# Disease-Oriented Evaluation of Dual-Bootstrap Retinal Image Registration

Chia-Ling Tsai[1], Anna Majerovics[2], Charles V. Stewart[1], and Badrinath Roysam[1]

[1] Rensselaer Polytechnic Institute, Troy, NY 12180–3590
[2] The Center for Sight, 349 Northern Blvd., Albany, NY 12204

**Abstract.** This paper presents a disease-oriented evaluation of two recent retinal image registration algorithms, one for aligning pairs of retinal images and one for simultaneously aligning all images in a set. Medical conditions studied include diabetic retinopathy, vein occlusion, and both dry and wet age-related macular degeneration. The multi-image alignment worked virtually flawlessly, missing only 2 of 855 images. Pairwise registration, the Dual-Bootstrap ICP algorithm, worked nearly as well, successfully aligning 99.5% of the image pairs having a sufficient set of common features and 78.5% overall. Images of retinas having an edema and pairs of images taken before and after laser treatment proved the most difficult to register.

## 1 Introduction

Images of the retina are used to diagnose and monitor the progress of a variety of diseases, including such leading causes of blindness as diabetic retinopathy, age-related macular degeneration, and glaucoma [7]. Registering images taken weeks, months or years apart can be used to reveal changes in the retina at the level of small regions and individual blood vessels. Multimodal registration can reveal the relationship between events seen on the surface of the retina and the blood flow shown in the angiography.

Many retinal image registration algorithms have been proposed in the literature [4, 5]. Recently, we have developed two extremely successful algorithms which together simultaneously align all images in a set of two or more images of the same retina [8, 3]. The goal of the current work, as the next step toward wide-spread use, is a clinically-oriented validation of these algorithms. Clinicians are interested in knowing an algorithm's capabilities on a variety of diseases, on a variety of stages of the diseases, and as a patient progresses through these stages. We therefore selected a set of diseases to study, focusing on leading causes of blindness for aged population. A set of patients was selected for each disease, and retrospective images were collected for each patient across the progression of the disease. These image sets form the basis for validating the performance of our registration algorithm in a clinically-oriented framework.

## 2 Pairwise and Joint Registration

Registering a set of images is done in stages which we call "pairwise" and "joint" registration. In pairwise registration, the new Dual-Bootstrap Iterative Closest Point (DB-ICP) algorithm (Fig. 1) [8] is applied to each pair of images in the set. For each successful registration, a set of matching constraints is produced. Joint registration [3], applied to sets involving three or more images, takes these constraints and produces a globally consistent set of transformations (Fig. 1(f)). Only images that could not be matched to any other image in all attempts at pairwise registration are left out. The quadratic transformation is a $2 \times 6$ parameter matrix $\mathbf{\Theta}$, which maps image location $\mathbf{p} = (x, y)^T$ in image $I_p$ to location $\mathbf{q} = \mathbf{\Theta}\mathbf{X}(\mathbf{p})$ in $I_q$, where $\mathbf{X}(\mathbf{p}) = (1, x, y, x^2, xy, y^2)$. Both pairwise and joint registration are feature-based techniques, using automatically detected blood vessel centerlines and their branching and cross-over points [2, 9].

The core idea of DB-ICP is to grow an image-wide registration starting from initial estimates that are only accurate in small, "bootstrap" image regions (Fig. 1) (see [8] for details). Bootstrap regions are generated from hypothesized landmark correspondences and their surrounding vasculature. Hypothesized correspondences are generated by matching invariant signatures. Initial bootstrap regions are grown into image-wide transformations by iterating a three-step process:

**Estimating the transformation:** The transformation is estimated only in the "bootstrap region" (shown as the white box in Fig. 1(c)-(e)), using a robust form of ICP [1]. ICP matches are generated between the vessel centerline points.

**Region bootstrapping:** Based on the uncertainty (covariance matrix) in the transformation estimate, the bootstrap region is expanded. Stable, accurate estimates cause rapid growth, while unstable, inaccurate estimates cause slow growth.

