A Feature-Based, Robust, Hierarchical Algorithm for Registering Pairs of Images of the Curved Human Retina*

Ali Can[1†] Charles V. Stewart[2] Badrinath Roysam[1] Howard L. Tanenbaum[3]

[1]Dept. of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, New York 12180–3590
cana,roysab@rpi.edu

[2]Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180–3590
stewart@cs.rpi.edu

[3]The Center for Sight
349 Northern Blvd.
Albany, NY 12204
how1@albany.net

May 25, 2001

Abstract

This paper describes a robust hierarchical algorithm for fully-automatic registration of a pair of images of the curved human retina photographed by a fundus microscope. Accurate registration is essential for mosaic synthesis, change detection, and design of computer-aided instrumentation. Central to the new algorithm is a 12-parameter inter-image transformation derived by modeling the retina as a rigid quadratic surface with unknown parameters, imaged by an uncalibrated weak perspective camera. The parameters of this model are estimated by matching vascular landmarks extracted by an algorithm that recursively traces the blood vessel structure. The parameter estimation technique, which could be generalized to other applications, is a hierarchy of models and methods: an initial match set is pruned based on a 0^{th} order transformation estimated as the peak of a similarity-weighted histogram; a 1^{st} order, affine transformation is estimated using the reduced match set and least-median of squares; and the final, 2^{nd} order, 12-parameter transformation is estimated using an M-estimator initialized from the 1^{st} order estimate. This hierarchy makes the algorithm robust to unmatchable image features and mismatches between features caused by large interframe motions. Before final convergence of the M-estimator, feature

*An earlier version of this work received the Best Paper Award at the 1999 IEEE Conference on Computer Vision and Patter Recognition[15]

†Dr. Can is now an Assistant Professor at Inonu University, Turkey

positions are refined and the correspondence set is enhanced using normalized sum-of-squared differences matching of regions deformed by the emerging transformation. Experiments involving 3000 image pairs (1024x1024 pixels) from 16 different healthy eyes were performed. Starting with as low as 20% overlap between images, the algorithm improves its success rate exponentially, and has a negligible failure rate above 67% overlap. The experiments also quantify the reduction in errors as the model complexities increase. Final registration accuracies less than a pixel are routinely achieved. The speed, accuracy, and ability to handle small overlaps compare favorably with retinal image registration techniques published in the literature.

Index terms: Robust estimation, registration, transformation estimation, image mosaic, retinal imaging, feature extraction, feature refinement, multiscale methods, ophthalmic image processing, biomedical image processing.

1 Introduction

The problem of registering two or more data sets is fundamental to many applications of computer vision and medical image analysis. These applications are as diverse as industrial inspection, aerial image analysis and cartography, virtual reality, surgical planning, change detection, and treatment monitoring. Solving the registration problem requires estimating the transformation(s) between data sets and applying them to place the data in a common coordinate system. The data may be two-dimensional intensity images [12], range images [10, 18], or volumetric images such as CT scans, MRIs, or confocal stacks [3, 21]. Many techniques have been proposed to solve the registration problem in many different forms. For reviews, see [12, 21, 32].

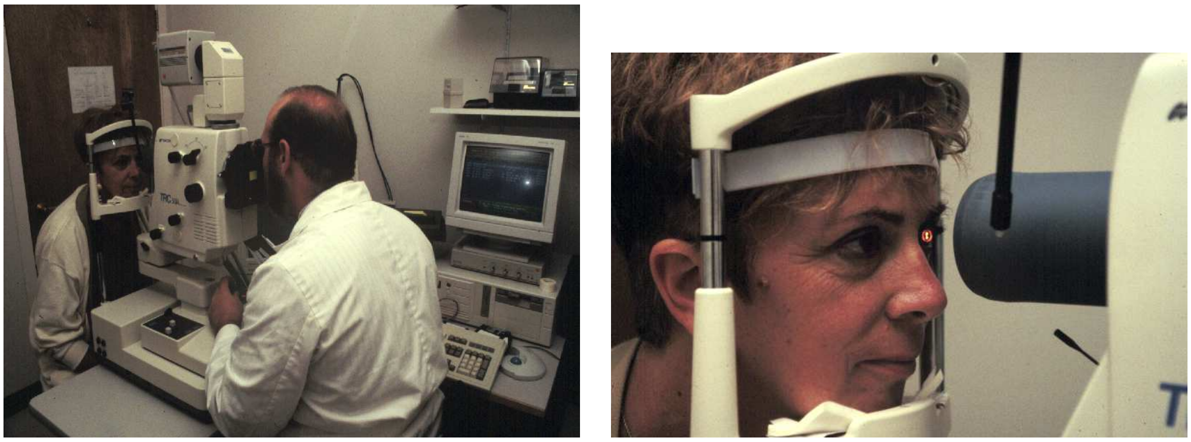
One medical application where registration techniques are increasingly important is in automated techniques to assist in the diagnosis and treatment of diseases of the human retina. For instance, two images taken before and after laser surgery can be registered to detect locations of scars and burns [7]. Two images taken of the same eye but at different times can be registered to track the progress of diseases such as macular degeneration, glaucoma and AIDS/CMV retinopathy [31, 48]. A series of images of the same retina can be registered to form a mosaic image, giving a complete view of the retina [33]. Finally, real-time (frame-rate) registration methods may be used as the basis for tools that assist ophthalmologists during laser surgery and other procedures [5, 8, 9].

