Confidence Estimation for Medical Image Registration Based On Stereo Confidences

Gorkem Saygili*, Marius Staring, and Emile A. Hendriks

***Abstract*—In this paper, we propose a novel method to estimate the confidence of a registration that does not require any ground truth, is independent from the registration algorithm and the resulting confidence is correlated with the amount of registration error. We first apply a local search to match patterns between the registered image pairs. Local search induces a cost space per voxel which we explore further to estimate the confidence of the registration similar to confidence estimation algorithms for stereo matching. We test our method on both synthetically generated registration errors and on real registrations with ground truth. The experimental results show that our confidence measure can estimate registration errors and it is correlated with local errors.**

***Index Terms*—Confidence estimation, medical image registration, stereo confidence.**

I. Introduction

IMAGE registration is widely used in medical image analysis to align different scans [1]–[4]. Although there are many registration algorithms, their accuracies vary between different image pairs and applications.

Quantifying errors in medical image registration is a crucial task. The quantity of errors indicate whether to trust the registration or not on a particular location. Furthermore, the parameters of a registration can be tuned adaptively on the erroneous regions to have a better alignment [5]–[8]. In general, the assessment of errors are done by the experts based on visual inspection or using residual (difference) image. However, expert-based assessment of the registration quality becomes infeasible when there are large sets of data from many subjects. Hence, it is important to develop fully-automatic confidence estimation methods for medical image registration. Such automatic methods can be used when the ground truth data is not available as well as to guide the medical expert while generating the ground truth through visual inspection.

Manuscript received July 17, 2015; revised September 01, 2015 and September 17, 2015; accepted September 19, 2015. Date of publication September 25, 2015; date of current version February 01, 2016. This research has been partially funded by the Human Brain Project from the European Union Seventh Framework Programme (FP7/2007-2013, no. 604102) and partially by the Dutch Technology Foundation STW (no. 13351). *Asterisk indicates corresponding author.*

*G. Saygili is with the Vision Lab, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands (e-mail: g.saygili@tudelft.nl).

M. Staring is with LKEB, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands (e-mail: m.staring@lumc.nl).

E. A. Hendriks is with the Vision Lab, Delft University of Technology (TU Delft), The Netherlands (e-mail: e.a.hendriks@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2015.2481609

Apart from image registration, confidence maps do also play an important role in the assessment of the so called 'disparity image' that are produced by stereo matching algorithms. Stereo matching has been an extensively researched topic of computer vision [9]–[11]. The main goal of stereo matching is to find the matching pixels between two images of the same scene from different viewpoints.

In order to improve the accuracy of stereo matching algorithms, recent research aims to extract confidence maps [12]–[14]. Without any ground truth depth measures, confidence maps are extracted by exploiting the shape of the matching cost space. An extensive overview of stereo confidences is presented in [10]. In general, full-search is computationally infeasible especially for non-rigid image registration since non-rigid medical image registration may have thousands, even millions of parameters to optimize compared to the disparity of the stereo matching. Hence, stereo confidences cannot be directly used in medical image registration. Furthermore, stereo confidences are not correlated with the amount of error in matching.

In this paper, we propose a fully-automatic confidence estimation method to measure the uncertainty in non-rigid registration. Our method relies on the fact that rigid and non-rigid deformations can be accurately modelled by local models such as translations estimated for each voxel [15] which is analogical to the non-rigid deformation field. Therefore, the quality of non-rigid registrations can be assessed by a local search for similar patterns in small neighbourhoods. As a first step, we extract local descriptors for each voxel [16], [17]. The cost space for each voxel is constructed by calculating the Euclidean distance between the extracted features of the voxels of the fixed image and their local neighbors in the moving image. We then calculate the confidence of the registration on each voxel using the proposed confidence function on the voxel's cost space.

In this paper, we make the following contributions:

- We introduce a confidence measure that requires no ground truth, no explicit model for the transformation, noise or images.
- Our confidence measure is directly correlated with the amount of mis-registration.
- Our confidence measure is independent from the type of registration algorithm as it only needs the fixed image and the output image of the registration.
- The full-search is applied over perpendicular directions. Since, full-search over each direction can be calculated independently, our method is very efficient in terms of GPU parallelization similar to stereo confidences.

Our paper is organized as follows: In Section II, we review the existing literature on error estimation in medical image reg-

istration and stereo confidence estimation. In Section III, we describe our method in detail. The experimental results are given in Section IV. Based on the experimental results, we elaborate on possible improvements for our method and draw conclusions in Section V.

II. Related Work

Confidence estimation for image registration has been the primary focus of many studies. Simonson *et al.* [18] proposed 2D-edge detection based matching with a McNemar test to find a confidence map assuming that the true transformation between images is only a rigid translation. Kybic [19] introduced a bootstrap-resampling based confidence estimation. In their method, image registration is performed several (100–1000) times. Similar to [20], [21], their method is computationally expensive. Additionally, Kybic [22] proposed a fast registration accuracy estimation method (FRAE) that is based on a Hessian similarity criterion for each transformation parameter. Although FRAE is faster than the bootstrap method, calculating the Hessian is computationally expensive. Furthermore, FRAE cannot measure the absolute registration error accurately [22]. In addition to these algorithms, there are other algorithms that can estimate the confidence of a registration as long as the utilized registration algorithm is formulated as a Bayesian framework [23], [24]. All of the above mentioned confidence estimation algorithms need to be integrated with the incorporated registration algorithm. Sofka *et al.* [25] included matching of key-points in an SVM-classifier to decide the correctness of an alignment. The features are extracted sparsely so a dense confidence estimation is not possible with their algorithm. Muenzing *et al.* [26], [27] employed a two-stage classifier cascade to classify the local alignment patterns into three classes: correct, poor, and wrong alignment. The main drawback of learning-based approaches is their limited number of classes to represent errors in the registration. Hence, the authors replaced their classifier with a regressor to obtain continuous confidence scores instead of discrete classes in [27]. However, they reported that using regressors instead of classifiers to obtain continuous measures did not achieve accurate results. Lotfi *et al.* [28], [29] used reinforcement learning to create an uncertainty measure for probabilistic image registration. The resulting measures were thresholded to indicate the degree of error into three classes; low, medium and high error. However, no explicit correlation with the amount of registration error was indicated. Learning techniques have also been used to predict errors in medical image segmentation. Kohlberger *et al.* [30] used a generic learning approach based on regression to estimate the Dice coefficient and the overlap error. The drawback of their algorithm is that the result does not indicate the spatial location of the errors. Crum *et al.* [31] proposed a residual-image based error detection algorithm, which uses a Gaussian scale-space to determine the scale of the registration error. As the main drawback, detection of the spatial locations of the registration error is not possible with their algorithm. Fedorov *et al.* [32] proposed using the robust Hausdorff distance on the edges of registered images. Their algorithm can assess the accuracy of an alignment at the edges of the image pairs. Park *et al.* [8] extracted the mismatch (residual) maps using mutual information (MI) and intensity-based radial dilation between the aligned pairs.

In this paper, we introduce a confidence measure based on stereo confidence measures that is densely computed for each voxel and is correlated with the registration error. Furthermore, our method is independent from the type of registration algorithm and requires no ground truth or user intervention.

III. Method

Our algorithm is composed of three steps: feature extraction, matching (full-search) and confidence estimation. In the following subsections, we describe each of these steps in detail.

A. Feature Extraction

In stereo matching, choosing robust features for constructing the cost space is an important step. The features should be representative and dense to find an estimation of the disparity for each pixel. It has been shown that the complexity and the accuracy of the features are closely related [33]–[35]. Although simple features such as intensity can be easily extracted, they cannot perform as well as the complex ones such as mutual information [36] and normalized cross correlation [37]. Sotiras *et al.* [3] described different features for image registration many of which are also used in stereo matching. It has been shown that a similar relation between complexity and accuracy of similarity measures also exist for image registration [4].

