An Automatic Quantitative Measurement Method for Performance Assessment of Retina Image Registration Algorithms

Ee Ping Ong[1], Jimmy Addison Lee[1], Guozhen Xu[1], Beng Hai Lee[1], Damon W.K. Wong[1]

Abstract—**This paper presents a novel automatic quantitative measurement method for assessment of the performance of image registration algorithms designed for registering retina fundus images. To achieve automatic quantitative measurement, we propose the use of edges and edge dissimilarity measure for determining the performance of retina image registration algorithms. Our input is the registered pair of retina fundus images obtained using any of the existing retina image registration algorithms in the literature. To compute edge dissimilarity score, we propose an edge dissimilarity measure that we called "robustified Hausdorff distance". We show that our proposed approach is feasible as designed by drawing comparison to visual evaluation results when tested on images from the DRIVERA and G9 dataset.**

I. Introduction

Image registration is the process where an image is aligned to another image taken from the same object or scene but in different situations, such as at different time, with different illumination, different viewing angle, and/or different imaging modalities [12, 2, 11, 8]. Image registration is important as a prior processing step before other processes such as image fusion, mosaicking, and retina verification could be carried out. Image registration methods are usually composed of three main components, namely, feature selection, feature correspondence, and transformation-model estimation. The registered image is then recovered by transforming one of the images using the transformation model obtained by the image registration algorithm.

As there is no universally accepted method and gold standard for automatic performance assessment of image registration, the usual means for performance assessment to date is to use either manual visual evaluation or manually annotated control points' method. In the performance assessment of image registration methods using manual visual evaluation, the reference and registered images are often superimposed in a checkerboard patterns to assist in human's visual evaluation [11, 8]. While visual evaluation is useful, it is insufficient to capture and quantitatively assign an error value (particularly for small errors/differences), and too slow and too laborious to compare the registration performance for many different image registration methods. Thus, there is a need for automatic quantitative measurement method. The availability of such automatic method can help to speed up research in image registration algorithms. It could also help to piece-and-match the different types of features, feature correspondence, and transformation model estimation methods to design a "best" image registration method for the image registration task on hand.

Also, it has been acknowledged that the manually annotated control points' method is not a good quantitative error measure [12] because the control points need to be manually annotated, the number of control points is usually limited due to the need to perform manual annotation which is a tedious and laborious process, and they do not capture the errors of all or at least most parts of the registered images since error measurement is just focusing on the limited number of control points only. It should be noted that using the average difference of the two registered image pair as a performance measure will not work because of illumination differences etc. between the image pair.

In [8], an automatic image registration performance approach has been proposed based on the goodness of superposition of a network of blood vessel centerline, where superposition is defined as the presence of another centerline pixel into a fixed size window centered on a reference centerline pixel (The authors empirically used a 3x3 window). The superposition percentage is the quantitative criterion of the registration performance. However, there are 2 major problems with this approach in [8]. Firstly, their method requires prior detection of the network of blood vessel centerline in both the reference and registered images. Accurate automatic detection of blood vessel centerline is itself a difficult problem. Secondly, the accuracy of the computed registration errors depends on the setting of the window size and it cannot measure the errors progressively. The larger the window is, the larger the registration error and thus failure to differentiate registration errors of varying degrees to different degrees of registration mismatch. For example, if a pair of registered images whose registered centerline pixels are all misaligned by 1 pixel each, their proposed method will still regard the registration result as being 100% perfect, as compared to another registered image pairs with perfectly superimposed network of vessel centerline. Hence, their proposed method is unable to give progressively varying quantitative errors to differentiate these cases.