Despite the importance of the problem and prior research effort, retinal image registration has remained a difficult problem [5, 20, 26, 33, 35]. Several challenges must be addressed in developing reliable automatic registration techniques. These challenges are special cases of difficulties that arise in other applications.

- The surface of the retina is curved, almost spherical. The inter-image transformation model must take this into account.
- The images are acquired using an imaging system whose optics include the unique optics of each eye, making prior calibration of the system difficult.
- Illumination, which comes from outside the eye, is necessarily viewpoint dependent and can cause glaring as well as fade-outs. As a result, a region of a retina might have substantially different intensity properties in different images (Figures 1 and 2).
- Image overlap may be small due to large changes in viewpoint between images.
- Large regions of retinal images are relatively textureless. The predominant features are the blood vessel structures and, in some images, the optic disk.
- Blood vessel widths can be as narrow as two or three pixels. Precise registration is therefore necessary for accurate change detection and mosaic formation.

This paper is the first of two devoted to solving the retina image registration problem. The current paper addresses the problem of registering a pair of retinal images. The second paper [16] addresses the problem of simultaneously registering multiple retinal images to form a mosaic. Together, they describe a set of algorithms that are currently being tailored to several ophthalmic applications.

Three main contributions are described in this paper. The first is the derivation of a 12-parameter inter-image transformation model that accounts for the curvature of the retinal surface and the motion of the eye between images. The second is a hierarchical, robust, algorithm for estimating the parameters of this model given a pair of retinal images. The algorithm, which could be generalized to other applications, uses branching points and cross-over points in the retinal vasculature detected using a recursive tracing algorithm [14]. The final contribution is a method to refine feature positions while estimating the transformation parameters. This refinement step is necessary for achieving the final accuracy of the registration. Extensive experimental results demonstrate the effectiveness of our model and registration techniques, and provide important quantitative insights into the accuracy and reliability of the algorithms. Overall, the algorithm is faster and provides more accurate registration on higher-resolution images than algorithms published in the retinal image analysis literature.



a b

Figure 1: The fundus camera imaging set-up: (a) shows the patient in an imaging harness, and (b) shows the ring of illumination.

Figure 2: Three example 1024×1024 images of the retina surface. These images have relatively little overlap between them.

2 Image Transformation Model

2.1 Retinal Imaging Instrumentation

Current methods for imaging the retina include the standard fundus microscope, the slit-lamp microscope, and the laser-scanning ophthalmoscope [22]. All of the images for this study were obtained using a TOPCON IMAGENET digital fundus camera (Figure 1(a)). This instrument illuminates the dilated retina in a ring pattern (Figure 1(b)). The illuminated retina is imaged onto a CCD array with a 1024×1024 pixel resolution and 8 bits per pixel. Figure 2 shows three sample images of a healthy human retina. The image on the left also serves to illustrate glare — a common artifact. Asking the patient to fixate on a lighted target (Figure 1(b)), and adjusting the angle of the instrument relative to the eye can enable the photographer to take pictures of the retina from a variety of angles.

2.2 Modeling Assumptions

The retinal surface is almost, but not quite, spherical [22]. A quadric surface model is a good approximation to its shape. We restrict this further to a quadratic surface, as described below, for mathematical simplicity.

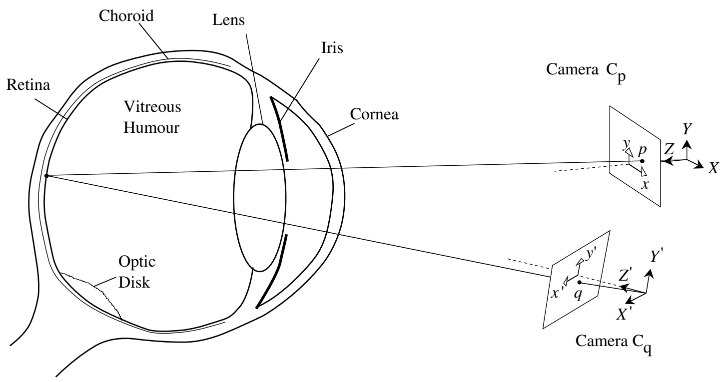


Figure 3: An illustration of the eye and retina, with two different camera viewpoints, C_p and C_q.

Next consider the relative motion between the eye and camera (Figure 3). During imaging, a patient's pupil is dilated and his/her forehead is held against a harness (Figure 1). Small shifts in head position are likely, inducing translations and rotations of the eye. Eye movements themselves, incompletely constrained during diagnosis or surgery, are almost exclusively rotational and occur about two axes at rates of up to 180° per second [39]. Significantly, neither axis of rotation is the optical axis. Finally, except for detached retinas, we may reasonably assume the retina is rigidly attached to the back of the eye. Together, these observations imply that the apparent motion of the retina should be modeled as a general rigid motion. They also imply, however, that some components of the motion — rotation about the camera's optical axis in particular — will be small. This fact will be used in the estimation technique described in Section 4.

The final assumption concerns the image formation model. Cameras used for fundus imaging have long focal lengths, and have planar CCD area sensors. The portion of the retinal surface visible in any image is roughly parallel to the image plane of the camera, even as the eye rotates. This suggests that perspective effects are minor and a weak perspective camera model may be sufficient. On the other hand, the optics of the eye are part of the imaging system since light reflected off the retina passes through the lens of the eye before reaching the camera. This suggests that prior camera calibration may not be possible.

