	1.23mm ± 0.43	40hours	15sec
									Dice	0.905		
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	5.67mm	n/a	2.50sec
									FR	2.7%		
Fan et al. [204]	3D MRI	3D CT	rigid	global	PTR-UnSupervised	GAN	private	medical	mTRE	1.57mm ± 0.44		
									Dice	0.86		n/a
DSAC [205]	3D scene	2D image	rigid	global	PTR-Reinforcement learning	CNN	7-Scenes [86]	general	mTRE	4.1cm, 1.1°	n/a	0.1sec
									SR	58.5%		
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	0.567 ± 0.038		
									MI	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-

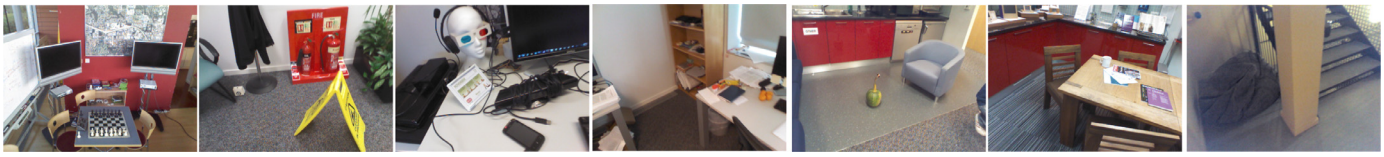

Fig. 8. 7-Scenes dataset sample images from left to right: Chess, Fire, Heads, Office, Pumpkin, Red Kitchen and Stairs.

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 Extent (m)	# Frames Train	# Frames 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.

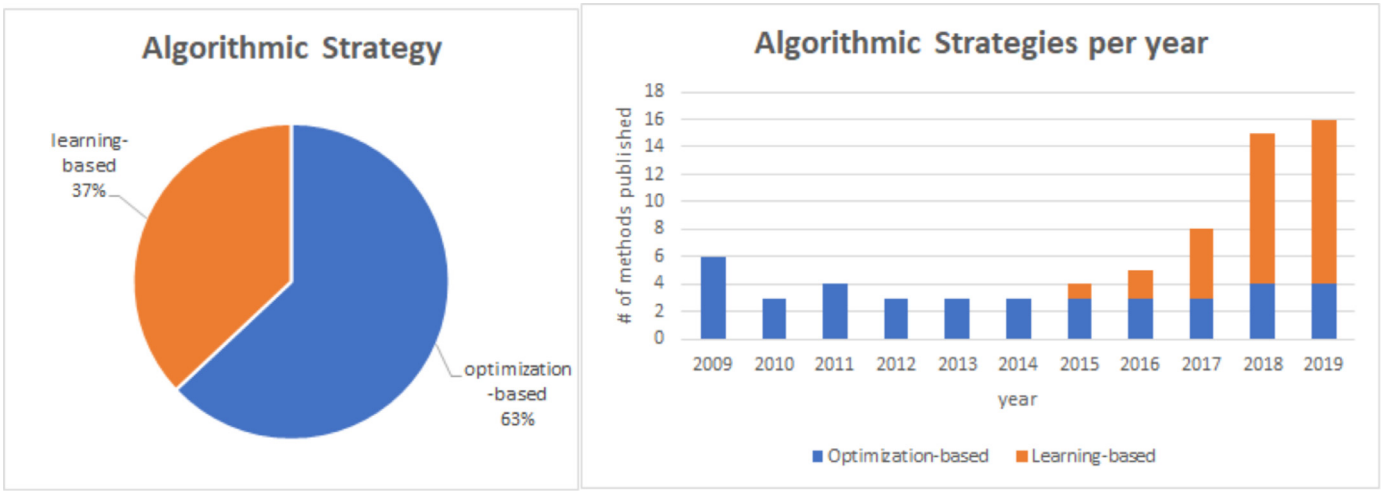

Fig. 9. Overview of the number of publications in multimodal 3D registration based on their algorithmic strategy .