[1] E.P. Ong, J.A. Lee, G. Xu, B.H. Lee, and D.W.K. Wong are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore { epong, jalee, xug, benghai, wkwong }@i2r.a-star.edu.sg

Here, we propose an automatic quantitative measurement method to assess the performance of image registration algorithms designed for registering retina fundus images. We exploit the availability of edges on blood vessel structures in retina images to aid in the design. To this end, we proposed an edge dissimilarity measure called "robustified Hausdorff distance" to achieve our aim. Compared to [8], our proposed method also overcomes the 2 problems faced by them. Firstly, we do not need to automatically detect the network of blood vessel centerline, which is difficult to extract accurately and robustly. Secondly, our approach can give progressive and distinctive error score for registered image pairs of varying mismatch in registration (and there is no empirical window size to set that affects registration accuracy and effectiveness). While our approach uses edges to replace the use of blood vessel centerline, edges are relatively easy to detect and our method does not require the edges to be detected accurately and robustly since our proposed "robustified Hausdorff distance" method is sufficiently robust against outliers and noise. Our solution is also more general and mathematically tractable, being based on the "Hausdorff distance" mathematical metric as opposed to the heuristic empirical fixed window-based method proposed in [8].

II. Proposed Method

To apply our approach, first we use any existing image registration algorithm to be assessed to obtain the registered image pair. Subsequently, we extract edges from the registered image pair, followed by applying our edge dissimilarity measure to obtain a quantitative dissimilarity score. Details are described in the following sections.

A. Edge Extraction

Any of the state-of-the-art edge detection techniques may be used to extract the edges. In our implementation, the Canny's edge detector [1] has been used to extract edges from the reference retina image and the registered retina image due to its reasonably good performance.

B. Edge Dissimilarity Measurement

The original directed Hausdorff distance metric [6] to measure the distance between 2 point sets is given by:

$$d(A,B) = \max_{a\in A} \min_{b\in B} d(a,b) \qquad (1)$$

where *a* and *b* belong to point sets $A = \{a_1,...,a_{N_A}\}$ and $B = \{b_1,...,b_{N_B}\}$ respectively, and *d*(*a*,*b*) defines the distance between the 2 points, *a* and *b*. Here, we will assume that this distance *d*(*a*,*b*) can be defined as the Euclidian distance between *a* and *b*, i.e. $d(a,b) = \|a-b\|$. We define the distance between a point *a* and a set of points *B* as $d(a,B) = \min_{b\in B} d(a,b)$. Then, the directed distance between 2 point sets *A* and *B* can be described by the above equation *d*(*A*,*B*). The Hausdorff distance, in essence, measures how far two subsets of a metric space are from each other.

The directed distances *d*(*A*,*B*) and *d*(*B*,*A*) between 2 point sets, *A* and *B*, can be combined to define an undirected Hausdorff distance measure, *D*(*A*,*B*), which is given by:

$$D(A,B) = \max\left(d(A,B), d(B,A)\right) \qquad (2)$$

Several variations of this set metric have been proposed as alternatives to the traditional Hausdorff distance matching so that the metric will be less prone to outliers in the data. These include the Hausdorff average [4] and Hausdorff ranked [7]. Huttenlocher et al. [7] proposed the ranked modified Hausdorff distance measure to handle outliers and occlusions as follows:

$$d(A,B) = {}^{x}K^{th}_{a\in A} d(a,B) \qquad (3)$$

where ${}^{x}K^{th}_{a\in A}$ represents the K^{th} ranked distance of *d*(*a*, *B*) ($\forall a \in A$, in ascending order), such that $K/N_A = x\%$, N_A being the number of points in the point set *A*. For example, ${}^{50}K^{th}_{a\in A}$ corresponds to median of distances *d*(*a*,*B*).