2.3 Transformation Model Derivation

Consider two images of the retinal surface, I_p and I_q, taken from two different viewpoints, or equivalently, by two different cameras, C_p and C_q (see Figure 3). Let $\mathbf{P}^T = (X, Y, Z)$ and $\mathbf{Q}^T = (X', Y', Z')$ denote the same point on the retina surface, but expressed in 3-D coordinate systems attached to C_p and C_q, respectively. The quadratic surface equation, which relates X, Y and Z, is:

$$Z = A_1X^2 + A_2XY + A_3Y^2 + A_4X + A_5Y + A_6, \tag{1}$$

in the coordinate system of C_p. Here, $A_1, \ldots, A_6$ are (unknown) parameters. The assumption of a rigid transformation between viewpoints relates $\mathbf{P}$ and $\mathbf{Q}$ as follows:

$$\mathbf{Q} = \mathbf{RP} + \mathbf{t}, \qquad \text{or,} \qquad \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}, \tag{2}$$

where $\mathbf{R}$ is the orthonormal rotation matrix. Finally, let the weak-perspective camera projection matrices be:

$$\mathbf{M}_p = \begin{pmatrix} \alpha_x & 0 & 0 & c_x \\ 0 & \alpha_y & 0 & c_y \\ 0 & 0 & 0 & s \end{pmatrix}, \quad \text{and} \quad \mathbf{M}_q = \begin{pmatrix} \alpha'_x & 0 & 0 & c'_x \\ 0 & \alpha'_y & 0 & c'_y \\ 0 & 0 & 0 & s' \end{pmatrix}, \tag{3}$$

in the coordinate systems of I_p and I_q respectively. Here α_x, α_y, α'_x, and α'_y are pixel dimension parameters, (c_x, c_y) and (c'_x, c'_y) are the centers of projection (nodal points) in the two cameras, and s and s' are scaling parameters of the weak perspective projection. Based on these equations (and using homogeneous coordinates), the image projections of $\mathbf{P}$ in image I_p and of $\mathbf{Q}$ in image I_q are:[1]

$$\begin{pmatrix} wx \\ wy \\ w \end{pmatrix} \cong \mathbf{M}_p \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} w'x' \\ w'y' \\ w' \end{pmatrix} \cong \mathbf{M}_q \begin{pmatrix} X' \\ Y' \\ Z' \\ 1 \end{pmatrix},$$

which yields:

$$\mathbf{p} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{\alpha_x X + c_x}{s} \\ \frac{\alpha_y Y + c_y}{s} \end{pmatrix}, \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \frac{\alpha'_x X' + c'_x}{s'} \\ \frac{\alpha'_y Y' + c'_y}{s'} \end{pmatrix}. \tag{4}$$

[1] Using homogeneous coordinates, the symbol $\cong$ means equality up to a scale factor.

Our goal, using the foregoing definitions, is to derive the transformation that maps image coordinates $\mathbf{p}$ in I_p to image coordinates $\mathbf{q}$ in I_q. First, inverting the relationship between 2D image coordinates and 3D coordinates in (4) gives

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \frac{sx - c_x}{\alpha_x} \\ \frac{sy - c_y}{\alpha_y} \end{pmatrix}. \tag{5}$$

Substituting these into (1) allows us to write Z in terms of image coordinates x and y as follows:

$$Z = a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6,$$

where the parameters a_i combine the original parameters A_i from equation (1), as well as the camera parameters. (For example, $a_1 = A_1 s^2/\alpha_x^2$ and $a_2 = A_2 s^2/(\alpha_x \alpha_y)$.) Next, substituting the equations for X, Y, and Z into (2) gives expressions for X' and Y':

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} r_{11}\frac{sx - c_x}{\alpha_x} + r_{12}\frac{sy - c_y}{\alpha_y} + r_{13}(a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6) \\ r_{21}\frac{sx - c_x}{\alpha_x} + r_{22}\frac{sy - c_y}{\alpha_y} + r_{23}(a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6) \end{pmatrix}.$$

Substituting these equations into (4) gives an expression for pixel coordinates $(x', y')^T$ in image I_q in terms of the two sets of camera parameters, the rigid transformation parameters, the quadratic surface parameters, and, most importantly, pixel coordinates (x, y) in image I_p. Since all of these parameters are unknown except the image coordinates, we can consolidate the resulting equation as

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} & \theta_{26} \end{pmatrix} (x^2,\ xy,\ y^2,\ x,\ y,\ 1)^T \tag{6}$$

where the matrix entries θ_{ij} are functions of the camera, surface and rigid transformation parameters. It is important to note that we only need to know the θ_{ij} values in order to transform the pixels from one image to the next. These are the parameters estimated by our algorithms. Finally, we may summarize the derived transformation as

$$\mathbf{q} = \begin{pmatrix} \boldsymbol{\Theta}_1^T \\ \boldsymbol{\Theta}_2^T \end{pmatrix} \mathbf{X}(\mathbf{p}) = \boldsymbol{\Theta}\, \mathbf{X}(\mathbf{p}). \tag{7}$$

where $\boldsymbol{\Theta}_i^T = (\Theta_{i1}, \ldots, \Theta_{i6})$, $i = 1, 2$, and $\mathbf{X}(\mathbf{p}) = (x^2, xy, y^2, x, y, 1)^T$.

