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

**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.

© 2020 The Authors. Published by Elsevier Ltd.

This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

The 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

© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)
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)
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.

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.
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 (DOF), 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 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 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** [88]: 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 \( T_{\text{ground}} \) and the estimated registration \( T_{\text{reg}} \). 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.

\[
\text{TRE} = \| T_{\text{reg}}(P) - T_{\text{ground}}(P) \| \quad (1)
\]

- **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.

\[
\text{mTRE} = \frac{1}{N} \sum_{i=1}^{N} \| T_{\text{reg}}(P_i) - T_{\text{ground}}(P_i) \| \quad (2)
\]
• **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 $\hat{n}$.

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

(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.

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

(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 overlap between the two datasets, to 1, indicating complete overlap and thus a successful registration. Given two datasets $X$ and $Y$ to be registered, the DSC is defined as in Eq. 3, where $|X|$ and $|Y|$ refer to the cardinalities of the respective datasets [117].

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

(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.

### 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-
ment transformation) given the correspondences. These two sub-problems 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.

![Multimodal Registration Framework](image)

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

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 Ol (Orthogonal Iteration) [171], is employed. The SoftOI algorithm first introduces an assignment matrix that describes the correspondences for the Ol 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality A</th>
<th>Modality B</th>
<th>Nature of Correspondence</th>
<th>Domain of Transform.</th>
<th>Type Of Correspondence</th>
<th>Modality Gap Strategy</th>
<th>Optimization-based Dataset Strategy</th>
<th>Initial Application</th>
<th>Evaluation Measure</th>
<th>Value of Eval.Measure</th>
<th>Execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parmehr et al [77].</td>
<td>3D model (LIDAR point cloud)</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity: NCMi</td>
<td>mapping one modality to another</td>
<td>intensity-based distance optimization</td>
<td>private</td>
<td>urban navigation</td>
<td>0.12m - 0.0051°</td>
<td>n/a</td>
</tr>
<tr>
<td>Sottile et al [78].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity: MC</td>
<td>mapping one modality to another</td>
<td>distance optimization intensity-based</td>
<td>private</td>
<td>general</td>
<td>4.8 pixels</td>
<td>2sec</td>
</tr>
<tr>
<td>Wachowiak et al [121].</td>
<td>3D volume MRI</td>
<td>2D US</td>
<td>rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>distance optimization Stochastic/ PSO</td>
<td>BrainWeb [98,122], NLM-NIH VHP [101]</td>
<td>medical</td>
<td>TRE</td>
<td>2.36mm</td>
</tr>
<tr>
<td>Wachowiak et al [121].</td>
<td>3D volume MRI</td>
<td>2D CT</td>
<td>rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>Stochastic / PSO</td>
<td>BrainWeb [98,122], NLM-NIH VHP [101]</td>
<td>medical</td>
<td>TRE</td>
<td>2.14</td>
</tr>
<tr>
<td>Schwab et al [116].</td>
<td>3D volume MRI</td>
<td>3D volume CT</td>
<td>rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / PSO</td>
<td>medical</td>
<td>TRE</td>
<td>9.57mm - 78%</td>
<td>n/a</td>
</tr>
<tr>
<td>Chen et al [123,124].</td>
<td>3D volume MRI</td>
<td>3D volume CT</td>
<td>rigid, non rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / PSO</td>
<td>medical</td>
<td>TRE</td>
<td>2.36mm</td>
<td>1893.637sec</td>
</tr>
<tr>
<td>Lin et al [125].</td>