**Model bootstrapping:** The similarity transformation model used in the initial bootstrap region is automatically switched to a higher-order model (eventually the quadratic) as the bootstrap region grows to cover the entire image.

The process terminates with success if one of the initial bootstrap regions tested can be expanded to a sufficiently-accurate, stable image-wide transformation.

The key idea of joint registration is that pairs of corresponding centerline points (in the final bootstrap region) from image pairs aligned by DB-ICP be mapped consistently when simultaneously transformed into any other image. This produces constraints on the final transformation estimates that ensures global consistency in all transformations, even image pairs that DB-ICP did not register. See [3] for details.

## 3 Pairwise Acceptance Criteria

The accuracy of the pairwise registration between two images is defined by the alignment of the vessel centerlines, termed the centerline error measure (CEM).

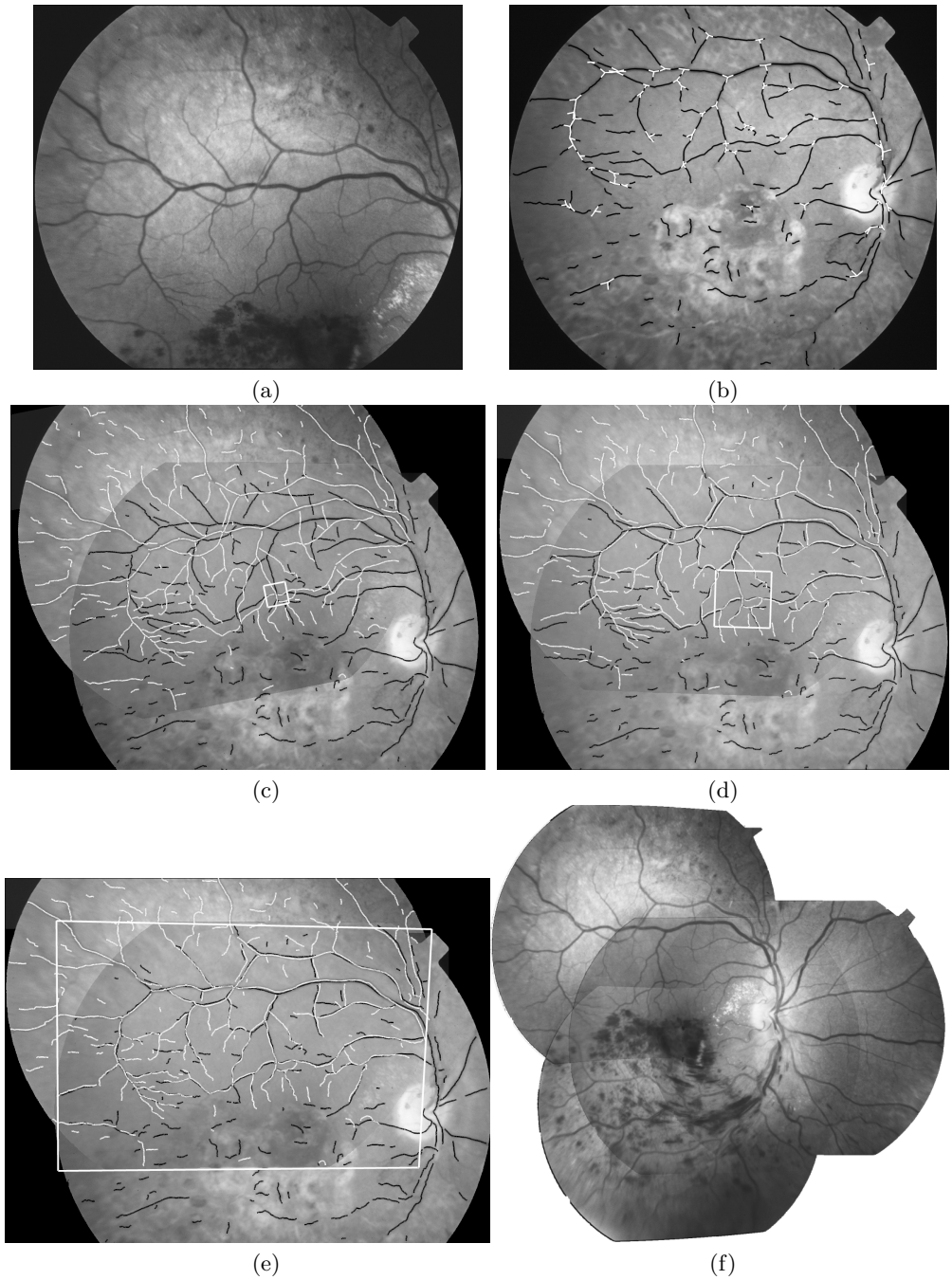


**Fig. 1.** Retinal image registration results on images of a patient with branch vein occlusion. Images in (a) and (b) were taken 3.5 years apart. Panel (b) is overlaid with the automatically extracted vessel centerlines and landmarks (branch and cross-over points). Panel (c) shows an initial alignment of the two images based on small regions ("bootstrap regions" — the white rectangle) surrounding a landmark correspondence. Extracted blood vessel centerline points from the two different images are shown in black and in white. Panels (d) and (e) show intermediate and final alignment results of Dual-Bootstrap ICP. Panel (f) shows the joint registration of all the images that taken at the same time as (a).

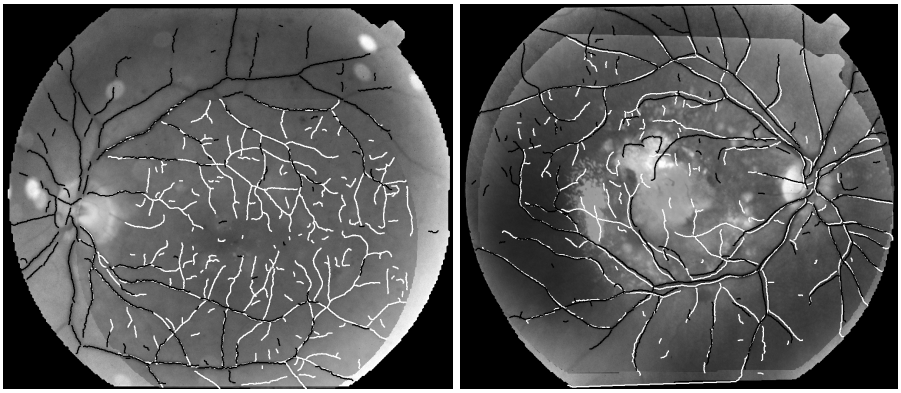