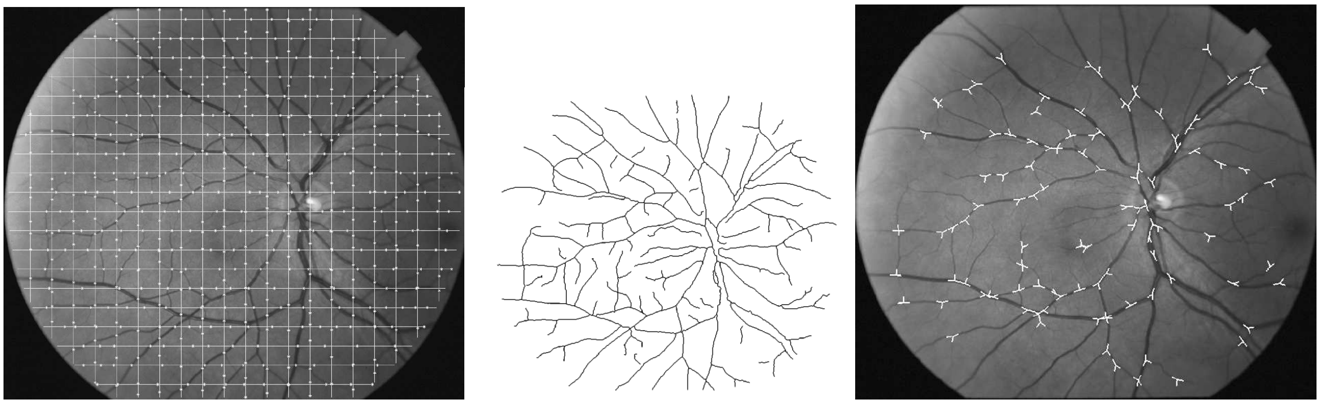
Equation 7 has several important properties. It generalizes the 2D affine transformation model induced by the rigid motion of a planar surface [6], which is given by:

$$\mathbf{q} = \mathbf{A}\mathbf{p} + \mathbf{t}. \tag{8}$$

(See [46, 54] for other recent generalizations.) Observe, also, that (7) is the second-order Taylor series expansion of the general image transformation equation, while the affine model (8) is the first-order expansion.

As analyzed in detail in Section 6.4, registration errors using the 12-parameter model average 0.83 pixels on 1024×1024 images. By contrast, the average registration error using the affine model is 2.2 pixels and the average registration error using translation only is 5.0 pixels. These results are significant since reported prior work [40, 56] has been based on affine or simpler models, tested on image sizes 512×512 or smaller. Matsopoulos et al. [35] studied the use of bilinear and projective transformation models, but could not show improvement over the affine model (see Table 1 in their paper). Current clinical instruments are already producing images of size 1024×1024 or larger, and are scheduled to approach 2500×2500 [30]. Our proposed model is expected to hold up much better with these higher-resolution images.

3 Feature Extraction



(a) (b) (c)

Figure 4: Illustrating the recursive tracing of the vasculature and detection of vascular landmarks: (a) the initial grid, (b) the traced centerlines, and (c) the detected landmarks.

The main problem in registering a pair of images, I_p and I_q, is estimating the twelve parameters of $\mathbf{\Theta}$, which determines the inter-image transformation. In designing an algorithm to solve this problem, we must first choose the image primitives — pixels, regions or features — on which to base the estimation process. For several reasons, a feature-based method is attractive. First, as discussed in the introduction, I_p and I_q may be captured under vastly different illumination conditions, and parts of the retinal surface may undergo changes as well, making direct, image-wide comparison of intensities or even intensity gradients less than ideal. Second, outside the vasculature, the intensity structure is relatively homogeneous. Third, the vasculature and its bifurcations are currently used as spatial landmarks by ophthalmologists and in much of the published literature in this area [4, 35, 56]. Finally, except for detachments, the retina is rigidly attached to the back of the eye. Hence, the branching and cross-over points of the blood vessel structure serve as our features. The remainder of this section summarizes our algorithm to detect these features.

The obvious approach to detecting the vascular landmarks is to apply an edge detector to the input images, pair antiparallel edges to detect the blood vessel boundaries, and then extract landmarks by edge following and corner detection [25, 56, 40]. Indeed, a variation of this approach was used in our prior work [4]. Recently, however, we have published a dramatically faster approach based on the notion of exploratory image analysis [14]. Edge detection principles are applied, but as part of an exploratory procedure that traces out the vascular structure from seed points. The seed points are found using 1-D edge detection along a series of evenly spaced rows and columns. The exploratory procedure traces out the vasculature structure using elongated edge detection filters, recording centerline locations along the way. Landmarks are placed at intersections of traces and at locations where three or more traces meet. These landmarks are characterized by the orientations and widths of the associated blood vessels. Typically, 30-50 landmarks are found in each image. An example result is shown in Figure 4.

Overall, the algorithm concentrates computation only on the relevant subsets of the pixels; upwards of 75% of the pixels are never touched. A version of the algorithm has been implemented to run at frame rates on an SGI workstation [1]. In a recent paper, we have shown how to optimally prioritize the exploratory tracing for real-time performance [47].

4 Estimating the Transformation Parameters and the Feature Correspondences

The next step is to use the vascular landmarks to estimate the transformation parameters for a pair of retinal images. Several challenges must be overcome. First, the overlap between images may be small. Second, vascular landmarks are not uniquely identifiable because branching angles at bifurcations are similar across the retina [37, 38]. Combined, the large inter-image motion between frames and resulting small image overlap, shows there will be many different possible matches for each landmark, yet many landmarks may have no correct matches at all. In turn this means we cannot solve the problem of establishing correspondence between landmarks prior to estimating the transformation; *the correspondence and parameter estimation problems must be solved jointly.* The third challenge is the high dimensionality of the transformation model, which makes exhaustive search impractical. Overall, these observations imply that the matching and transformation estimation problem in registering retinal images requires a more sophisticated algorithm than has been proposed for related problems such as fundamental matrix estimation [53, 57].