<td>3D volume MRI</td>
<td>3D volume CT</td>
<td>rigid, non rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / PSO</td>
<td>medical</td>
<td>TRE</td>
<td>14.18mm - 1.55°</td>
<td>40sec - 200sec</td>
</tr>
<tr>
<td>Liu et al [126].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>features: points</td>
<td>mapping one modality to another reconstruction</td>
<td>Multiview with SFM [127]</td>
<td>cultural</td>
<td>TRE</td>
<td>3.19cm - 0.26°</td>
<td>1140sec - 24960sec</td>
</tr>
<tr>
<td>Corsini et al [120].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>features: points</td>
<td>modality strategy reconstruction</td>
<td>Multiview with SFM [128,129]</td>
<td>heritage cultural</td>
<td>TRE</td>
<td>3.19mm</td>
<td>1140sec - 24960sec</td>
</tr>
<tr>
<td>Pintus and Gobbetti [130]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>features: points</td>
<td>modality strategy reconstruction</td>
<td>Multiview with SFM [128,129]</td>
<td>heritage cultural</td>
<td>TRE</td>
<td>3.19mm</td>
<td>1140sec - 24960sec</td>
</tr>
<tr>
<td>Klima et al [131].</td>
<td>3D volume CT</td>
<td>2D x-rays</td>
<td>non rigid</td>
<td>local</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>NL / LM method</td>
<td>heritage medical</td>
<td>mTRE</td>
<td>1.23mm</td>
<td>15.77sec</td>
</tr>
<tr>
<td>DePose [132]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>Stochastic / GA</td>
<td>private</td>
<td>mTRE</td>
<td>0.6m - 1.0°</td>
<td>1.25sec - 1.99sec</td>
</tr>
<tr>
<td>EvoPose [133]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>Stochastic / GA</td>
<td>private</td>
<td>mTRE</td>
<td>1.28 cm - 2.2°</td>
<td>0.68sec - 4.11sec</td>
</tr>
<tr>
<td>Crombez et al [134].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>intensity: MI</td>
<td>mapping one modality to another</td>
<td>Stochastic / PSO</td>
<td>private</td>
<td>mTRE</td>
<td>5.65cm - 0.61°</td>
<td>n/a</td>
</tr>
<tr>
<td>Toth et al [135].</td>
<td>3D volume MRI</td>
<td>2D x-rays</td>
<td>rigid</td>
<td>global</td>
<td></td>
<td>mapping one modality to another</td>
<td>BnB [52]</td>
<td>medical</td>
<td>mTRE</td>
<td>3.87 ± 1.22mm</td>
<td>95.24sec</td>
</tr>
<tr>
<td>Wang et al [136].</td>
<td>3D volume MRI</td>
<td>2D x-rays</td>
<td>rigid</td>
<td>global</td>
<td></td>
<td>intensity-based distance optimization</td>
<td>intensity-based distance optimization</td>
<td>medical</td>
<td>mTRE</td>
<td>0.17mm</td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
Table 2 (continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Schaffert et al [138]. [140,141]</td>
<td>3D volume</td>
<td>2D x-rays</td>
<td>rigid</td>
<td>global</td>
<td>mapping one modality to another</td>
<td>Multiview with SFM</td>
<td>[139] medical mTRE</td>
<td>SR</td>
<td>98.4%</td>
<td>0.22mm</td>
<td>7.0sec-35.0sec 36sec</td>
</tr>
<tr>
<td>SoftPosit [142]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: points, lines feature: lines</td>
<td>EM-based private</td>
<td>general SR</td>
<td>SR</td>
<td>75%</td>
<td></td>
<td>72sec-100sec</td>
</tr>
<tr>
<td>David et al [143].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity: joint entropy</td>
<td>EM-based private</td>
<td>general SR</td>
<td>SR</td>
<td>70%</td>
<td></td>
<td>6.50sec-15.0sec</td>
</tr>
<tr>
<td>Mastin et al.</td>
<td>3D model</td>
<td>(LIDAR point cloud) 2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity: WNMI</td>
<td>NL / Downhill Simplex</td>
<td>private SR</td>
<td>urban</td>
<td>98.5%</td>
<td></td>
<td>1500sec</td>
</tr>
<tr>
<td>Parmehr et al [76].</td>
<td>(LIDAR point cloud)</td>
<td>(aerial photograph) 2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity-based</td>
<td>distance optimization BnB</td>
<td>general SR</td>
<td>navigation</td>
<td>96%</td>
<td></td>
<td>2sec-4sec</td>
</tr>
<tr>
<td>Enqvist et al [144].</td>
<td>(LIDAR point cloud)</td>
<td>(aerial photograph) 2D image</td>
<td>rigid</td>
<td>global</td>
<td>features: points</td>
<td>BnB</td>
<td>[145] general SR</td>
<td>SR</td>
<td>96%</td>
<td></td>
<td>2sec-4sec</td>
</tr>
<tr>
<td>Brown et al.</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>features: points</td>
<td>BnB</td>
<td>[146,147] general SR</td>
<td>SR</td>
<td>25%</td>
<td></td>
<td>500sec-1000sec</td>
</tr>
<tr>
<td>[148,149]</td>
<td>GOPAC [150]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>BnB</td>