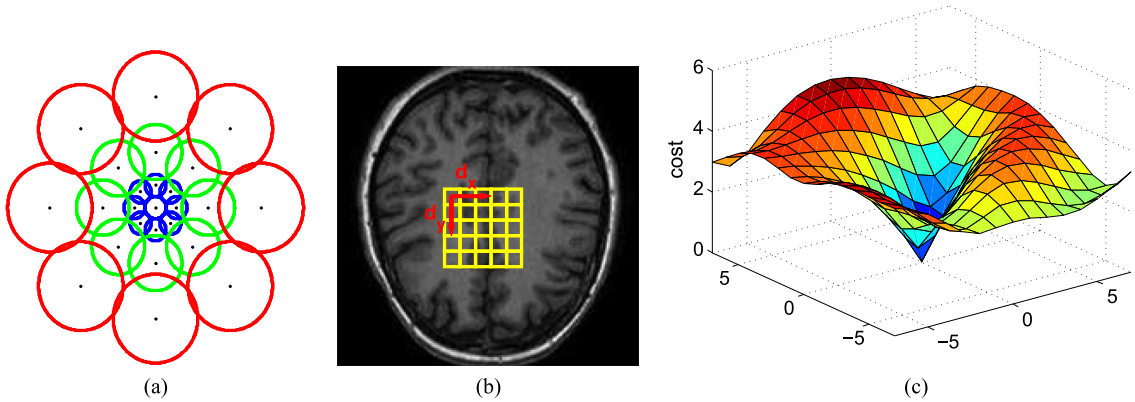
Daisy features [17] are robust descriptors that are designed for wide-baseline (long distance between camera pairs) stereo matching. It extracts gradient orientation histograms similar to SIFT [38] and GLOH [39]. Though, Daisy convolves the extracted histograms with different-sized Gaussian kernels to obtain weighted sums which makes it very efficient to compute. Furthermore, varying-sized Gaussian kernels provide descriptors at multiple scales. Fig. 1(a) represents the structure of the Daisy descriptor.

The Modality Independent Neighborhood Descriptor (MIND) [40] is a feature that is specifically designed for multi-modal image registration. For each voxel, MIND features are calculated using absolute intensity differences and the variation inside a local neighborhood. The calculated features can be incorporated in a similarity measure for both mono and multi-modal matching by using the Euclidean distance [28].

Even non-rigid deformations can be accurately modelled by local translations [15]. Hence, we opt to measure how well aligned the two registered images are by applying a full search in local neighbourhoods of all voxels. To do so, we use the Daisy and MIND features to obtain robust and dense descriptors which can be effectively computed and are prone to radiometric differences between the image pairs.

B. Matching—Full Search

Stereo matching algorithms aim to find corresponding points between the reference and target images. In stereo matching, the search for corresponding pixels is applied along a 1D horizontal line [41]. The distance function that calculates the difference between the extracted features creates a 1D cost space for each pixel. The confidence can be estimated for each pixel by using functions that analyze the shape of this cost space [10].



(a) (b) (c)

Fig. 1. (a) Structure of the Daisy descriptor [17], (b) full search over 2D local region. x and y axes, (c) corresponding cost space.

Medical image registration aims to align one (moving) image to another (fixed) image. Depending on the type of registration, algorithms may have thousands of parameters in total. Compared to stereo matching which has only one parameter to optimize, the disparity, it is a more challenging problem. Since stereo matching algorithms aim to find just one disparity for each pixel, it is computationally feasible to search all possible disparities and select the disparity that achieves the minimum cost. However for most image registration algorithms, full-search for each parameter is not feasible because of the excessive number of parameters. Therefore, stereo confidence measures cannot be applied directly in image registration algorithms. As a solution, we propose to compute our confidence measure from the cost space that we obtain by applying full-search over spatial directions for each voxel. Since any deformation can be modelled as local translations [15], we extract stereo confidences from the cost space that we obtain from a full-search over the local neighbourhood for each image element.

In contrast to stereo, corresponding voxels in medical images can exist in any location in their local neighborhood. Since there is no epipolar constraint in image registration, full-search should be applied over a 2D space as shown in Fig. 1(b), or over a 3D space. An example of the obtained cost space for 2D search is shown in Fig. 1(c). Since the aim of registration was to align the two images, the corresponding voxels (the global minimum of the cost space) should be located at the center of the search space for a correct registration.

After extracting features for each voxel, we calculate the Euclidean distance between the extracted features of fixed and moving images at each location in a local neighborhood. Taking the mean of the cost space for each voxel in a local neighbourhood is a common procedure of almost all stereo matching algorithms since the cost space of an individual pixel is noisy. Therefore, we also take the mean of each cost space over a fixed local neighborhood $N(\boldsymbol{x})$, to reduce noisy results.

Let $\boldsymbol{\psi}^f(\boldsymbol{x}_i)$ and $\boldsymbol{\psi}^m(\boldsymbol{x}_i - \boldsymbol{d})$ denote the extracted features of the voxels in the local neighborhood $N(\boldsymbol{x})$ of the center voxel at $\boldsymbol{x} = (x, y, z)$ and the target voxels that are located at $\boldsymbol{x}_i - \boldsymbol{d}$ in the full-search direction, in the fixed and deformed moving images, respectively. $\boldsymbol{d}$ indicates the amount of shift in 2D or 3D similar to the disparity in stereo matching. The cost space for the voxel at $\boldsymbol{x}$ is constructed as:

$$C(\boldsymbol{x}, \boldsymbol{d}) = \frac{1}{|N(\boldsymbol{x})|} \sum_{\forall \boldsymbol{x}_i \in N(\boldsymbol{x})} \left\| \boldsymbol{\psi}^f(\boldsymbol{x}_i) - \boldsymbol{\psi}^m(\boldsymbol{x}_i - \boldsymbol{d}) \right\|. \quad (1)$$

Let the global minimum of the cost space be at $\boldsymbol{d}^*$:

$$\boldsymbol{d}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{d}} \left(C(\boldsymbol{x}, \boldsymbol{d}) \right). \quad (2)$$

If the registration is optimal, we expect $\boldsymbol{d}^*$ to be at the center, $\mathbf{0}$, of the search space as in Fig. 1(c). Hence, correct alignment is achieved at that location.

C. Confidence Estimation

The shape of the cost curve (as depicted in Fig. 2) indicates how distinctive the matching voxels are compared to the neighboring voxels. The cost space like in Fig. 2(e) may indicate correct registration, whereas random shapes like in Fig. 2(k) point out possibly wrong alignments. On the contrary, voxels that share similar gradient along an edge may have a valley-like cost space shape as in Fig. 2(h).

Registration Maximum Likelihood (RML) is the confidence measure that we propose for image registration specifically. RML explores both the steepness and the location of the global minimum in the cost space of each voxel and derives a confidence measure for each voxel that reveals the quality of alignment.