**Fig. 2.** Examples of image pairs to demonstrate the weighted error measure. The left shows registration of two Diabetic Retinopathy images taken 8 years apart, sharing 30% common traces. $\sigma_m$ is 4.2 pixels and $\sigma_w$ is 1.15 pixels. The right shows the alignment of two images of wet AMD, with $\sigma_w = 3.12$ pixels. This result is in the "grey area" of accuracy, which should be manually validated.

In [4], CEM is defined as the median of the alignment errors of the final trace point matches, and an empirical evaluation on images of healthy retinas produced a threshold of 1.5 pixels. Unfortunately, on images of diseased eyes, the median is not robust enough and the 1.5 pixel threshold does not accommodate changes caused by disease progression. We rectify the first problem here by proposing a weighted CEM. We address the second with an empirical study in Section 4.

Let $\{r_i\}$ be the set of alignment errors on final match set $\{(\mathbf{p}_i, \mathbf{q}_i)\}$. Each $r_i$ is defined as a point-to-line distance between the transformed $\mathbf{p}_i$ and the linear approximation of the centerline at $\mathbf{q}_i$. We assign a weight, $w_i$, to each match using the product of a robust registration error distance weight (from the Beaton-Tukey biweight function [6]) and a feature similarity weight. The new weighted CEM for error set $\{r_i\}$ is $\sigma_w = \left(\sum_i w_i r_i\right)/\left(\sum_i w_i\right)$. We denote the old median CEM as $\sigma_m$. Figure 2(a) clearly shows the superiority of $\sigma_w$. Numerically, we have found overall performance much better using $\sigma_w$, and it is used throughout our experiments.