<td>DATA61/2D3D general TRE</td>
<td>2.30m - 2.08°</td>
<td>477sec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BlindPnP [119]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: points</td>
<td>EM-based [83] private</td>
<td>general CR</td>
<td>SR</td>
<td>82%</td>
<td>65%</td>
<td>20sec-100sec</td>
</tr>
<tr>
<td>Sanchez et al [151].</td>
<td>3D model</td>
<td>2D image</td>
<td>non rigid</td>
<td>local</td>
<td>feature: points</td>
<td>EM-based private</td>
<td>general CR</td>
<td>CR</td>
<td>90%</td>
<td></td>
<td>600sec-1500sec</td>
</tr>
<tr>
<td>SoftOI [152]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: points</td>
<td>EM-based private</td>
<td>general CR</td>
<td>CR</td>
<td>75%</td>
<td></td>
<td>10sec-60sec</td>
</tr>
<tr>
<td>Corsini et al [153].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity:MI</td>
<td>NL / Powell's method</td>
<td>private cultural SP</td>
<td>CR</td>
<td>80%</td>
<td></td>
<td>4sec</td>
</tr>
<tr>
<td>Palma et al [154].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity:MI</td>
<td>NL / Powell's method</td>
<td>private heritage cultural SP</td>
<td>CR</td>
<td>70%</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>Yang et al [155].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td></td>
<td>common space</td>
<td>Stochastic / GA private</td>
<td>general FS</td>
<td>25%</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>Marques et al.</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>feature: points</td>
<td></td>
<td>common space</td>
<td>NL / Linear Regression private</td>
<td>general FS</td>
<td>20%</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>Enqvist et al [156].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>features: points</td>
<td>BnB</td>
<td>[157] general FS</td>
<td>general FS</td>
<td>20%</td>
<td></td>
<td>5sec-15sec</td>
</tr>
</tbody>
</table>
### Table 2 (continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kisaki et al [158].</td>
<td>3D volume CT</td>
<td>3D volume MRI</td>
<td>rigid</td>
<td>local</td>
<td>intensity: NMI</td>
<td>mapping one modality to another</td>
<td>NL / LM method</td>
<td>private</td>
<td>medical</td>
<td>MI</td>
<td>0.294</td>
</tr>
<tr>
<td>Talbi et al [159].</td>
<td>3D volume MRI</td>
<td>3D volume CT</td>
<td>rigid</td>
<td>global</td>
<td>similarity measure</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / HPSO</td>
<td>private</td>
<td>medical</td>
<td>MI</td>
<td>0.6349</td>
</tr>
<tr>
<td>Talbi et al [159].</td>
<td>3D volume MRI</td>
<td>3D volume SPECT</td>
<td>rigid</td>
<td>global</td>
<td>similarity measure</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / HPSO</td>
<td>private</td>
<td>medical</td>
<td>MI</td>
<td>0.6789</td>
</tr>
<tr>
<td>Talbi et al [159].</td>
<td>3D volume MRI</td>
<td>3D volume PET</td>
<td>rigid</td>
<td>global</td>
<td>similarity measure</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / HPSO</td>
<td>private</td>
<td>medical</td>
<td>MI</td>
<td>0.6431</td>
</tr>
<tr>
<td>Khoo and Kapoor [160]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>global</td>
<td>mapping one modality to another</td>
<td>NL / Conv [37], private medical datasets</td>
<td>private</td>
<td>medical</td>
<td>RMSD</td>
<td>6.9mm</td>
<td>n/a</td>
</tr>
<tr>
<td>Ayatollahi et al [161].</td>
<td>3D volume MRI</td>
<td>3D volume CT</td>
<td>rigid</td>
<td>global</td>
<td>intensity: MNMI</td>
<td>learning multimodal similarity measure</td>
<td>Stochastic / HPSO</td>
<td>private</td>
<td>medical</td>
<td>RMSD</td>
<td>44%</td>
</tr>
<tr>
<td>Zhao et al [75].</td>
<td>3D range scans (aerial photograph)</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>intensity: CMI</td>
<td>mapping one modality to another</td>
<td>distance optimization</td>
<td>KITTI [80]</td>
<td>urban</td>
<td>projection</td>
<td>error</td>
</tr>
<tr>
<td>RANSAC [67]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: points</td>
<td>mapping one modality to another</td>
<td>Stochastic</td>
<td>private</td>
<td>general</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Beveridge et al [163].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: lines</td>
<td>mapping one modality to another</td>
<td>EM-based</td>
<td>private</td>