However, Dubuisson and Jain [4] found that Huttenlocher et al. [7]'s algorithm still present some problems for object matching and thus proposed the following modified Hausdorff distance measure:

$$d(A,B) = \left(\frac{1}{N_A}\right)\sum_{a\in A} d(a,B) \qquad (4)$$

In this paper, we propose a robustified Hausdorff distance (RHD) measure that is robust to outliers and noise points. The outliers and noise points could be the result from occlusions, dis-occlusions, edge extraction errors, differences in point sets due to degradation, and differences in input images etc.). The RHD is given by:

$$d(A,B) = \left(\frac{1}{K}\right)\sum_{i=1}^{K} \tilde{d}(a_i,B), \quad RHD(A,B) = \max\left(d(A,B), d(B,A)\right) \qquad (5)$$

where $\tilde{d}(a_i,B)$ represents the ranked distance of $d(a_i,B)$ ($\forall a_i \in A$, in ascending order), such that $K/N_A = z\%$, N_A being the number of points in point set *A*. With our formulation, the modified Hausdorff distance proposed by Dubuisson and Jain [4] becomes a special case of our RHD. If $K = N_A$, then we will obtain the same modified Hausdorff distance as [4], thus providing our proposed RHD the characteristic of being robust to noise (in comparison to [7]). In order to obtain sufficient robustness against outliers, occlusions, dis-occlusions etc., *K* is usually set less than $N_A/2$, thus giving our proposed method an outlier rejection/tolerance of 50% outliers or more. Hence, it can be seen that our proposed RHD incorporates the desired properties of being robust to noise and outlier points. Our proposed RHD is applied to the set of edges, *A* and *B*, obtained from the reference and the registered retina image respectively. The *RHD*(*A*,*B*) score determines the dissimilarity (i.e. quantity of differences) of the blood vessel structures between the registered retina image pair. The smaller this *RHD*(*A*,*B*) value is, the more similar is the similarity between the edge structure of the registered image pair (and hence the better the image registration process).

III. Results

We demonstrate our proposed method by applying it on two image registration techniques that are known to perform well on retina fundus images [2, 10] and then draw comparison to results of visual evaluation using checkerboard display. The image registration techniques we tested are:

1. Method 1: Using features known as bifurcation and cross-over points extracted from segmented vessels and feature points matching using point pattern matching approach with 4-parameter similarity transformation estimation (e.g. [10]);
2. Method 2: Using blob features such as scale invariant feature transform (SIFT) [9] extracted from the images and related nearest-neighbor matching method (where nearest-neighbor is defined as the feature point with minimum Euclidean distance for the SIFT descriptors) followed by RANSAC [5] with 6-parameter affine transformation estimation (e.g. [2]).

In all cases, the estimated transformation model is finally utilized to obtain the registered image. Also, in all our experiments, we set K (in equation (5)) to $N_A/3$, a choice which is a trade-off between robustness against outliers and fidelity of measurements (i.e. using more data versus using less data for the error score computation). As there isn't any standard retina image registration test dataset, we tested our proposed approach on retina fundus images from the DRIVERA dataset from Condurache et al. [3]. The DRIVERA dataset contains 280 retina images and is very challenging because of the numerous types of distortions and large variations present in the images.

Table 1 shows the quantitative performance measures of the image registration methods on DRIVERA dataset. Note that Method 1 gives numerous wrongly registered pairs of retina images. We manually examined the results of these wrongly registered pairs and found that they are mainly because two points or less have been matched. The similarity transformation model cannot be estimated using less than two matched points, and for those with only two matched points, the estimation is poor. On the other hand, Method 2 (with affine transformation estimation) uses SIFT that gives significantly more feature points that can be matched successfully for estimating the affine transformation model. In addition, the average value and the standard deviation of the quantitative error score of Method 2 is significantly much smaller than Method 1, hence indicating that Method 2 quantitatively performs better image registration than Method 1.

We also tested our proposed approach on another dataset called G9, consisting of 151 real-life retina fundus images captured from 18 different identities using a TopCon color fundus camera, with each identity having 8 or 9 images captured under different acquisition conditions. We rescaled all these retina fundus images to standard size of 1000 x 670 pixels in width and height. Table 2 shows the quantitative performance measures of the retina image registration methods on G9 dataset. Similar to results on DRIVERA dataset, the average value and standard deviation of the quantitative error score of Method 2 is still smaller than Method 1, hence indicating that Method 2 performs better image registration than Method 1. Also, it can be seen that Method 2 performs very much significantly better than Method 1 on DRIVERA dataset than G9 dataset. This may be because DRIVERA dataset is very much more challenging and also that Method 2 uses affine transformation model versus Method 1 using similarity transformation model.