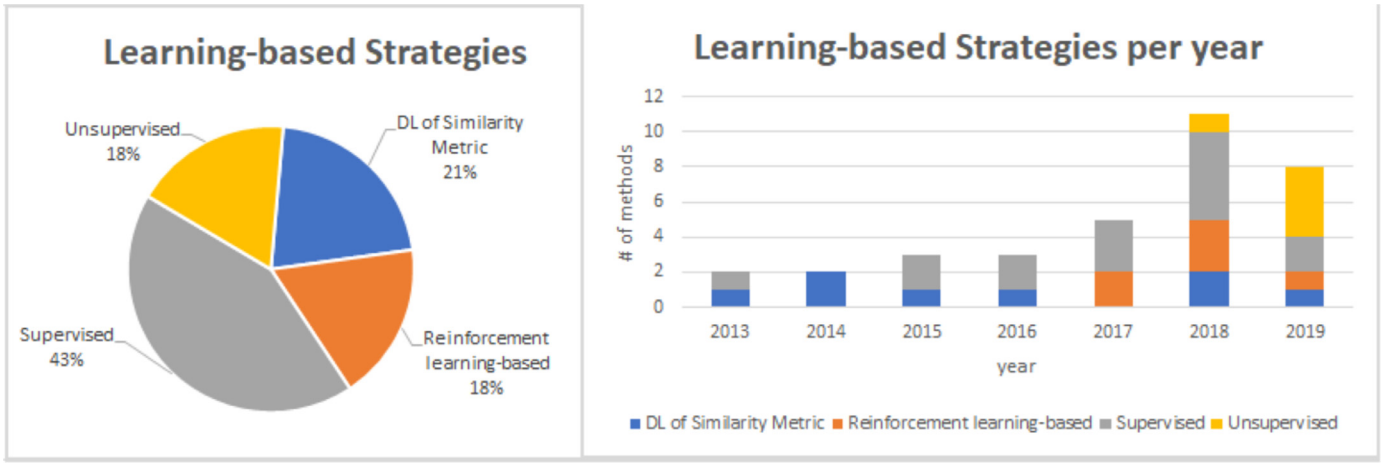

Fig. 10. Overview of the number of proposed learning-based methods for multimodal 3D registration .

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,

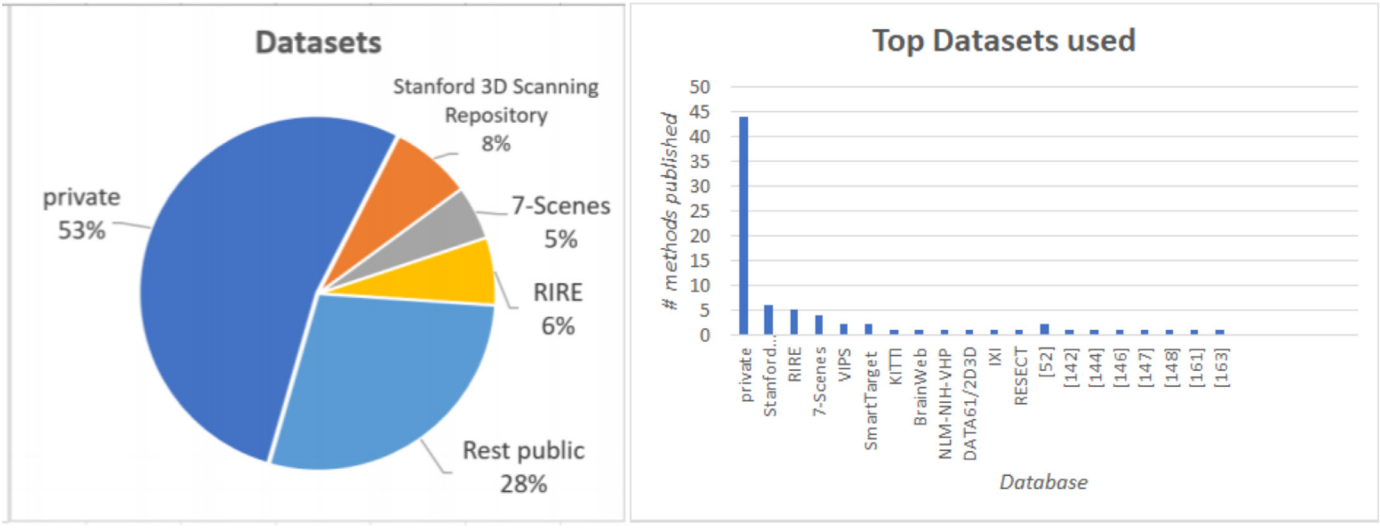

Fig. 11. Overview of the datasets used to implement/test the presented techniques.