<td>urban</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>David et al [164].</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: lines</td>
<td>mapping one modality to another</td>
<td>EM-based</td>
<td>private</td>
<td>navigation</td>
<td>general</td>
<td>n/a</td>
</tr>
<tr>
<td>SoftSI [165]</td>
<td>3D model</td>
<td>2D image</td>
<td>rigid</td>
<td>local</td>
<td>feature: points</td>
<td>mapping one modality to another</td>
<td>EM-based</td>
<td>private</td>
<td>general</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Pan et al [166].</td>
<td>3D x-rays</td>
<td>2D x-rays</td>
<td>rigid</td>
<td>global</td>
<td>mapping one modality to another</td>
<td>BnB</td>
<td>private</td>
<td>medical</td>
<td>n/a</td>
<td>n/a</td>
<td>4.12sec-12.09sec</td>
</tr>
<tr>
<td>Zhao et al [167].</td>
<td>4D video</td>
<td>3D point cloud</td>
<td>local</td>
<td>features: points</td>
<td>reconstruction</td>
<td>Multiview with SFM</td>
<td>private</td>
<td>general</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>
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 Levenberg-Marquardt 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. Kikaki 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 Fischer 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/best/sI". 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 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 the 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 swarm optimization (HPSO) method incorporates sub-population and crossover from GAs into the conventional PSO. The particles are not standalone, but are divided into a number of sub-populations. 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 space 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 inner 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 inner 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>
<thead>
<tr>
<th>Table 3</th>
<th>Overview of Learning-based Registration Methods, grouped by evaluation measure and dataset used.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning-based Registration Methods</td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Modality A</td>
</tr>
<tr>
<td>Haskins et al. [189]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Zheng et al. [190]</td>
<td>3DCT</td>
</tr>
<tr>
<td>Ma Kai et al. [191]</td>
<td>3D CT</td>
</tr>
<tr>
<td>Miao et al. [192]</td>
<td>3D volume</td>
</tr>
<tr>
<td>Hu et al [193]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Yan et al. [194]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Salehi et al. [195]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Sedghi et al. [196]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Lee et al. [197]</td>
<td>3D CT</td>
</tr>
<tr>
<td>Lee et al. [198]</td>
<td>3D PET</td>
</tr>
<tr>
<td>Hu et al [199]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Hu et al [199]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Chou et al. [200]</td>
<td>3D CBCT</td>
</tr>
<tr>
<td>Wright et al. [201]</td>
<td>3D CT</td>
</tr>
<tr>
<td>Cao et al. [202]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Pei et al. [203]</td>
<td>3D CBCT</td>
</tr>
<tr>
<td>POINT [118]</td>
<td>3D CT/CBCT</td>
</tr>
<tr>
<td>Fan et al. [204]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>DSAC [205]</td>
<td>3D scene</td>
</tr>
<tr>
<td>PoseNet [206]</td>
<td>3D scene</td>
</tr>
<tr>
<td>Moehl et al. [207]</td>
<td>3D scene</td>
</tr>
<tr>
<td>Kendall et al. [91]</td>
<td>3D scene</td>
</tr>
<tr>
<td>Sun et al. [208]</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Shotton et al. [86]</td>
<td>3D scene</td>
</tr>
<tr>
<td>Miao et al. [209]</td>
<td>3D model</td>
</tr>
<tr>
<td>Miao et al. [47]</td>
<td>3D CT</td>
</tr>
<tr>
<td>Yu et al. [210]</td>
<td>3D CT</td>
</tr>
<tr>
<td>DenseRegNet [211]</td>
<td>3D CT</td>
</tr>
</tbody>
</table>
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 wrapper, 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 reward 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 RANSAC 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 diluted 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 Melkhov 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-
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 2D 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