A cost space with a steep global minimum as depicted in Fig. 2(e) indicates a correct alignment which also pinpoints the distinctiveness of a voxel at a particular location. Any shift of the global minimum from the center indicates a misalignment (error in registration). In order to make RML correlated with the registration error, we use a Gaussian function, $S_D(\boldsymbol{x})$, that is centered at the center of the search space. $S_D(\boldsymbol{x})$ penalizes the dislocation of the voxel at $\boldsymbol{x}$ proportional to the Euclidean distance of its shifted location to the center of search space, $\mathbf{0}$. Hence, independent from the type of deformation, as the global minimum gets shifted from the center of the search space, $S_D(\boldsymbol{x})$ decreases:

$$S_{\mathrm{D}}(\boldsymbol{x}) = \exp\left(\frac{-\left\| \boldsymbol{d}^*(\boldsymbol{x}) \right\|^2}{2\sigma_{\mathrm{D}}^2} \right). \quad (3)$$

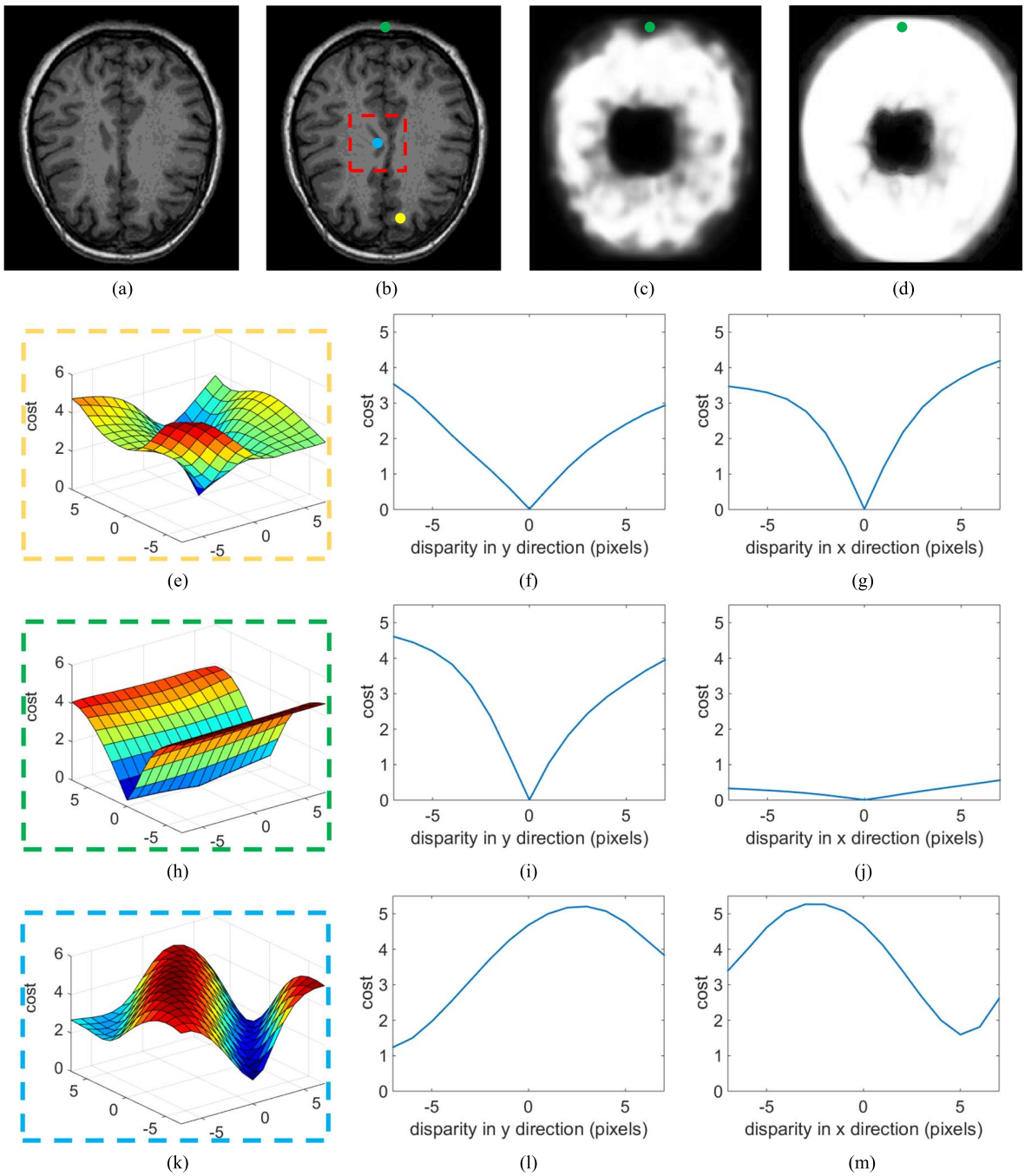


Fig. 2. Cost space examples extracted using Daisy features. The red region is the area with misalignment. All areas outside the red box are aligned perfectly. The yellow, green and blue points indicate locations with correct alignment not on edge, correct alignment on an edge, and misalignment, respectively: (a) original image, (b) synthetically deformed image, (c) RML confidence without cross-section, (d) RML confidence with cross-section, (e) cost space of the yellow spot, (f-g) cross-sections on y and x directions, (h) cost space of the green spot and (i-j) its cross-sections, (k) cost space of the blue spot and (l-m) its cross-sections at the center, respectively.

Similar to stereo confidences [10], RML should favour the cost spaces with a steep global minimum as in Fig. 2(e). However, voxels sharing similar gradients (structure) along the edges creates a valley-like shape in the cost space as depicted with green in Fig. 2. Favouring only cost spaces like Fig. 2(e) inevitably enforces the correctly aligned edge locations to have low confidence as depicted in Fig. 2(c). To obtain high confidence on the correctly-aligned edges, we take the cross-sections of the cost space along all search directions, $\boldsymbol{p}$, at the center of the search space as shown in Fig. 2(f) and Fig. 2(g). For each cross-section, we analyze the steepness of the cost curve around the center. If the registration is optimal, we expect to have at least one cross-section that has its steep global minimum at the center. Hence, we take the maximum confidence of all search directions:

$$\theta_{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{d}) = \frac{-f\left(C^{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{d}), C(\boldsymbol{x},\boldsymbol{0})\right)^2}{2\sigma^2_{\mathrm{RML}}}, \tag{4}$$

$$S^{\boldsymbol{p}}_{\mathrm{RML}}(\boldsymbol{x}) = \frac{S_{\mathrm{D}}(\boldsymbol{x})}{\left\|\exp\left(\boldsymbol{\theta}_{\boldsymbol{p}}(\boldsymbol{x},\boldsymbol{d})\right)\right\|}, \tag{5}$$

$$S_{\mathrm{RML}}(\boldsymbol{x}) = \max_{\boldsymbol{p}} S^{\boldsymbol{p}}_{\mathrm{RML}}(\boldsymbol{x}), \tag{6}$$

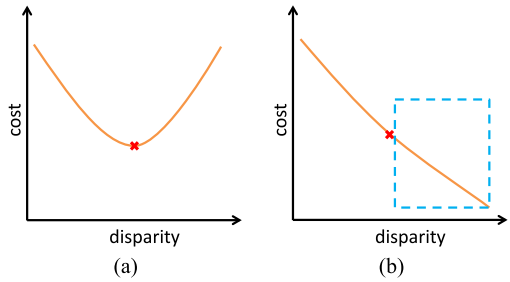


Fig. 3. Example cost functions to demonstrate the necessity of using $f(a,b)$: (a) Cost function with a steep global minimum at the center (indicated with red), (b) non-bell shaped cost function. Blue region indicates violation of global minimum at the center.

where $f(a,b)$ is a function to penalize the cost values that are smaller than the center value. A correct alignment in image registration is only possible when the global minimum of the cost space resides at the center of the search space, at $\mathbf{0}$, as depicted in Fig. 3(a). Therefore, RML should check the shape of the cost space at the center of the search space and favor the global minimum if it resides at the center. In case of any violation, RML needs to penalize the cost values that is smaller than the center value as indicated with blue in Fig. 3(b). $S_{\mathrm{RML}}(\boldsymbol{x})$ is maximal when $f(a,b)=0$ and gets lower as $f(a,b)$ increases or decreases:

$$f(\mathrm{a},\mathrm{b}) = \begin{cases} \mathrm{b}-\mathrm{a}, & \text{if a} > \mathrm{b} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The result of our RML confidence measure is depicted in Fig. 2(d). Different from Fig. 2(c), we obtain high confidence also on the correctly-aligned edge locations by using the cross-sections.