4.1 Hierarchical Estimation Algorithm

Successfully estimating the parameters of the high-dimensional transformation $\mathbf{\Theta}$ requires good initialization, especially since correspondences can not be established reliably in advance. The natural approach is to obtain initial estimates from lower-order models. This suggests a hierarchical estimation strategy, such as adopted in much of the closely related literature [6, 29, 32, 43, 44]. To estimate $\mathbf{\Theta}$, we adopt a three-level hierarchy, beginning with a rough estimate of 2-D translations, continuing to an affine estimate, and ending with the quadratic transformation, $\mathbf{\Theta}$. At each level, a different robust estimation technique is applied and the set of possible correspondences is culled. Use of the initial translation estimate assumes approximately constant scale and small rotations. Scale changes can be kept small by ophthalmic photographers. Rotation in the image plane is generally small because the patient is constrained in a harness, eye rotations about the optical axis are small, and eye rotations about the other axes mimic translation except over broad image regions. The affine model is appropriate because the retinal surface is roughly planar over small regions [4]. This explains why the affine transformation model yields modest errors and has been used in much of the related work [35, 40, 56].

4.2 0^{th} Order, Translation

A weighted histogram technique is used to estimate the translation vector, $\mathbf{t}_0$, representing the 0^{th}-order transformation:

$$T_0(\mathbf{p}; \mathbf{t}_0) = \mathbf{p} + \mathbf{t}_0. \tag{9}$$

Let the vascular landmark sets from the two images be denoted $\mathbf{P}$ and $\mathbf{Q}$, with $N_p = |\mathbf{P}|$ and $N_q = |\mathbf{Q}|$ points respectively. The algorithm forms a two-dimensional histogram of translation vectors with relatively large bin sizes — large enough to accommodate both feature position errors and the modeling error induced by using a 0^{th} order model. An initial correspondence set, $\mathbf{C}_0 = P \times Q$, is established containing all possible correspondences. For each $(\mathbf{p}_i, \mathbf{q}_j) \in \mathbf{C}_0$, both $\mathbf{t}_{i,j} = \mathbf{q}_j - \mathbf{p}_i$ and a similarity measure $s_{i,j}$, described below, are calculated. Then, $s_{i,j}$ is entered in the histogram bin containing $\mathbf{t}_{i,j}$. After doing this for all correspondences in $\mathbf{C}_0$, the histogram is smoothed and the peak is detected (Figure 5). Finally, a new correspondence set $\mathbf{C}_1$ is formed that contains all matches $(\mathbf{p}_i, \mathbf{q}_j)$ such that $\mathbf{t}_{i,j}$ is within twice the histogram bin width of $\hat{\mathbf{t}}_0$. Some landmarks will have several correspondences in $\mathbf{C}_1$, while others will have none.

The similarity measure, $s_{i,j}$, for a possible correspondence $(\mathbf{p}_i, \mathbf{q}_j)$, is computed in two steps. First, a raw measure is computed based on the orientations of the vascular centerlines that meet at the intersection points. For landmark location $\mathbf{p}_i$ from image I_p, these direction vectors are denoted by $\hat{\mathbf{u}}_{i,1}, \ldots, \hat{\mathbf{u}}_{i,n_i}$, and for landmark location $\mathbf{q}_j$ from image I_q, these are denoted by $\hat{\mathbf{v}}_{j,1}, \ldots, \hat{\mathbf{v}}_{j,n_j}$. Importantly, n_i and n_j, the number of direction vectors for landmarks $\mathbf{p}_i$ and $\mathbf{q}_j$ respectively, may be different, even for correctly matching landmarks, because the tracing procedure cannot be guaranteed to detect all vessels. Define $\mathbf{\Gamma}$ to be an $n_i \times n_j$ binary matrix determining correspondences between centerline trace directions for $\mathbf{p}_i$ and $\mathbf{q}_j$. $\mathbf{\Gamma}$ is constrained so that each row and each column has at most one non-zero entry. Then, the raw similarity measure $s'_{i,j}$ is defined as a maximum over all correspondence matrices, as follows:

$$s'_{i,j} = \frac{1}{2\min(n_i, n_j)} \max_{\mathbf{\Gamma}} \sum_{k,l} \mathbf{\Gamma}(k,l)(\hat{\mathbf{u}}_{i,k} \cdot \hat{\mathbf{v}}_{j,l} + 1). \tag{10}$$

It is straightforward to show that $0 \leq s'_{i,j} \leq 1$.

The second step to calculating the similarity measure is to convert $s'_{i,j}$ to a prior probability that $(\mathbf{p}_i, \mathbf{q}_j)$ is correct. Experimental analysis shows that $s'_{i,j}$ must be almost exactly 1.0 for the correspondence to be correct, and that the probability that $(\mathbf{p}_i, \mathbf{q}_j)$ is correct given the raw similarity measure, $s'_{i,j}$, is well-approximated by a power transform. Thus, the final similar measure is

$$s_{i,j} = {s'_{i,j}}^{\alpha},$$

with α empirically set at 100. For example, $s'_{i,j} = 0.99$ converts to $s_{i,j} = 0.37$. See [13] for details.