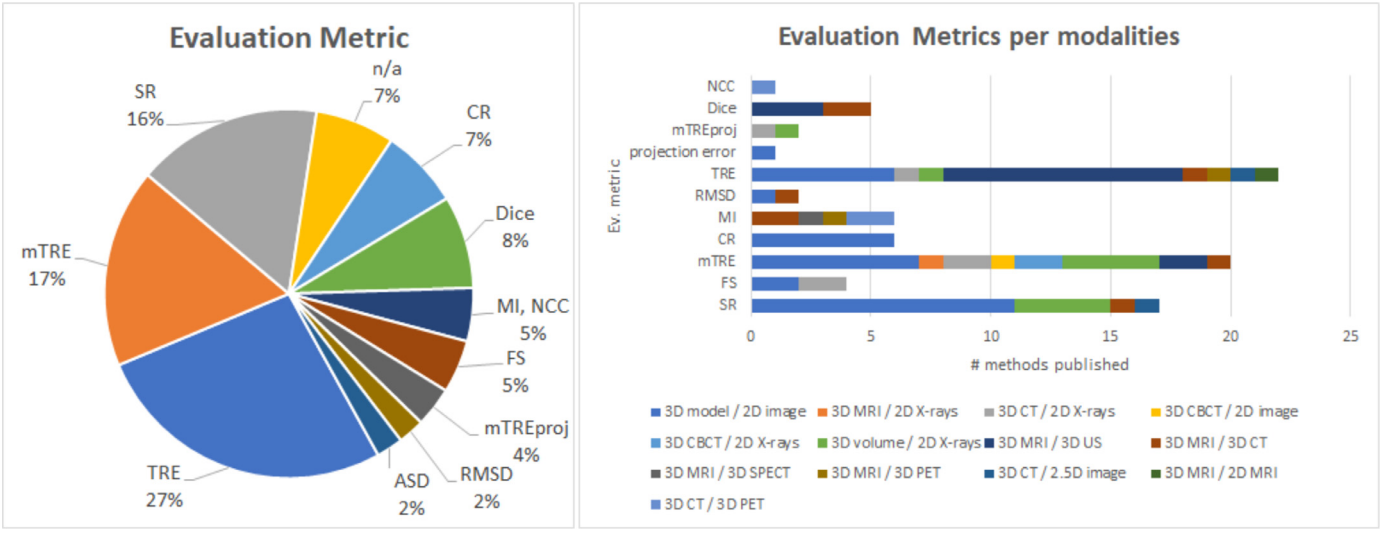

Fig. 12. Overview of evaluation measures used in the presented multimodal 3D registration methods.

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

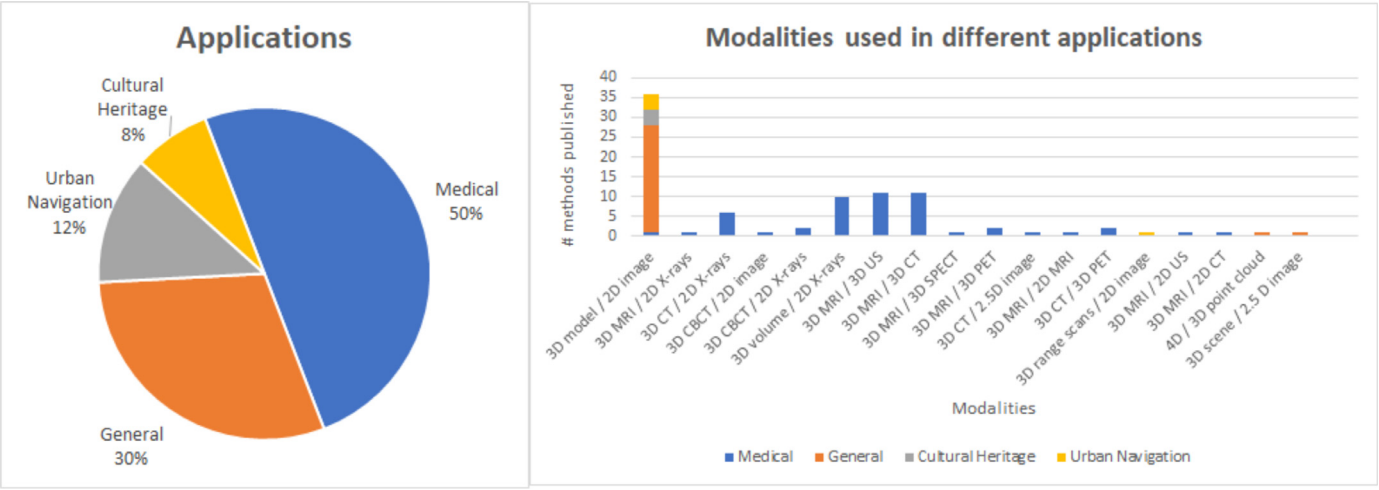

Fig. 13. Pie Charts of applications and modalities registered per application.

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] Klauduiny 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.