IV. Experiments

We conducted several experiments to test the performance of RML compared to other commonly used confidence measures. In our tests, we used both synthetical deformations and real registration ground-truth. In the following sections, we first introduce the details of our experiments and the dataset, then we elaborate on the experimental results.

A. Data and Experimental Setup

1) Data: In our experiments, we used three different datasets that include both lung CT and brain MRI scans. Lung CT images from 21 patients were recorded in the SPREAD study [42]. Each scan has around $446 \times 315 \times 129$ voxels. Additionally, there are 100 expert-marked ground truth points for each scan that we used for our real registration experiments. In addition to SPREAD, we used the adult maximum probability brain atlas (HAMMERS with atlas) dataset [43]–[45]. This dataset contains brain scans from 30 healthy adults and 83 manually segmented regions for each scan which we used as ground truth in our experiments. We also utilized the HAMMERS dataset in our synthetic deformation experiments. Finally, for our multi-modality experiments, we used the T1 and T2 brain MRI images from the RIRE dataset [46].

2) Compared Methods: For our expriments, we implemented two versions of our method. The first one applies full-search in horizontal and vertical directions. The second version also considers deformations between slices and constructs the cost function in three dimensions. The latter version requires more computational time and resources. Additionally, we used two different features, the Daisy and MIND descriptors for the full search. We conducted our synthetic simulations with both 2D and 3D deformations. To compare our method with existing confidences, we implemented confidence measures with the Hausdorff distance, MLM [47] and confidence obtained from the Daisy residual image [16]:

Hausdorff Distance. The Hausdorff distance is commonly used in medical image registration for error estimation. Fedorov *et al.* [32] incorporated the Hausdorff distance to measure confidence. In this paper, we calculated the Hausdorff distance between voxel intensities of the fixed and moving images.

MLM [47]. Maximum likelihood measure is among the best performing stereo confidence measures. It is similar to our confidence measure in terms of the underlying function. We incorporated the Daisy cost space to calculate the MLM confidence.

Residual (Difference) Image. Residual images are used to find the difference between two images. Park *et al.* [8] proposed adaptive registration algorithm based on residual images (mismatch maps) similar to confidence maps. Since we incorporated both the Daisy [17] and the MIND [40] features, we measured the residual images by taking the absolute difference of the extracted Daisy and MIND features of two images. We converted the residual images into confidence maps by subtracting each location from the maximum residual.

3) Error and Quality Measures: Medical image registration algorithms are often quantified using two different methods. One of these methods is the use of expert control points. Medical experts mark corresponding points between the two images. Each pair is expected to be at the same location after the registration. Hence, the amount of dislocation indicate the registration error. The second method is the use of segmentation maps. The transformation that is obtained after registration is applied to the segmentation maps of the two registered images. The Dice Similarity Coefficient (DSC) is calculated by measuring the overlap between the corresponding transformed segments. Both DSC and the amount of shift of expert control points are commonly used as quantitative measures of quality in medical image registration.

In addition to these methods, registration algorithms are often tested on synthetically created deformations [28], [31], [32]. One of the main reasons of using synthetically created deformations is its full control over the amount of induced error which is not affected by segmentation errors and expert-related mistakes.

A confidence measure is expected to indicate a wrong alignment in a synthetically deformed region with lower confidence than its correctly registered neighbors. Hence, the wrong alignment can be clearly presented to the observer. Fig. 4(d)–(f) show the results of different confidence estimation algorithms that are applied to the image pairs in Fig. 4(a) and 4(c). The brighter the region, the higher the confidence and vice versa. The result of RML in Fig. 4(f) clearly indicate the deformed region with low

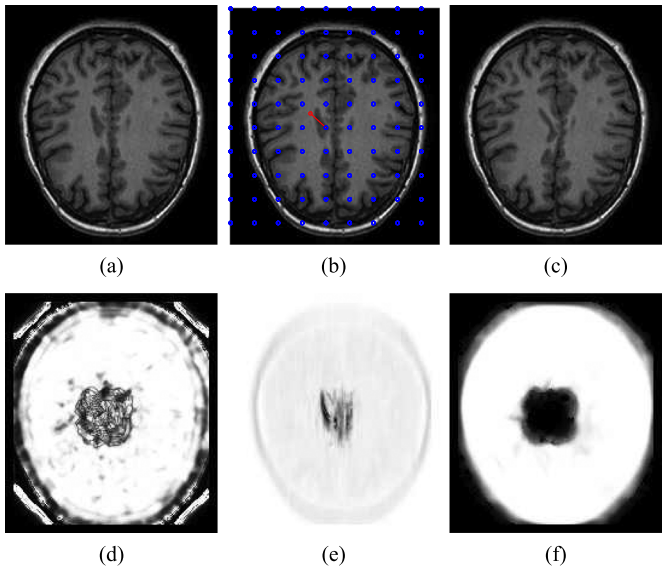


Fig. 4. Synthetically deformed image generation: (a) initial image, (b) synthetically-padded deformation (indicated with red), (c) resulting deformed image; confidence estimation results for: (d) stereo confidence (MLM [47]), (e) Hausdorff distance [32], (f) RML, respectively.

confidence. In order to measure how distinct the deformed region is, we computed the ratio of the mean confidence of the deformed region, $\boldsymbol{r}$, with the mean confidence of its non-deformed neighbor, $\boldsymbol{r}'$, and use this as the distinctiveness score, $S_T(\boldsymbol{r})$:

$$S_{\mathrm{T}}(\boldsymbol{r}) = \exp\left(\frac{-\bar{\boldsymbol{S}}_{\mathrm{RML}}(\boldsymbol{r})}{\bar{\boldsymbol{S}}_{\mathrm{RML}}(\boldsymbol{r}')}\right). \tag{8}$$

4) Experiments: We generated non-rigid deformations on random regions synthetically from the original images using grid points. To add deformation, one of the grid points was shifted from its original position to a different location as indicated with red in Fig. 4(b). Non-rigid registration was applied over the grid points (as ground truth landmarks) to align the two images. Since the two images were identical initially, mis-alignment occurred only in the neighborhood of shifted grid point as shown in Fig. 4(c). In our synthetical experiments, we randomly chose a grid point and dislocated it with a random shift. We repeated this step 50 times and calculated the mean and standard deviation of the errors with respect to the amount of shift on the grid point. As a result of our synthetical experiments, we expect to find the confidence measure that can represent the deformed regions more distinctively than the others. Furthermore, we opt to find a correlation between distinctiveness and the amount of mis-alignment in our synthetical experiments.

As for one of our real registration experiment, We explored if there is any correlation between confidences and registration error by considering the expert ground truth points after a real non-rigid registration using SPREAD lung CT data. We applied a rigid transformation that was followed by a B-spline non-rigid registration to align follow-up scans of patients. The obtained transformation after registration was applied to the expert control points in order to find their locations after the registration. The amount of dislocation was measured for every point to quantify the registration error. We expect to find a correlation between the dislocation and the confidences.

As our final experiment, we tested our confidence measure on ground truth segmentation maps using the median dice overlap score. Similar to the experiment with expert control points, non-rigid registration was applied between the two different images and the obtained transformation was used to align the segmentation maps of these images. As the result of this experiment, we opt to find a correlation between the DSC score and our confidence measure.

We implemented our method using Matlab and partially C++. We used fixed parameters for all of our experiments. The Daisy parameters were chosen the same as the default parameters in [17] except the radius, R, which was set to 5. Since we used two different features, we normalized the cost space first before applying our confidence measure. We set the parameters of RML experimentally. For the confidence measures, σ_{RML} was fixed to 0.02. σ_D was set to 3. Finally, the neighborhood for aggregation and the search radius were both set to 5. All of the parameters are held constant through the experiments.