The above similarity measure is translation and scale invariant. Zana and Klein [56] proposed a similar measure, based on minimizing the angular displacement between landmarks. Our measure improves upon this by handling bifurcations as well as trifurcations, by not requiring explicit calculation of angles, and

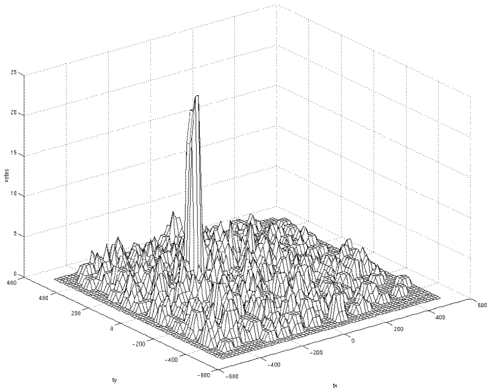


Figure 5: Sample plot of the 2D similarity-weighted histogram for an image pair.

especially, by using the power transform to obtain an estimate of the probability the match is correct given the raw similarity measure.

4.3 1st-Order, Affine Model

The next level of the hierarchy estimates an affine transformation,

$$T_1(\mathbf{p}; \mathbf{A}, \mathbf{t}_1) = \mathbf{A}\mathbf{p} + \mathbf{t}_1,$$

starting from the reduced correspondence set $\mathbf{C}_1$ determined by the similarity-weighted histogram technique. The technique used to estimate $\mathbf{A}$ and $\mathbf{t}_1$ is a slightly modified least-median of squares algorithm (LMS) [41, 50].

Let $\mathbf{P}_1 \subseteq \mathbf{P}$ contain the features from I_p having at least one match in $\mathbf{C}_1$, and for each $\mathbf{p} \in \mathbf{P}_1$ let $\mathbf{C}_1(\mathbf{p}) = \{\mathbf{q} \mid (\mathbf{p}, \mathbf{q}) \in \mathbf{C}_1\}$. Note that if the overlap between images is small, $\mathbf{P}_1$ will be much smaller than $\mathbf{P}$. The LMS estimate of the affine parameters is

$$(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1) = \operatorname*{argmin}_{\mathbf{A}, \mathbf{t}_1} \operatorname*{median}_{\mathbf{p} \in \mathbf{P}_1} \min_{\mathbf{q} \in \mathbf{C}_1(\mathbf{p})} \|\mathbf{q} - \mathbf{A}\mathbf{p} - \mathbf{t}_1\|^2. \tag{11}$$

In other words, the objective function for a given $(\mathbf{A}, \mathbf{t}_1)$ is calculated by finding the minimum distance match for each transformed $\mathbf{p} \in \mathbf{P}_1$ and then taking the median of the resulting squared (error) distances. This differs from other well-known uses of LMS — e.g. in estimating the fundamental matrix [53, 57] — because uniqueness in the correspondence set is not (and should not be) enforced at this level of the computation.

The objective function in (11) is not differentiable, partially due to use of the median and partially due to the selection process in choosing the best correspondence for each $(\mathbf{A}, \mathbf{t}_1)$. As a result, we use a random sampling search technique [23, 41, 50] to find the approximate minimum of the objective function. The method differs only slightly from the techniques proposed in the original LMS [41] and Random Sample Consensus (RANSAC) [23] papers. Triples of landmark locations, $\mathbf{p}_i$, $\mathbf{p}_j$, $\mathbf{p}_k$, are chosen at random from $\mathbf{P}_1$. For each such triple, each possible combination of matches, one from $\mathbf{C}_1(\mathbf{p}_i)$, one from $\mathbf{C}_1(\mathbf{p}_j)$, and one from $\mathbf{C}_1(\mathbf{p}_k)$, is used to calculate a unique $\mathbf{A}$ and $\hat{\mathbf{t}}_1$ parameter set. The objective function in (11) is analyzed for each set. The set minimizing the objective function is retained. This process is repeated for S triples, and the parameter set minimizing the objective function over all triples becomes the affine estimate $(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1)$. S, the number of sample triples required to ensure with high probability that at least one triple contains landmarks each having a correct correspondence in $\mathbf{C}_1$, is determined using standard methods (see [23, 41, 50]).

After minimization, a robust scale value, $\hat{\sigma}_1$, is calculated from $(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1)$ using the median absolute deviation (MAD) scale estimator [42]:

$$\hat{\sigma}_1 = 1.4826\left[1 + \frac{5}{|P_1| - 3}\right] \sqrt{\operatorname*{median}_{\mathbf{p} \in \mathbf{P}_1} \min_{\mathbf{q} \in \mathbf{C}_1(\mathbf{p})} \|\mathbf{q} - \hat{\mathbf{A}}\mathbf{p} - \hat{\mathbf{t}}_1\|^2}. \tag{12}$$

The factor $1 + 5/(|P_1| - 3)$ is an experimentally determined finite-sample correction factor proposed by Rousseeuw and Leroy [42] to avoid artificially low values of $\hat{\sigma}_1$ caused by small point sets. The constant 1.4826 ensures that $\hat{\sigma}_1$ is "consistent with a normal distribution" — in other words, if the errors $\|\mathbf{q} - \hat{\mathbf{A}}\mathbf{p} - \hat{\mathbf{t}}_1\|$ were independent and normally distributed then $\hat{\sigma}_1^2$ would be an unbiased estimate of the variance of this distribution [50].

