ABSTRACT

Image registration is a common task for many biomedical analysis applications. The present work focuses on the benchmarking of registration methods on differently stained histological slides. This is a challenging task due to the differences in the appearance model, the repetitive texture of the details and the large image size, between other issues. Our benchmarking data is composed of 616 image pairs at two different scales — average image diagonal 2.4k and 5k pixels. We compare eleven fully automatic registration methods covering the widely used similarity measures (and optimization strategies with both linear and elastic transformation). For each method, the best parameter configuration is found and subsequently applied to all the image pairs. The performance of the algorithms is evaluated from several perspectives — the registrations (in)accuracy on manually annotated landmarks, the method robustness and its processing computation time.

Index Terms— Image registration, benchmarking, pathology, histological slides, stained tissue.

1. INTRODUCTION

Computer-assisted diagnosis algorithms are currently developed for disease detection, diagnosis and prognosis prediction on digitized tissue histopathology images [1]. Commonly, consecutive tissue sections are differently stained to extract complementary information. The registration (i.e., alignment) of those sections can help to infer the correspondences between the spatial distributions of different markers (e.g., gene expression signature [2]) on the tissue. Image registration is a common task in many biomedical image analysis applications [3]. A few registration benchmarks have already been performed focusing on intra-modality clinical imaging data [4,8] but as per our knowledge, there has not been a previous initiative tailored to inter-modality (i.e., with a clear appearance difference) data.

In the last years, a number of methods have been specifically developed for the registration of histological sections [11-20]. This is a challenging task due to the following aspects: (a) the vast size of a digitized tissue section (e.g. 50k × 50k pixels) [9,10], (b) the repetitive nature of the texture and observed patterns [11], (c) the elastic deformations, occlusions and missing parts, caused by sample preparation [9,11].

The first attempts to register histological sections deal with same-stain images (typically H&E stain) [12-15]. More recently, registration methods were extended to deal with the alignment of differently stained histological sections [9,11,16-18]. This is a multi-modal registration since the change in appearance among the differently stained images is large (see Fig. 1). So far, a wider experimental comparison between methods on the same dataset with a known ground truth is missing.

The aim of this work is to fill this gap and present our preliminary results concerning such a benchmark. In the previous registration benchmarks, the simplest evaluation metric used was the similarity measure value (indicative of algorithm convergence) [6]. Another widely used evaluation metric relies on the segmentation performed by a human operator and measures the overlap between the annotations of the corresponding slides (e.g. by the Dice, kappa, or Jaccard index) [5,8]. In our cases, since an expert segmentation is not available, we perform a landmark-based evaluation [7]. Namely, we measure the (in)accuracy on the transformed positions of a sufficiently large set of landmarks (about a hundred) manually set by an expert, after registration.

The structure of this work is as follows: in Sec. 2 we present the main aspects of the evaluated image registration methods; in Sec. 3 we describe the datasets and present the performance metrics; and finally, in Sec. 4 we discuss the experimental results. Our dataset containing the images together with the landmark positions and the evaluation source code is freely available.

This work was partially supported by the Czech Science Foundation project 17-15361S, a Research Center for Informatics project CZ.02.1.01/0.0/0.0/16_019/0000765, and by the Spanish Ministry of Economy and Competitiveness (TEC2013-48552-C2-1-R, TEC2015-73064-EXP and TEC2016-78052-R).
2. REGISTRATION METHODS

Image registration involves spatially transforming the source (reference) images to align with the target image. The main components of image registration algorithms are: the geometrical transformation (linear or non-linear), the similarity measure to be maximized, and the optimization technique. We have classified the algorithms with respect to the similarity measure as those in which: (i) the criterion is formulated in the original image domain (see Sec. 2.1), (ii) the criterion measures spatial relationships between previously extracted features (see Sec. 2.2), and (iii) the criterion is formulated in the segmentation domain (see Sec. 2.4). In addition, there are methods that use both features and similarity measures (see Sec. 2.3). We chose methods for which an open source implementation is available, see Table 1.

2.1. Intensity-based registration methods

Intensity-based methods compare intensity patterns in images using a given measure [13, 19, 20]: sum of squared differences (SSD), cross-correlation (CC), Mattes mutual Information (MMI) and normalized mutual Information (NMI). Commonly, the optimal transformation is found by a local continuous descent optimization, for example the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS). These methods may suffer from local minima such as when repetitive texture patterns are presented (see Fig. 1). We chose the following methods: Elastix [20] and Advanced Normalization Tools (ANTS) [19] modules which are both based on the Insight Segmentation and Registration Toolkit (ITK) [21]; NiftyReg [22] which uses a block-matching technique to find a coarse preliminary transformation which is than followed by a non-linear refinement using the free-form (B-spline) deformation scheme; and bUnwarpJ [13] which calculates a non-linear invertible deformations represented by B-splines (note that landmarks are not used).

2.2. Feature-based registration methods

The critical task of these methods is the extraction of descriptive features robust to the difference in appearance between stains. Commonly, the well-known scale-invariant feature transform (SIFT), speeded up robust features (SURF), or maximally stable extremal regions (MSER) are used. Then, the optimal transformation is estimated e.g. by a random sample and consensus (RANSAC) [23]. We choose Feature-based registration (openCV) and TrakEM2 [23] as representatives of this class.

2.3. Feature and intensity-based registration methods

The purpose of this class of methods is to address the weaknesses of the simple methods and increase their convergence speed and robustness. A common approach is to use the feature-based method for pre-registration to estimate an initial alignment [11] or combine both approaches in a single global criterion [13]. As representatives, we chose Feature-based + Elastix [11], DROP [24] which uses discrete optimization, and Register Virtual Stack Slices (RVSS) [13].