B. Results

1) Confidence at the Synthetically-Deformed Regions Under 2D Deformation: The most important aspect of confidence estimation is to indicate erroneous regions. Therefore, the resulting confidence images have low intensity at deformed regions compared to their non-deformed regions. Fig. 5 shows the confidence maps between the image and its deformed replica with small (Fig. 5(a)) and large (Fig. 5(e)) deformations. The deformation is non-rigid and the deformed region is indicated in red ((Fig. 5(b), 5(f)). The confidence estimations are obtained using the proposed RML confidence with MIND and Daisy features. The confidence results show that, as the amount of deformation increases, the deformed region becomes darker in the confidence image. Hence, the distinctiveness of the deformed region in the confidence image is perceptually correlated with the amount of added deformation to the image.

2) Confidence at the Synthetically-Deformed Regions Under 3D Deformation: One of the challenges in medical image registration is the anatomical variances between image pairs. Any structural (anatomical) variance between image pairs can have a substantial importance and should be indicated by the confidence measure with low confidence. In order to test against structural differences between image pairs, we created inter-slice deformation and tested different confidence measures on the center slice. The results of the confidence estimation on the center slice after 3D deformation is depicted in Fig. 6. Different from the deformation in 2D, new shapes might appear as shown in red because of the inter-slice deformation. The results show that all the confidence measures can indicate the existence of the deformation. However the results of these algorithms are noisy since they also give high confidence in the deformed regions. The proposed algorithm clearly detects the inter-slice deformation with least noise compared to other confidences.

3) Distinctiveness Under 2D Synthetical Deformation: Fig. 7(a) shows the result of distinctiveness of the confidence measures for intra-slice deformations (in 2D). The higher the distinctiveness, the better the confidence measure. According to the results, stereo confidence performs similarly with the Hausdorff distance [32] and the residual-based confidences perform

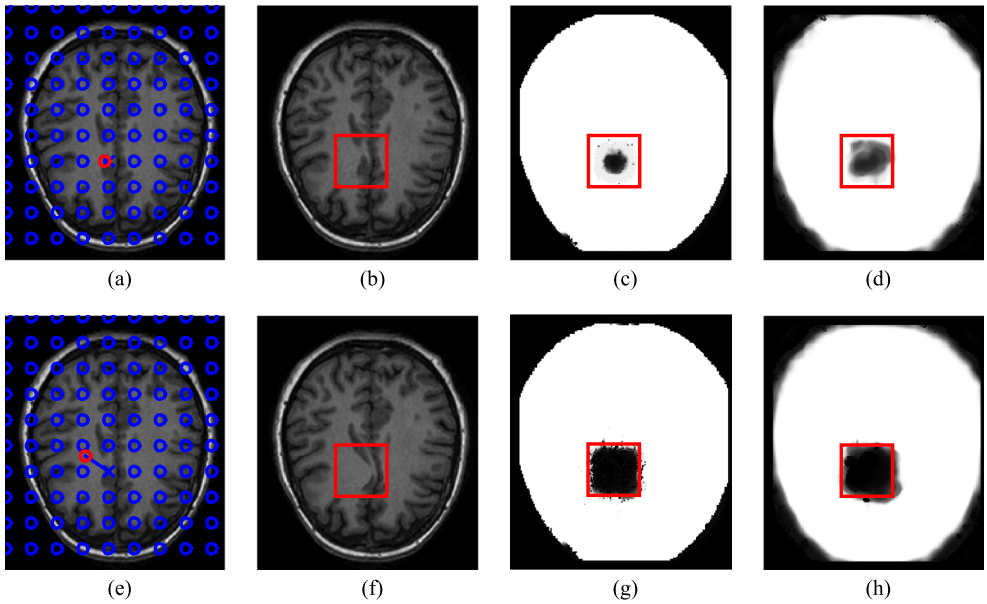


Fig. 5. The resulting confidence images under: (a) small, (e) large deformations, respectively: (b–f) resulting deformed images, (c–g) RML confidence results with MIND features, (d–h) RML confidence results with Daisy features.

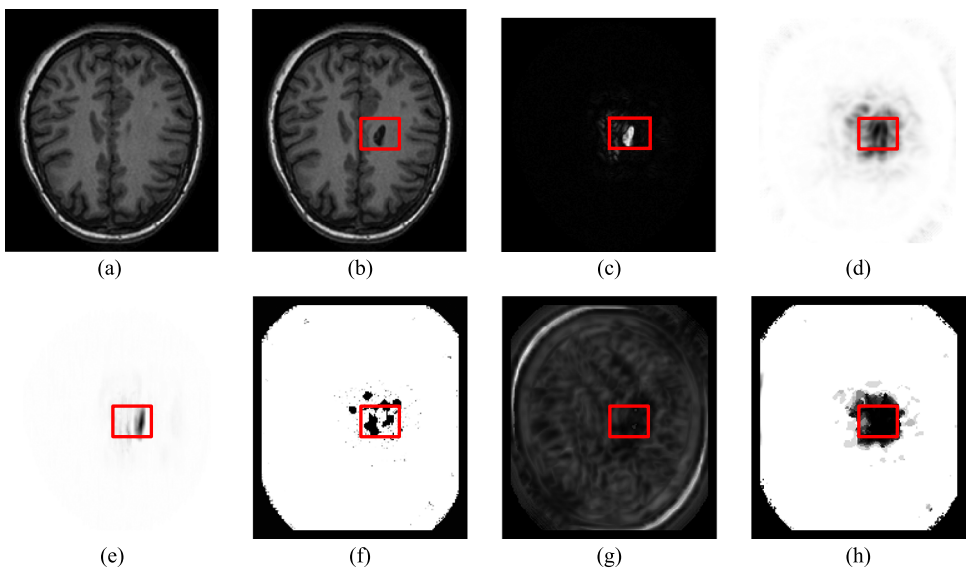


Fig. 6. Performance of confidence measures under 3D deformation: (a) original, (b) deformed images, (c) intensity difference, (d) Daisy residual, (e) Hausdorff, (f) MLM, (g) WMN, (h) proposed confidences, respectively. Red indicates the region with a significant deformation.

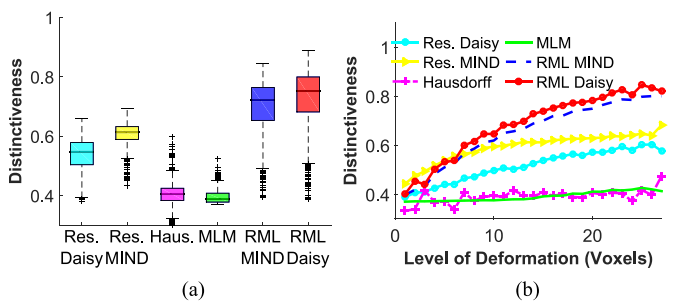


Fig. 7. Distinctiveness scores under 2D deformation, with single modality: (a) Distinctiveness of the region, (b) Distinctiveness with respect to the degree of deformation in voxels.

slightly better. Since we incorporate knowledge that the global minimum should reside at the center of the cost space as in (2), our confidence measure outperforms all other confidences. In this experiment, we incorporate various deformations between 0–25 voxels as depicted in Fig. 7(b). For a good confidence measure, the distinctiveness has a high correlation with the amount of deformation as shown in Fig. 7(b). The observed variance for the RML results in Fig. 7(a) is due to the fact that we span a large range of deformations. Note that each box represents 1500 points, and RML with Daisy and with MIND have 48 and 38 points below the lower whisker, respectively.

Fig. 7(b) shows the distinctiveness score with respect to the level of deformation. The result clearly shows that distinctiveness increases as the deformation amount increases with RML confidence. Furthermore, RML substantially outperform all other measures as the amount of deformation increases to higher levels.