## 4 Validation

The test dataset was formed from retrospective images of patients with four common diseases: Diabetic Retinopathy (DR), Vein Occlusion (VO), dry AMD and wet AMD. Ten representative retinas were chosen for each disease and six healthy retinas were added, giving a total of 46 retinas. Each retina (patient) had at least 3 visits over a time period as long as 5 years. Pathologies appearing in the diseased retinas include flame-shaped hemorrhage for VO, fibrosis for wet AMD, neovascularization for DR, and RPE detachment for dry AMD. Color slides were

pulled from the records at the Albany Center for Sight, scanned, and resized to $1024 \times 1024$ pixels. 855 images (producing 14,924 image pairs) were acquired. An additional 61 digital fluorescein angiogram sequences (Topcon IMAGENET) were obtained of different eyes, each with two digital red-free images.

***Results on Joint Registration*** We present results on joint registration first, even though it depends on the results of pairwise registration. The reason is that we use joint registration to validate pairwise — joint registration can recover from failures in pairwise. We need to know the performance of pairwise and joint separately because datasets can often involve a small number of images, perhaps just two. Also, when we say joint registration here, we mean joint registration of the entire image set for each eye, even images separated in time by five years.

Taking $\sigma_w = 1.5$ pixels for pairwise registration and then applying joint registration based on the resulting aligned image pairs, all images matched to some other image were aligned accurately by joint registration. The only errors were images completely unmatched by pairwise registration. These images all show small, but significant (for registration) changes in the position of the vasculature over time. Relaxing the accuracy to 3 pixels allowed all but two images to be jointly registered for the entire data set. (The missing two were of a single patient who had developed a fibrosis that obscured the entire retina.) We manually validated the resulting transformations.

The virtually flawless results of joint registration allow us to develop approximate upper bounds on the performance of pairwise registration. For any retinal image pair we can start from the "correct" transformation from the joint registration, and find an approximation to the correct set of correspondences (again, with the feature sets fixed). From there we can determine the covariance of the transformation estimate. If the condition number of this matrix indicates that the transformation is sufficiently stable and $\sigma_w$ is less than 1.5, we say that an accurate feature-based pairwise registration possible. We term such a transformation "stable". As a result for each evaluation below we give two success rates. We use the notation $\mathcal{S}_a(\mathcal{S}_s)$, where $\mathcal{S}_a$ is the absolute success rate by considering all pairs, and $\mathcal{S}_s$ is the stable success rate considering only pairs with stable transformations. Note that at this point we do not consider increasing the threshold on CEM beyond 1.5.

***Pairwise — Overall Results*** The overall success rate of DB-ICP pairwise registration is 78.5%(98.5%). If we only consider pairs that have at least one common detected landmark (and therefore one possible starting point for DB-ICP), the result is 99.5%. There are two reasons for the difference between the absolute and "stable" success rates. The first is a lack of common features between image pairs. There is little that can be done in the DB-ICP algorithm about this (though perhaps something can be done in feature extraction). The second is the effect of shifts in position of the vasculature over time. We return to this later. The overall conclusion is that the DB-ICP algorithm itself is virtually flawless in finding a correct registration from the feature-sets if one is possible to find.