Method	Average score	Standard deviation of score	Wrongly registered pairs
Method 1	0.5576	0.3441	25
Method 2 (+ affine transform)	0.1040	0.1744	0
Method 2 + similarity transform	0.7293	0.2507	0
Method 2 + projective transform	0.0857	0.1522	0

Table 1. Results for DRIVERA dataset (total 1820 pairs). (Note that wrongly registered image pairs are not used to compute the scores here).

Method	Average score	Standard deviation of score	Wrongly registered pairs
Method 1	0.0396	0.0851	0
Method 2 (+ affine transform)	0.0310	0.0767	0
Method 2 + similarity transform	0.0318	0.0787	0
Method 2 + projective transform	0.0291	0.0765	0

Table 2. Results for G9 dataset (total 572 pairs).

In addition to the above quantitative error measurement, we also performed visual evaluation on the obtained results. We used a checkerboard display (similar to [11]) interspersing alternate square regions of both the original reference image and the registered image to help us in our visual evaluation. We discovered that on those registered pairs with high error score, our visual evaluation assessment confirms that these are indeed those registered pairs with visually poorer registration results. From our visual evaluation experiments, we found that our visual assessment conforms to the results obtained using the automatic quantitative measurement.

For illustration, Figure 1 shows the results of one of the registered retina image pair from DRIVERA dataset obtained using Method 1, while Figure 2 shows the corresponding results obtained using Method 2. In Figure 1, (a) and (b) show the feature points extracted from the two input images (where the cross-over points are shown as blue "*" and bifurcation points shown as green "x"). In Figure 2, (a) and (b) show the SIFT feature points (shown as blue "x") extracted from the two input images. Meanwhile, (c) shows the matched feature points (shown as "*" and "x" joined by yellow lines) plotted on the two concatenated input images; and (d) shows the checkerboard illustration of the registered image pair. Visually, we can see that Method 2 provides better registration results than Method 1 (see Figure 1(d) and Figure 2(d)) in this registered retina image pair. Similar results can be seen for other registered retina image pairs from DRIVERA dataset that we had conducted the visual assessment.

To observe the effect of the transformation model, we conducted another experiment where Method 2 has been modified to be used with a 4-parameter similarity transformation model (instead of the 6-parameter affine transformation model). The results can be seen in Table 1 and 2. It can be seen that the performance of Method 2 with similarity transformation model deteriorates significantly to the extent that it now performs much poorer than Method 1.

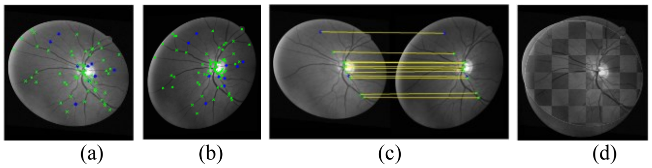

Figure 1. Results of Method 1: (a) Extracted feature points from image 1; (b) Extracted feature points from image 2; (c) Matched points; (d) Checkerboard display

We conducted another experiment again, this time where Method 2 has been modified to be used with a higher order transformation model, the 8-parameter projective transformation model. Table 1 and 2 show that Method 2 with projective transformation estimation gives better quantitative results than Method 2 with affine transformation estimation. However, when we performed visual assessment of the checkerboard display of registered image pairs using projective transformation model (e.g. see Figure 3(b)), our eyes generally are not able to detect any visible difference from those results produced using affine transformation estimation (e.g. see Figure 2(d)). When the number of matched points is sufficiently large, projective transformation model will, in theory, be expected to give better results than affine transformation model on image pairs with more complex transformation, but this difference cannot be easily detected using visual evaluation, thus highlighting the limitations of manual visual evaluation.