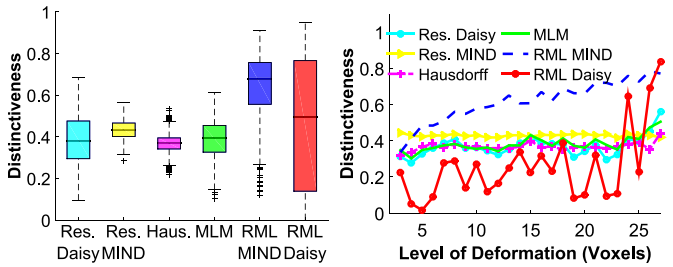


Fig. 8. Distinctiveness scores 2D synthetic simulation on multi-modal data: (a) Distinctiveness of the region, (b) Distinctiveness with respect to the degree of deformation in voxels.

4) Distinctiveness Between Images of Multiple Modality: Daisy features are designed to be used for wide-baseline stereo matching algorithms. In stereo matching, the images are from the same modality. However, images from multiple modalities can often be registered in medical image registration tasks. In contrast to Daisy, MIND features are designed specifically for medical image registration and build to be robust against the differences caused by multiple modalities. Therefore, it is important to test the performance of our confidence estimation algorithm between images from multiple modalities with both Daisy and MIND features. We performed our synthetical experiment between T1 and T2 intra-patient scans from RIRE dataset. The results are depicted in Fig. 8. Different from our previous results, RML confidence with Daisy features does not provide similar results with RML confidence with MIND features. The main reason for this result is the diversity of intensity gradient directions in between different modalities. Since Daisy relies on gradient orientations as a feature, the change of these directions in different modalities are recognized as a difference in alignment. Fig. 9 shows the effect of cross-modality and polarity differences on the qualitative results of RML with Daisy and MIND features. The region marked in blue indicates a region with a polarity change. Although there is no deformation in that region, RML with Daisy wrongly indicates low confidence. As depicted in Fig. 9(c), RML with MIND provides more accurate results on both non-deformed and deformed regions (indicated with red). Hence, RML with Daisy features has a noisy behaviour as in Fig. 8(b). In contrast to Daisy, RML with MIND features provide the best performance over all other confidence measures because of MIND's robustness.

5) Distinctiveness Under Gaussian Noise: The noise between two images is not just occur in between different modalities. To test our confidence measure against Gaussian noise, we performed two synthetical experiments that are applied on images with zero mean Gaussian noise. Fig. 10(a) and Fig. 10(b) show the results for sigmas of 0.01 and 0.1, respectively. The results indicate that RML with Daisy is the most robust against Gaussian noise and outperforms all other confidences. Both RML with MIND features and MIND residual confidences are outperformed by Daisy as the noise level increases. Even though the level of Gaussian noise affect the performance of RML with Daisy features, the correlation between the level of deformation and the distinctiveness is still observable as the sigma is increased to 0.1. In contrast, this correlation does not exist for RML with MIND features.

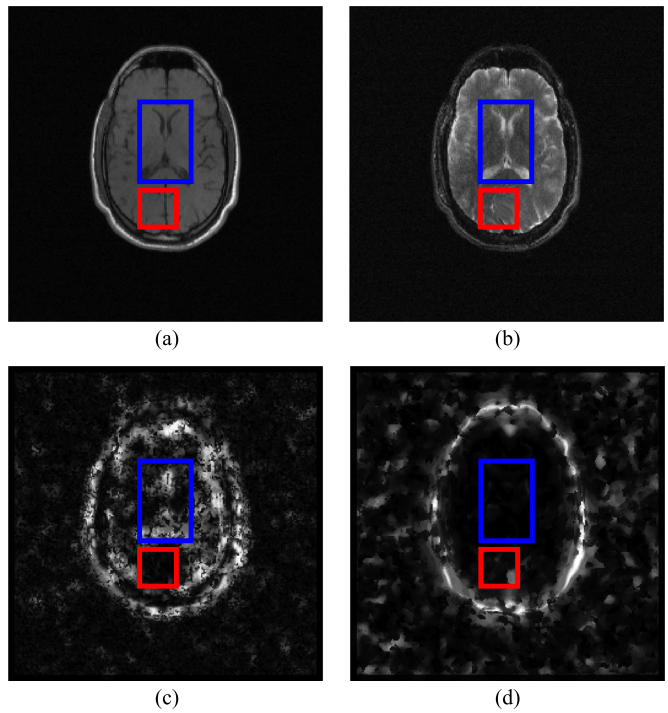


Fig. 9. The qualitative effect of multi-modality on RML confidence: (a) T1 and (b) T2 MRI scans, RML confidence results with (c) MIND, and (d) Daisy features, respectively. The blue region indicates an example of polarity difference. The red region indicates the region of deformation.

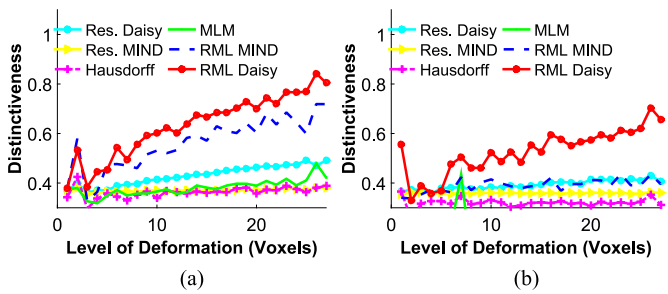


Fig. 10. Distinctiveness with 2D synthetic simulation under zero-mean Gaussian noise with sigma: (a) 0.01, (b) 0.1.

6) Distinctiveness Under 3D Synthetical Deformation: In this experiment, we explore the effect of deformations in three dimensions. Hence, we used 3D full search rather than 2D with both Daisy and 3D MIND features.

Fig. 11(a) shows the results of the ratio of the deformed and non-deformed regions for the simulation under 3D deformation. In both cases, our confidence measures outperform the other confidences. RML confidences based on MIND and Daisy features perform similarly.

Fig. 11(b) shows the result of our 3D simulation. Similar to our previous result for 2D, both of the RML results significantly outperform the other measures as the amount of deformation increases.

7) Correlation of Confidence With the Expert Ground Truth Data on Real Non-Rigid Registration: We tested the confidence measures with the ground truth data points that are marked by the experts with 3D deformations. The results are depicted in Fig. 12. Confidence is shown until 7 mm error, after which only

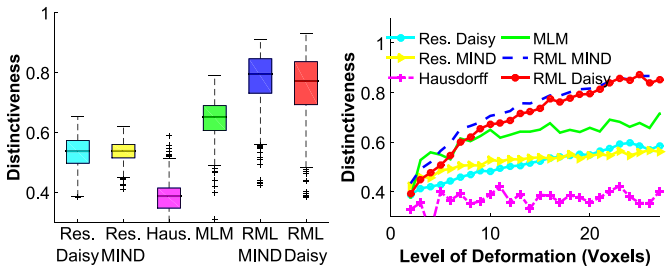


Fig. 11. Distinctiveness under 3D deformation: (a) Distinctiveness of the region, (b) Distinctiveness with respect to the degree of deformation in voxels.

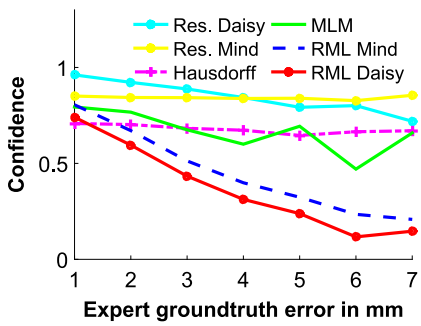


Fig. 12. The confidence levels for 3D full-search with respect to expert ground truth error in mm on the SPREAD lung data [42].

a few data points were available (<30), i.e. insufficient to reliably estimate confidence. Since we expect a strong correlation between a confidence measure and the amount of error in registration, the amount of confidence should decrease as the error increase. MLM confidence, Hausdorff distance measure and MIND residual do not show any correlation with the registration error. The Daisy residual shows weak correlation compared to the performance of our confidence measure. The results indicate that 3D RML confidence measures with both MIND and Daisy features are strongly correlated with the registration error from expert ground truth markers. We believe that the strong correlation is a consequence of our assumption that the matching points should reside at the center of the search range. This assumption cannot be used in stereo matching therefore direct use of stereo confidences do not show a correlation with the registration error. Since our confidence measures penalize the amount of translation between the positions of the corresponding points, they are able to correlate the registration error and the amount of confidence.