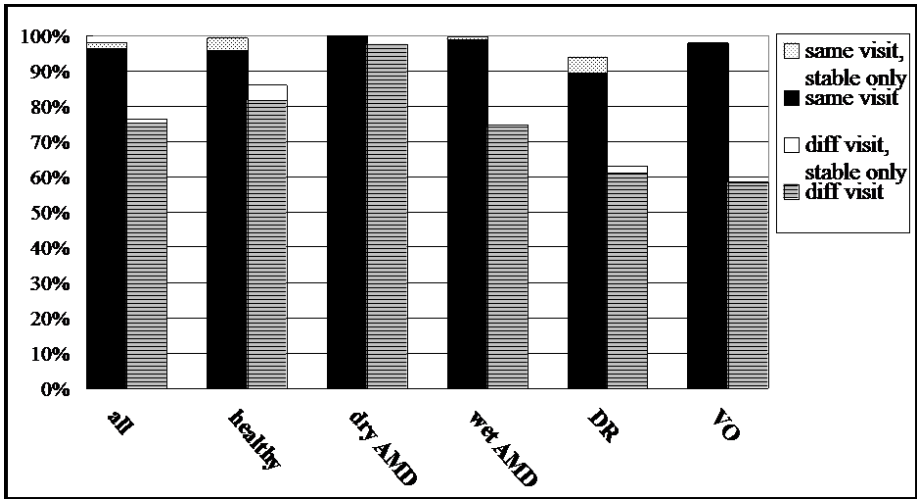


**Fig. 3.** Validation results by medical conditions. The plot shows the results of registering image pairs taken from the same visit and pairs from different visits, separately. Each bar shows the percentage succeeded pairwise, with the success rate for stable transformations added to the top.

***Results by Medical Condition*** Focus on individual medical conditions, the highest success rate is in dry AMD — 97.93%(100%) — which is expected since dry AMD does not have pathologies that affect the position and appearance of the vasculature. The success rates for healthy, DR, wet AMD and VO are 83.8%(95.1%), 65.5%(96.7%), 77.5%(99.1%), and 65.9%(99.7%), respectively. The lower $\mathcal{S}_s$'s for healthy retinas and DR result are an artifact of the image acquisition process: wider coverage of the retina is required for evaluation and therefore there is lower overlap between images.

***Same Visit / Different Visit*** We can further break down the results into "same visit" and "different visit" to analyze the effect of longitudinal changes on the algorithm (see plot 3(b)). The overall success rates are 96.4%(98.3%) and 75.3%(98.5%), respectively. The difference in $\mathcal{S}_a$ is mainly due to longitudinal changes. As the medical condition progresses, the retinal surface and vasculature tend to undergo changes as a result of edema, fibrosis, appearance/disappearance of pathologies, etc. The exception to this is the low $\mathcal{S}_s$ for the healthy set — surprisingly these images were generally of much poorer quality (and there are fewer of them in the clinical records!).

***Results by Medical Events*** Analyzing results by medical conditions is important but relatively difficult in a retrospective study of the current size. We examine the results in two ways. First, we partition the diseases into those that can cause an edema — as swelling of the retina surface due to build up of fluid — and those that can not. (Note that we don't have records of whether or not an edema was present.) The significance of this is that an edema causes a

| | all | w/ edema | w/o edema | same visit | different visit |
|---|---|---|---|---|---|
| $\sigma_w < 1.5$ | 78.5 | 70.5 | 93.5 | 96.4 | 75.3 |
| $\sigma_w < 4$ | 92.2 | 89.3 | 97.6 | 97.2 | 91.3 |

**Table 1.** Absolute success rates ($\mathcal{S}_a$) using CEM thresholds of 1.5 and 4 pixels. As expected, the biggest differences are for image pairs of diseases that cause edema and for image pairs taken at different visits.

misalignment due to inconsistencies with the quadratic surface model underlying the transformation. Our results show that the edema causing diseases — wet AMD, VO and DR have lower success rates — 70.5%(98.4%) — than non-edema causing conditions 93.5%(98.5%).

The second analysis based on medical condition is more precise — effect of laser surgery. We compare two sets of image pairs: the first contains pairs having one image before and one after surgery, and the second contains pairs before or without surgery. The success rates are 63.1% (98.1%) and 91%(99%) respectively. The difference is that laser treatment causes swelling and scarring, which shifts the position of the vasculature, thereby causing misregistration.

***Results on Fluorescein Angiograms*** The circulation of fluorescein defines five successive stages of the FA image sequence: arterial, arteriovenous, venous, late venous and recirculation. We define the success rate for a specific phase as a fraction of sequences for which joint registration successfully aligned all images in the sequence up to and including the phase. 100% success rate was achieved up to venous phase, 92% for late venous and 75% for recirculation. Failures are caused by obscuring of the vasculature due to leakage in vessels and a resulting pooling of the fluorescein dye.