The final step in LMS and other "high-breakdown" robust estimation procedures — i.e., those relatively unaffected by large numbers of outliers — is usually to refine the estimate by gathering inliers and calculating a least-squares estimate. This step is left out here because our next step is to switch to estimation of the full quadratic transformation parameters.

Before discussing this, however, it is important to consider why other robust procedures are not used in estimating $(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1)$. The obvious alternative to LMS is a RANSAC method [23]. This would estimate $\hat{\mathbf{A}}$ and $\hat{\mathbf{t}}_1$ by maximizing the number of landmarks having a correspondence whose error is less than a given threshold D_1. Stated explicitly, the estimate would be

$$(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1) = \operatorname*{argmax}_{\mathbf{A}, \mathbf{t}_1} \ | \{\mathbf{p} \in \mathbf{P}_1 \ : \ \exists \mathbf{q} \in \mathbf{C}_1(\mathbf{p}), \ \|\mathbf{q} - \mathbf{A}\mathbf{p} - \mathbf{t}_1\| < D_1\} \ | \ . \tag{13}$$

The problem is establishing the distance threshold, D_1. Clearly it must depend on errors in the estimate of landmark positions. More importantly, it must depend on the modeling error introduced by using an affine model. This error is difficult to predict, however, because it varies with the amount of overlap between images as well as the positions of the landmarks. The LMS objective function is more reliable because the landmark set $\mathbf{P}_1$, which is calculated from the initial translation estimate, generally includes landmarks taken from near or within the region of overlap and as a result at least 50% of the landmarks in $\mathbf{P}_1$ have correct matches. If this fraction became a concern we would switch to generalizations of LMS such as MUSE [36] that adaptively choose the inlier fraction, including fractions well-below 50%.

4.4 2^{nd} Order, Quadratic Model

The final level of the hierarchy estimates the quadratic transformation,

$$T_2(\mathbf{p}; \mathbf{\Theta}) = \mathbf{\Theta}\,\mathbf{X}(\mathbf{p}),$$

(see Equation 7 and discussion) using an M-estimator [27, 50]. We describe first a straightforward instantiation of the M-estimator and then show several important modifications.

For each $\mathbf{p_i} \in \mathbf{P}_1$, let

$$\mathbf{q}_i = \operatorname*{argmin}_{\mathbf{q} \in C_1(\mathbf{p_i})} \|\mathbf{q} - \hat{\mathbf{A}}\mathbf{p}_i - \hat{\mathbf{t}}_1\|^2.$$

In other words, $\mathbf{q}_i$ is the best match for $\mathbf{p}_i$ based on the estimated affine transformation. Then the M-estimate of $\mathbf{\Theta}$ is

$$\hat{\mathbf{\Theta}} = \operatorname*{argmin}_{\mathbf{\Theta}} \sum_{\mathbf{p}_i \in \mathbf{P}_1} \rho(\|\mathbf{q}_i - \mathbf{\Theta}\,\mathbf{X}(\mathbf{p}_i)\|/\hat{\sigma}), \tag{14}$$

where ρ is a "robust loss function" which grows subquadratically, and $\hat{\sigma}$ is a scale estimate. Here we use the Beaton-Tukey biweight function [2] (Figure 6):

$$\rho(u) = \begin{cases} \frac{a^2}{6}\left[1 - \left(1 - \left(\frac{u}{a}\right)^2\right)^3\right] & |u| \le a \\ \frac{a^2}{6} & |u| > a \end{cases}$$

where $u = \|\mathbf{q} - \mathbf{\Theta}\,\mathbf{X}(\mathbf{p})\|/\hat{\sigma}$ is a "scale normalized residual." (Typically $a \approx 4.0$ [28].) Equation 14 may be solved using iteratively-reweighted least-squares (IRLS) [28], with weight function $w(u) = \rho'(u)/u$. For

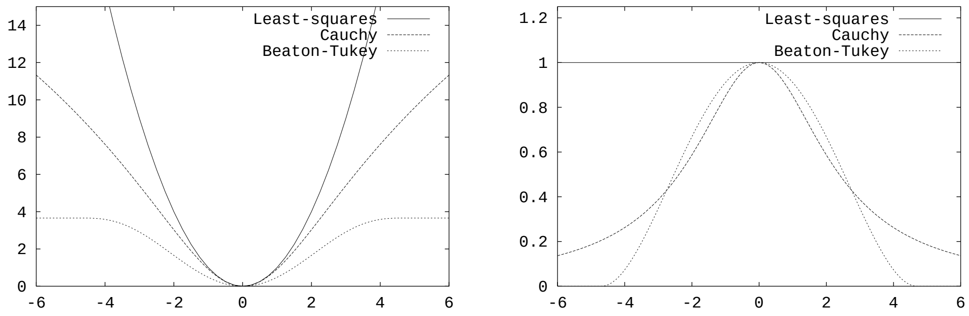


Figure 6: Plots of the robust loss function ρ (a) and weight function (b) for the Beaton-Tukey biweight, the Cauchy loss function and the quadratic error function used in least-squares. The Beaton-Tukey is chosen because it most aggressively rejects outliers.

the Beaton-Tukey biweight (Figure 6)

$$w(u) = \begin{cases} \left[1 - \left(\frac{u}{a}\right)^2\right]^2 & |u| \leq a \\ 0 & |u| > a. \end{cases}$$

The value u_i for each match in each iteration is calculated using the estimate of $\mathbf{\Theta}$ obtained in the previous iteration. A robust starting point for IRLS is crucial, as has been confirmed experimentally in vision applications [53].