8) Confidence Estimation Correlation With DSC on Real Registration: We conduct another experiment using non-rigid registration with different parameters in order to obtain a variety of good and bad registrations. We used one of the most commonly used quantitative score, the Dice Score (DSC), to find if RML has any correlation with registration error. Fig. 13(a) and Fig. 13(b) represents the results of RML with MIND and Daisy features, respectively.

As the overlap between registered segments increases, the DSC increases, which indicates better registration. In both of the results, we see that the RML confidence increases as the DSC

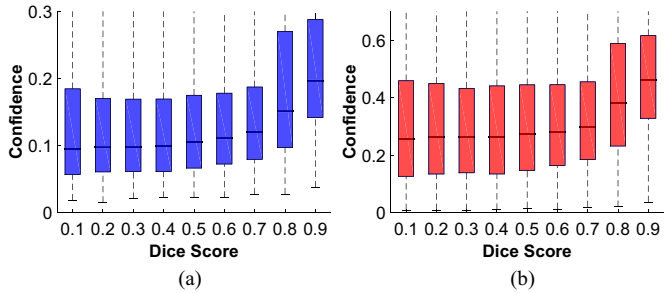


Fig. 13. Correlation of confidence with respect to regional Dice scores with real registration: RML with (a) MIND, and (b) Daisy features.

increase. The correlation is low for lower DSC scores and becomes more recognizable for higher values. This result together with the results from the expert ground control points proves that RML confidence measure can be used on real registrations as a quantification score.

Fig. 14 represents qualitative results of different confidence estimators on a real registration. The red-bounded regions indicate some of the locations with substantial perceptual deformation between fixed and moving images. The results show that RML with both Daisy and MIND features successfully locate these regions with low confidence. MLM and Daisy residual confidences can detect some of these regions whereas the remaining confidence maps fail to represent the erroneous regions. The result show that, in addition to synthetical deformations, RML can also qualitatively indicate deformations in real registrations.

V. Discussion and Conclusion

Our results show that stereo confidences and confidence estimation based on the residual image and the Hausdorff distance do not show strong correlation with the amount of registration error. Furthermore, their confidence images do not indicate the location of the error distinctively. On the contrary the proposed confidence measure can distinctively indicate the region of error and the amount of confidence is strongly correlated with the amount of registration error. The main reason for both observations is the implicit incorporation of the amount of disparity in the calculation of confidence.

In contrast to the discrete quality classes of [26], [27], the proposed confidence provides continuous confidence measures. The proposed confidence measure can potentially be clustered in discrete classes to represent e.g. high, medium and low confidence. The cluster boundaries can be chosen depending on the application at hand, as a post-processing step, which is not possible with [26], [27] Different from [31], the proposed confidence can distinctively indicate the location of the error and its calculation is computationally cheaper and highly parallelizable compared to [19] and [22]. For 2D search on 315×446 image, proposed algorithm executes in 29 seconds whereas 3D search executes in 321 seconds without any parallelization on an Intel i7 quad core CPU at 2.4 GHz.

We incorporated the Daisy and MIND features because they are efficient to compute and can densely extract important structural information for each voxel. From our experiments, we realize that the polarity of the gradient is important for the Daisy

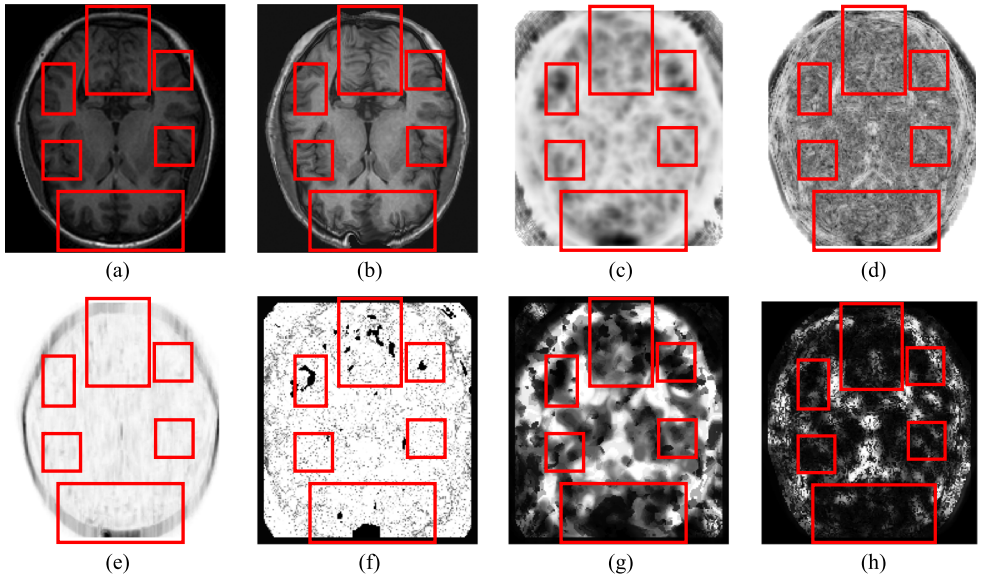


Fig. 14. Qualitative result for real registration: (a) fixed image, (b) moving image, (c) Daisy residual, (d) MIND residual, (e) Hausdorff, (f) MLM, (g) RML Daisy, (h) RML MIND, respectively. Red indicates the region with a substantial deformation.

features and MIND is more robust to multi-modality registration than Daisy. On the other hand, Daisy provides better performance under Gaussian noise than the MIND features. These results shows the importance of similarity measure in our confidence estimation. As a future step, more similarity measures with better robustness should be tested with our confidence measure to achieve better performances.

Apart from noise, there are other artefacts that affect the quality of registration such as non-standardness and intensity inhomogeneity [48]. In our experiments, we did not test the effect of these artefacts on our confidence estimation algorithm. In addition, we did not use any pre-processing of the input images. As a future study, we plan to first analyse their effect and then use standardization and non-uniformity correction approaches [49] as a pre-processing step to overcome such disturbances.

In our experiments, we used non-rigid deformations since such deformations have higher degrees of freedom compared to rigid and affine deformations. Any type of local deformation induces a shift from the center location and RML considers the amount of shift in terms of the Euclidean distance from the center location. Hence, independent from the type of deformation, the confidence at that location is penalized relative to the amount of the shift as in (3). Besides, RML is applied separately from the registration algorithm and only uses the fixed image and the output of the registration, it does not have a dependency on the type of registration. Hence, we expect that RML can be successfully used with all type of registrations including affine and rigid. Specifically, in case of a global transformation where all voxels are deformed similarly, we still expect locally distinctive confidence results since RML not only considers the amount of shift from the center, but also the steepness of the minimum of the cost space as in (5). Therefore, homogeneous (textureless) regions without a steep global minimum, such as the black corners of the images, will be penalized more than the regions with steeper global minimum. As a future work, we plan to explore the advantages of RML over other confidence measures with different types of transformation models.