***Upper Bound on CEM for Pathological Data*** A CEM of 1.5 pixels causes some correct registrations to be labeled as incorrect. Diseased eyes where an edema is present cause slight misregistration because the quadratic surface model becomes a less accurate representation. The scarring following laser treatment causes tractional movement of blood vessels. Algorithmically, the misregistrations appear as mis-alignment of the traces in small image regions (see Figure 2(b)) — "regional mis-alignment". Raising the CEM threshold allows these registration to be classified as correct, but the danger is that some true misalignments would then be called correct as well. Therefore, we'd like to determine a second, "gray area" threshold. Calling this threshold $C_2$ and calling the original threshold $C_1$, we develop a three-part classification to registration results: when $\sigma_w \leq C_1$ the registration is accepted as correct; when $C_1 < \sigma_w \leq C_2$, the registration is provisionally accepted and presented to the clinician for verification; when $C_2 < \sigma_w$ the registration is rejected. We empirically determine this threshold as $C_2 = 4.0$ by comparing the pairwise and joint transformations. Table 1 shows the fraction of image pairs (all pairs) below $C_1$ and below $C_2$.

Using the new threshold, the improvement on $\mathcal{S}_a$ is shown in Table 1. The results fit our intuition, since most mis-alignments are from surface deformation.

## 5 Discussion and Conclusion

This disease-oriented evaluation has demonstrated the capabilities of our two-part registration technique — the Dual-Bootstrap ICP pairwise algorithm and the multi-image joint registration algorithm that builds on DB-ICP results — in aligning retinal images in a clinical framework. The multi-image algorithm aligned all but 2 out of 855 images. DB-ICP pairwise registration aligned 78.5% of all pairs, 98.5% of the pairs for which a stable transformation exists (based on the exisiting set of features), and 99.5% of the stable pairs having at least one landmark in common. No incorrect registrations were accepted. Raising the acceptable threshold to 4.0 allowed image pairs to be considered in a grey zone that clinicians should check for accuracy. 92.2% of all pairs had a registration error of at most 4.0. Edema and the longitudinal effects of laser treatment caused the significant misalignments, and fibrosis completely obscured the vasculature in the two images that failed completely. Our overall conclusion is that the Dual-Bootstrap ICP pairwise and multi-image joint registration algorithms are robust and reliable enough for a variety of clinical uses.

## References


[1] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. on PAMI*, 14(2):239–256, 1992.

[2] A. Can, H. Shen, J. N. Turner, H. L. Tanenbaum, and B. Roysam. Rapid automated tracing and feature extraction from live high-resolution retinal fundus images using direct exploratory algorithms. *IEEE Trans. on Info. Tech. for Biomedicine*, 3(2):125–138, 1999.

[3] A. Can, C. Stewart, B. Roysam, and H. Tanenbaum. A feature-based algorithm for joint, linear estimation of high-order image-to-mosaic transformations: Mosaicing the curved human retina. *IEEE Trans. on PAMI*, 24(3):412–419, 2002.

[4] A. Can, C. Stewart, B. Roysam, and H. Tanenbaum. A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina. *IEEE Trans. on PAMI*, 24(3):347–364, 2002.

[5] W. Hart and M. Goldbaum. Registering retinal images using automatically selected control point pairs. In *Proc. IEEE Int. Conf. on Image Processing*, volume 3, pages 576–581, 1994.

[6] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Commun. Statist.-Theor. Meth.*, A6:813–827, 1977.

[7] J. Kanski. *Clinical Opthalmology.* Butterworth-Heinemam, 4 edition, 1999.

[8] C. Stewart, C.-L. Tsai, and B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Trans. on Medical Imaging*, accepted, to appear 2003.

[9] C.-L. Tsai, C. Stewart, B. Roysam, and H. Tanenbaum. Repeatable vascular landmark extraction from retinal fundus images using local vascular traces. *IEEE Trans. on Inf. Tech. in Biomedicine*, to appear 2003.