<table>
<thead>
<tr>
<th>Scene</th>
<th>Spatial Extent (m)</th>
<th># Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chess</td>
<td>$3 \times 2 \times 1$m</td>
<td>4000</td>
</tr>
<tr>
<td>Fire</td>
<td>$2.5 \times 1 \times 1$m</td>
<td>2000</td>
</tr>
<tr>
<td>Heads</td>
<td>$2 \times 0.5 \times 1$m</td>
<td>1000</td>
</tr>
<tr>
<td>Office</td>
<td>$2.5 \times 2 \times 1.5m$</td>
<td>6000</td>
</tr>
<tr>
<td>Pumpkin</td>
<td>$2.5 \times 2 \times 1$m</td>
<td>4000</td>
</tr>
<tr>
<td>Red Kitchen</td>
<td>$4 \times 3 \times 1.5$m</td>
<td>7000</td>
</tr>
<tr>
<td>Stairs</td>
<td>$2.5 \times 2 \times 1.5$m</td>
<td>2000</td>
</tr>
</tbody>
</table>

Shotton et al. in [86] also propose a method for aligning a 3D scene with a 2D 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 [211] 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 m TRE (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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chess</td>
<td>0.042m, 1.4°</td>
<td>0.022m, 1°</td>
<td>0.32m, 4.06°</td>
<td>0.13m, 4.48°</td>
<td>0.042m, 1.1°</td>
<td>9.43m, 1.1°</td>
<td>0.95m, 0.02°</td>
</tr>
<tr>
<td>Fire</td>
<td>0.371m, 2.1°</td>
<td>0.051m, 2.4°</td>
<td>0.47m, 7.33°</td>
<td>0.27m, 11.3°</td>
<td>0.067m, 3.1°</td>
<td>2.46m, 1.57°</td>
<td>0.72m, 1.09°</td>
</tr>
<tr>
<td>Heads</td>
<td>0.098m, 3.1°</td>
<td>0.125m, 5.1°</td>
<td>0.29m, 6.00°</td>
<td>0.17m, 13.0°</td>
<td>0.125m, 4.1°</td>
<td>5.85m, 1.72°</td>
<td>0.90m, 4.71°</td>
</tr>
<tr>
<td>Office</td>
<td>0.089m, 1.6°</td>
<td>0.046m, 1.4°</td>
<td>0.48m, 3.84°</td>
<td>0.19m, 5.55°</td>
<td>0.098m, 2.7°</td>
<td>4.26m, 1.26°</td>
<td>1.17m, 1.47°</td>
</tr>
<tr>
<td>Pumpkin</td>
<td>0.045m, 1.7°</td>
<td>0.065m, 3.7°</td>
<td>0.47m, 4.21°</td>
<td>0.26m, 4.75°</td>
<td>0.040m, 1.5°</td>
<td>9.94m, 1.35°</td>
<td>1.14m, 1.29°</td>
</tr>
<tr>
<td>Red Kitchen</td>
<td>0.087m, 2.4°</td>
<td>0.072m, 2.1°</td>
<td>0.59m, 4.32°</td>
<td>0.23m, 5.35°</td>
<td>0.078m, 2.6°</td>
<td>20.7m, 1.29°</td>
<td>0.64m, 1.18°</td>
</tr>
<tr>
<td>Stairs</td>
<td>0.05m, 3.2°</td>
<td>0.149m, 2.6°</td>
<td>0.47m, 6.93°</td>
<td>0.35m, 12.4°</td>
<td>0.493m, 3.1°</td>
<td>9.02m, 1.53°</td>
<td>1.00m, 1.48°</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chess</td>
<td>0.03m</td>
<td>0.032m</td>
<td>0.45m</td>
<td>0.24m</td>