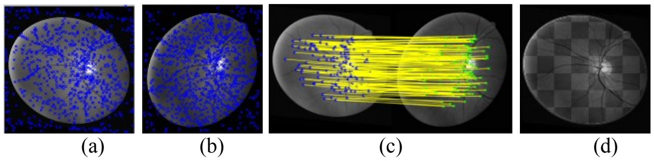

Figure 2. Results of Method 2 (with affine transformation model): (a) Extracted feature points from image 1; (b) Extracted feature points from image 2; (c) Matched points; (d) Checkerboard display

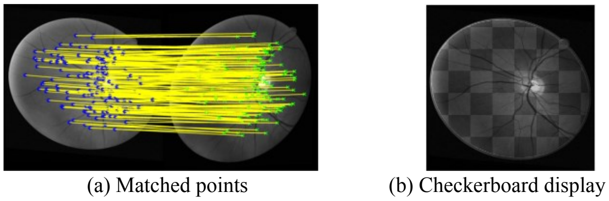

Figure 3. Results of Method 2 with projective transformation model.

In theory, Method 1 would benefit from adopting a higher order transformation model. However, in practice, it is not straight forward to modify the underlying original algorithm of Method 1 for use with higher order transformation model because its feature point matching is closely coupled to the transformation model used. On the other hand, it is relatively easy to modify Method 2 to be used with different transformation models, such as projective transformation model to get better results because the transformation model estimation is used in a second step with RANSAC after initial putative feature points matching has been performed.

IV. Conclusion

In this paper, we propose a novel automatic quantitative measurement method to automatically assess the performance of image registration algorithms designed for registering retina fundus images. We exploited the availability of blood vessel structures in retina images to help us to design this quantitative measurement method based on an edge dissimilarity measure, the "robustified Hausdorff distance". Our proposed method can help to give a quick and useful insight into the performance of retina image registration algorithms and can be used to complement/replace manual visual evaluation and also the manually annotated control points' method, both of which are tedious and laborious. Note that our experiments reported here only deal with mono-modal retina fundus images. Since blood vessels in retinal images are almost stationary structures and they are depicted in all modalities, they are reliable landmarks. Hence we believe our approach could be extended to deal with multi-modal images. This is an area that we will look into in our future work.

References

[1] Canny, J.: A computational approach to edge detection. *IEEE T. Pattern Analysis Machine Intell.*, Vol. 8, 679–698 (1986)

[2] Cattin, P.C., Bay, H., Gool, L.V., Szekely, G.: Retina Mosaicing Using Local Features. *MICCAI*, 185–192 (2006)

[3] Condurache, A.P., Kotzerke, J., Mertins, A.: Robust retina-based person authentication using the sparse classifier. *European Signal Processing Conf.*, pp. 1514–1518 (2012)

[4] Dubuisson, D.P., Jain, A.K.: A modified Hausdorff distance for object matching. *ICPR*, 566–568 (1994)

[5] Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Comm. of the ACM*, 24 (6), 381–395 (1981)

[6] Hausdorff, F.: Set Theory, Second Edition, Chelsea Publishing Co., New York (1962)

[7] Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. *IEEE T. Pattern Analysis Machine Intell.*, Vol. 15, 850–863 (1993)

[8] Laliberte, F., Gagnon, L., Sheng, Y.: Registration and Fusion of Retina Images – An Evaluation Study. *IEEE T. Medical Imaging*, Vol. 22(5), 661–673 (2003)

[9] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, Vol. 60(2), 91–110 (2004)

[10] Oinonen, H., Forsvik, H., Ruusuvuori, P., Yli-Harja, O., Voipio, V., Huttunen, H.: Identity verification based on vessel matching from fundus images. *IEEE ICIP*, 4089–4092 (2010)

[11] Ritter, N., Owens, R., Cooper, J., Eikelboom, R., van Saarloos, P.P.: Registration of stereo and temporal images of the retina. *IEEE T. Medical Imaging*, Vol. 18, 404–418 (1999)

[12] Zitova, B., Flusser, J.: Image registration methods - A survey. *IVC*, 21, 977–1000 (2003)