Both Daisy and MIND features provide features that are extracted over a local neighbourhood. Since the deformation for each voxel is calculated by considering this neighbourhood, the deformation at one voxel may affect the non-deformed neighbouring voxels. In order to circumvent this, we can chose a smaller radius of neighbourhood for both features. However, using a small radius may induce problems in homogeneous regions since it restricts the local interactions between neighbouring voxels. Therefore, an adaptive choice of this radius depending on the local structure may provide better results.

At edges, neighbouring voxels share a similar gradient and valley-like cost space as discussed in Section III-C. We solve this problem by taking cross-sections in x, y and (if exists) z directions. However, by taking a cross-section and estimating the confidence over multiple dimensions independently, the algorithm does not efficiently explore the whole cost space in all dimensions. Therefore, we cannot exploit the additional information we obtain from all dimensions fully. As a future work, we will further improve our confidence measure so that it can both perform accurately on valley-like cost space and can exploit information from all dimensions fully.

Since RML explores structural similarity after registration, any pathological differences between the registered pairs are indicated with low confidence. We believe this is a useful property of our strategy since the pathological variations may provide valuable information to the medical experts. As a further study, we opt to exploit this capability of our algorithm in terms of providing a feedback to diagnosis algorithms.

In this paper, we proposed a confidence estimation method for medical image registration. Our method does not require a ground truth and is independent from the incorporated registration algorithm. In all of our experiments, our method produces the most accurate confidence maps, both quantitatively and qualitatively.

ACKNOWLEDGMENT

The authors would like to thank Dr. Nora Baka for valuable discussions.

REFERENCES

[1] J. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.
[2] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 46, no. 3, p. R1, 2001.
[3] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.
[4] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.
[5] G. K. Rohde, A. Aldroubi, and B. M. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1470–1479, Nov. 2003.
[6] T. Rohlfing and C. R. Maurer, Jr., "Intensity-based non-rigid registration using adaptive multilevel free-form deformation with an incompressibility constraint," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2001, pp. 111–119, Springer.
[7] J. A. Schnabel *et al.*, "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations," in *MICCAI*, 2001, pp. 573–581, Springer.
[8] H. Park, P. H. Bland, K. K. Brock, and C. R. Meyer, "Adaptive registration using local information measures," *Med. Image Anal.*, vol. 8, no. 4, pp. 465–473, 2004.
[9] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.
[10] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Nov. 2012.
[11] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 297–304.
[12] P. Mordohai, "The self-aware matching measure for stereo," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1841–1848.
[13] K.-J. Yoon and I. S. Kweon, "Distinctive similarity measure for stereo matching under point ambiguity," *Comput. Vis. Image Understand.*, vol. 112, no. 2, pp. 173–183, 2008.
[14] P. Steingrube, S. K. Gehrig, and U. Franke, "Performance evaluation of stereo algorithms for automotive applications," in *Comput. Vis. Syst.*, 2009, pp. 285–294.
[15] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Int. Conf. Comput. Vis.*, 1995, pp. 374–381.
[16] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[17] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
[18] K. M. Simonson, S. M. Drescher, Jr, and F. R. Tanner, "A statistics-based approach to binary image registration with uncertainty analysis," *IEEE Trans. Pattern Anal. Mach. Intell,*, vol. 29, no. 1, pp. 112–125, Jan. 2007.
[19] J. Kybic, "Bootstrap resampling for image registration uncertainty estimation without ground truth," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 64–73, Jan. 2010.
[20] M. J. Murphy, F. J. Salguero, J. V. Siebers, D. Staub, and C. Vaman, "A method to estimate the effect of deformable image registration uncertainties on daily dose mapping," *Med. Phys.*, vol. 39, no. 2, pp. 573–580, 2012.
[21] J. Sykes, D. Brettle, D. Magee, and D. Thwaites, "Investigation of uncertainties in image registration of cone beam ct to ct on an image-guided radiotherapy system," *Phys. Med. Biol.*, vol. 54, no. 24, p. 7263, 2009.
[22] J. Kybic, "Fast no ground truth image registration accuracy evaluation: Comparison of bootstrap and hessian approaches," in *Proc. 5th IEEE Int. Symp. Biomed. Imag. From Nano to Macro*, 2008, pp. 792–795.
[23] P. Risholm, S. Pieper, E. Samset, and W. M. Wells, III, "Summarizing and visualizing uncertainty in non-rigid registration," in *MICCAI*, 2010, pp. 554–561, Springer.
[24] F. Janoos, P. Risholm, and W. Wells, III, "Bayesian characterization of uncertainty in multi-modal image registration," in *Biomed. Image Registrat.*, 2012, pp. 50–59, Springer.
[25] M. Sofka and C. V. Stewart, "Location registration and recognition (LRR) for longitudinal evaluation of corresponding regions in CT volumes," in *MICCAI*, 2008, pp. 989–997, Springer.
[26] S. E. Muenzing, K. Murphy, B. van Ginneken, and J. P. Pluim, "Automatic detection of registration errors for quality assessment in medical image registration," in *Proc. SPIE Med. Imag.*, 2009, pp. 72 590K–72 590K.
[27] S. E. Muenzing, B. van Ginneken, K. Murphy, and J. P. Pluim, "Supervised quality assessment of medical image registration: Application to intra-patient ct lung registration," *Med. Imag. Anal.*, vol. 16, no. 8, pp. 1521–1531, 2012.
[28] T. Lotfi, L. Tang, S. Andrews, and G. Hamarneh, "Improving probabilistic image registration via reinforcement learning and uncertainty evaluation," in *Mach. Learn. Med. Imag.*, 2013, pp. 187–194, Springer.
[29] T. L. Mahyari, "Uncertainty in probabilistic image registration," Ph.D. dissertation, Applied Sciences: School of Computing Science, Simon Fraser Univ., Burnaby, BC, Canada, 2013.
[30] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012*, 2012, pp. 528–536, Springer.
[31] W. R. Crum, L. D. Griffin, and D. J. Hawkes, "Automatic estimation of error in voxel-based registration," in *MICCAI*, 2004, pp. 821–828, Springer.
[32] A. Fedorov *et al.*, "Evaluation of brain MRI alignment with the robust hausdorff distance measures," in *Advances in Visual Computing*, 2008, pp. 594–603, Springer.
[33] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Ana. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sep. 2009.
[34] G. Egnal, Mutual Information as a stereo correspondence measure CIS, Tech. Rep., 2000, p. 113.
[35] G. Saygili, L. van der Maaten, and E. A. Hendriks, "Adaptive stereo similarity fusion using confidence measures," *Comput. Vis. Image Understand.*, vol. 135, pp. 95–108, 2015.
[36] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
[37] Y. S. Heo, K. M. Lee, and S. U. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 807–822, Apr. 2011.
[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[39] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
[40] M. P. Heinrich *et al.*, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, 2012.
[41] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[42] J. Stolk *et al.*, "Progression parameters for emphysema: A clinical investigation," *Respiratory Med.*, vol. 101, no. 9, pp. 1924–1930, 2007.
[43] A. Hammers *et al.*, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Human Brain Map.*, vol. 19, no. 4, pp. 224–247, 2003.
[44] I. S. Gousias *et al.*, "Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest," *Neuroimage*, vol. 40, no. 2, pp. 672–684, 2008.
[45] A. H. Hammers *et al.*, Adult brain maximum probability map [Online]. Available: www.brain-development.org
[46] J. West *et al.*, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, no. 4, pp. 554–568, 1997.
[47] P. Merrell *et al.*, "Real-time visibility-based fusion of depth maps," in *IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[48] U. Bağcı, J. K. Udupa, and L. Bai, "The role of intensity standardization in medical image registration," *Pattern Recognit. Lett.*, vol. 31, no. 4, pp. 315–323, 2010.
[49] A. Montillo, J. K. Udupa, L. Axel, and D. N. Metaxas, "Interaction between noise suppression and inhomogeneity correction in MRI," in *SPIE Med. Imag.*, 2003, pp. 1025–1036.