<td>0.04m</td>
<td>6.68m</td>
<td>2.94m</td>
</tr>
<tr>
<td>Fire</td>
<td>0.4m</td>
<td>0.045m</td>
<td>0.34m</td>
<td>0.45m</td>
<td>0.07m</td>
<td>4.26m</td>
<td>1.07m</td>
</tr>
<tr>
<td>Heads</td>
<td>0.12m</td>
<td>0.210m</td>
<td>0.52m</td>
<td>0.29m</td>
<td>0.14m</td>
<td>4.60m</td>
<td>1.09m</td>
</tr>
<tr>
<td>Office</td>
<td>0.07m</td>
<td>0.121m</td>
<td>0.67m</td>
<td>0.17m</td>
<td>0.19m</td>
<td>3.99m</td>
<td>3.56m</td>
</tr>
<tr>
<td>Pumpkin</td>
<td>0.03m</td>
<td>0.256m</td>
<td>0.49m</td>
<td>0.36m</td>
<td>0.03m</td>
<td>9.80</td>
<td>3.29m</td>
</tr>
<tr>
<td>Red Kitchen</td>
<td>0.09m</td>
<td>0.06m</td>
<td>0.61m</td>
<td>0.25m</td>
<td>0.06m</td>
<td>20.96m</td>
<td>5.55m</td>
</tr>
<tr>
<td>Stairs</td>
<td>0.75m</td>
<td>0.161m</td>
<td>0.58m</td>
<td>0.46m</td>
<td>0.04m</td>
<td>10.58m</td>
<td>3.17m</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Scene</th>
<th>SR of Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANSAC [67]</td>
<td>Shotton et al [86].</td>
</tr>
<tr>
<td>Chess</td>
<td>96.8%</td>
</tr>
<tr>
<td>Fire</td>
<td>71.8%</td>
</tr>
<tr>
<td>Heads</td>
<td>66.7%</td>
</tr>
<tr>
<td>Office</td>
<td>57.6%</td>
</tr>
<tr>
<td>Pumpkin</td>
<td>59.0%</td>
</tr>
<tr>
<td>Red Kitchen</td>
<td>40.1%</td>
</tr>
<tr>
<td>Stairs</td>
<td>12.8%</td>
</tr>
</tbody>
</table>

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 changes 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.
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,
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. 11. Overview of the datasets used to implement/test the presented techniques.](image1)

![Fig. 12. Overview of evaluation measures used in the presented multimodal 3D registration methods.](image2)
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, IIX [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 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


Fig. 13. Pie Charts of applications and modalities registered per application.
Yoshimura 2004.

Bhatnagar et al. 2015a;38(11):2241–54


Middell A, Scheler J. Visualization of large and unstructured data sets—applications in geospatial planning, modeling and engineering; 2011.


Enqvist. Gold and Powell. 3D.


Johnson K., Becker J. The whole brain atlas. [http://www.med.harvard.edu/brainatlas].