2.4. Segmentation based methods

Our registration methods ASSAR [17] and SegReg [18] are based on the registration of the segmented images, which leads to fast and robust approaches for multimodal registration. In ASSAR, the segmentation is performed using softmax regression and the registration is based on a triangular grid and discrete optimization using loopy belief propagation (BP) [17]. In SegReg [18], normal sampling is used for further speed-up.

3. MATERIALS AND EXPERIMENTAL SETTINGS

3.1. Dataset – images & landmarks

Consecutive tissue slices were stained with several different stains [2,9], e.g., Clara cell 10 protein (CC10), prosurfac-
Table 1. Main components of the representative registration methods used for our comparison. Note: Lagrangian Push-Forward (LPF), Levenberg-Marquardt (LM), Limited-memory BFGS (L-BFGS). For more acronyms see Sec. 2.

<table>
<thead>
<tr>
<th>Method</th>
<th>Criterion</th>
<th>Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenCV</td>
<td>SURF &amp; MSER</td>
<td>RANSAC</td>
</tr>
<tr>
<td>TrakEM2 [23]</td>
<td>SIFT</td>
<td>RANSAC</td>
</tr>
<tr>
<td>Elastix [20]</td>
<td>MMI</td>
<td>L-BFGS</td>
</tr>
<tr>
<td>ANTs [19]</td>
<td>MMI / CC</td>
<td>LPF</td>
</tr>
<tr>
<td>NiftyReg [22]</td>
<td>NMI</td>
<td>conj. gradient</td>
</tr>
<tr>
<td>OpenCV + Elastix</td>
<td>SURF &amp; MSER</td>
<td>RANSAC</td>
</tr>
<tr>
<td>RVSS [13]</td>
<td>SIFT + SSD</td>
<td>LM</td>
</tr>
<tr>
<td>ASSAR [17]</td>
<td>MIL</td>
<td>BP</td>
</tr>
<tr>
<td>SegReg [18]</td>
<td>MIL</td>
<td>BOBYQUA</td>
</tr>
</tbody>
</table>

tant protein C (proSPC), hematoxylin and eosin (H&E), antigen KI-67 (Ki67), platelet endothelial cell adhesion molecule (PECAM-1, also known as CD31), see Fig. 1. High-resolution (40x magnification) whole-slide of a tissue (e.g. breast and lung tumor, rat kidney, etc.) images were acquired - the original size of our images varies up to $45k \times 45k$ pixels [10] which is common in medical pathology. The acquired image are organized in 32 sets of consecutive sections where each slice was stained by a different dye. The number of used stains in a particular set varies between 4 and 9.

**Landmark annotation.** Significant structures in the tissue of consecutive stains were marked using uniformly spread landmarks. We obtained 616 image pairs to be registered. The annotation of a single image pair takes about 10 min.

**Dataset scaling.** For the purpose of characterizing the impact of the image size on the methods performance, we scaled the datasets to two sizes - $2k \times 2k$ (small (S)) and $4k \times 4k$ pixels (medium (M)).

### 3.2. Evaluation

The performance of the registration is evaluated using three metrics: registration inaccuracy, robustness and execution time. We have chosen not to give a global ranking of the methods as the relevance of each metric can vary depending on the application.

**Inaccuracy** is measured as the sum of the Euclidean distances in pixels between the positions of the landmarks in the reference and the transformed images. The inaccuracy is normalized by the size of the image diagonal. The inaccuracy statistics before registration are: mean=4.25%, median=2.79% and standard deviation=5.98%. The disagreement between the two experts performing the annotation is 0.4%.

**Robustness** is the percentage of cases in which the performed registration improved the initial relative inaccuracy measure. Note that only the results of the well-behaving registrations are used to compute the inaccuracy statistics.

**Execution time** is measured on a computer using a single CPU/thread. We measure only the execution time, not the time used to load and save the images. The experiments were performed on a server with an AMD K8 2000 processor and 128GB RAM.

Each method has several parameters with an impact on registration performance. The optimal parameters were set by experts (usually the method authors) using a few sample images provided by us. Note that we are using the same parameter configuration for all the image pairs, which is the desired scenario for an automatic analysis.

### 4. RESULTS

In this section, we present the experimental results of the selected image registration methods for the two image scales and the three metrics. In Table 2 and Figure 2 separate results are given for the image size (small (S) and medium (M))


Impact of the criterion type. The hybrid methods combining feature- and intensity-based criterion are robust and accurate at the expense of high processing time. The segmentation-based methods may have lower robustness if the segmentation fails, but for successful segmentation, they have high accuracy and low computational time.

Impact of image sizes. As expected, the execution time generally increases with the size of the registered images. However, the strongest effect seems to be due to the optimization algorithm (e.g., number of iterations needed). Feature-based methods are less affected by image size.

Linear vs. elastic transformation. The transformation model to use is also a common dilemma in medical practice. The free-form deformation has higher potential to capture the true deformation, although it may fall into a local minimum, and it is computationally demanding. The linear deformation for feature-based registrations turns to have higher robustness and often higher accuracy compared to elastic deformation.

5. DISCUSSION AND FUTURE WORK

We presented an experimental evaluation of publicly available image registration methods on challenging images of differently stained histological tissue. Compared to previous benchmarks we experimented with new challenging images, and utilized a more extensive dataset. The selected registration methods cover the most common similarity criteria and optimization techniques. Although the execution time of some methods is reasonably good (suitable for practical usage), the performances as measured by the proposed metrics show that the task is still not fully solved. Namely, the robustness does not approach 100% for any of the methods, and the mean accuracy is still far from that of a human annotator.

We make the datasets, the landmarks positions, and the source code of the evaluation methods publicly available. The benchmark can be easily extended with additional algorithms.

6. REFERENCES