Several modifications of the estimation equation (14) and IRLS search technique are important here:

1. The M-estimator is initialized from the affine estimate $(\hat{\mathbf{A}}, \hat{\mathbf{t}}_1)$ and scale $\hat{\sigma}_1$. These are used to compute initial scale normalized residuals

$$u_i = \|\mathbf{q}_i - \hat{\mathbf{A}}\mathbf{p}_i - \hat{\mathbf{t}}_1\| / \hat{\sigma}_1,$$

and these are used to compute initial weights to be used in IRLS.

2. A new (MAD) scale, $\hat{\sigma}_2$, is estimated from the residuals after each of the first few iterations of the IRLS procedure and then fixed for the remaining iterations.

3. Because $w(u) = 0$ for residuals (error distances) greater than about $4\hat{\sigma}_2$, there is no need to restrict the match set in (14). The entire original set $\mathbf{C}_0$ may be used, since matches with large error distances simply contribute zero weight. (In practice, loose restrictions are placed purely for computational reasons.) This allows recovery from earlier mistakes in reducing the match set.

4. The robust weights $w(u)$ are augmented in IRLS by the correspondence similarity measure (10) and, when more than one match for a given $\mathbf{p} \in \mathbf{P}$ has non-zero weight, these weights are normalized. For example, if $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}, \mathbf{q}')$ are two matches for $\mathbf{p}$, with robust weights w and w' and similarity measures s and s', then the actual IRLS weights will be:

$$w^* = s\,w \frac{s\,w}{s\,w + s'\,w'}, \qquad \text{and} \qquad w'^* = s'\,w' \frac{s'\,w'}{s\,w + s'\,w'}.$$

In effect, the 3^{rd} *and* 4^{th} *modifications allow decisions about correspondence to be deferred until the final transformation is estimated. This differs significantly from other robust matching and estimation algorithms [53, 57], and is an important reason why the algorithm is successful.* Correct correspondences are determined to be those having non-zero weights. More than one such correspondence for any $\mathbf{p}$ indicates a lingering ambiguity, which is usually caused by merging of blood vessel intersections into a single feature in one image but not in the other. This problem, together with uncertainties in the positions of the landmark features themselves, necessitates the final step:

5. After the estimate converges, the landmark positions are refined using the estimated transformation and then the correspondence set is increased by searching for landmarks detected in one image but missing in the other. These two steps are applied in sequence and the M-estimator is rerun to convergence after each. The methods used are detailed in the next section.

Convergence of the M-estimator generally requires only a small number of iterations each time: 5-10 initially, and 2-3 after refinement and increasing the correspondence set.

5 Methods for Refining Feature Positions and Enriching the Correspondence Set

Accurate pairwise registration requires repeatability in detected feature (landmark) locations (see discussion in [45]), especially when the overlap between images is small. Unfortunately, differences in illumination and viewpoint, as well as noise, can cause the same landmark in two images to vary in its detected location (after registration) by several pixels. The same problems can also cause a landmark to be detected in one image but missed in another. This results in a reduced correspondence set.

Both problems can be addressed as the pairwise registration converges. The transformation estimate can be used as the basis for refining landmark locations and for detecting landmarks [49] previously missed.

As already discussed, this can be used to refine the transformation estimate.

5.1 Feature Position Refinement

Suppose $\mathbf{p}_i$ and $\mathbf{q}_j$ form a landmark feature correspondence and let the estimated transformation parameters be $\hat{\mathbf{\Theta}}$. Let the transformed position of $\mathbf{p}_i$ from image I_p to image I_q be $\mathbf{q}_i^* = \hat{\mathbf{\Theta}}\mathbf{X}(\mathbf{p}_i)$. Differences between $\mathbf{q}_i^*$ and $\mathbf{q}_j$ are caused by three factors: modeling error, estimation error in $\hat{\mathbf{\Theta}}$, and discrepancies in the detected landmark positions $\mathbf{p}_i$ and $\mathbf{q}_j$. We would like to reduce the latter as much as possible. The idea is to transform a region surrounding $\mathbf{p}_i$ from I_p to I_q using $\hat{\mathbf{\Theta}}$, and locally match the transformed region to I_q by minimizing a sum-of-squared-differences (SSD) measure. (This region generally includes the current location for $\mathbf{q}_j$.) The SSD match becomes the new location for $\mathbf{q}_j$. Local intensity normalization in each image is used to partially correct for illumination differences.

The most important question is how to map the intensity region R_i from image I_p to image I_q. The simplest method is to ignore the warping induced by the quadratic transformation and simply translate the region between images — specifically by the vector $\hat{\mathbf{\Theta}}\mathbf{X}(\mathbf{p}_i) - \mathbf{p}_i$. We can do substantially better using a centered affine approximation to the quadratic transformation. This gives an invertible transformation that simplifies the mapping process while still accounting for the distortion induced by the higher-order transformation.

The centering process is straightforward. First, decompose $\hat{\mathbf{\Theta}}$ into

$$\hat{\mathbf{\Theta}} = \big(\mathbf{B}|\mathbf{A}|\mathbf{t}\big),$$

where $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{t}$ are 2×3, 2×2 and 2×1, respectively, and give the 2^{nd}, 1^{st} and 0^{th} order terms of the transformation (Equation 7). Also, let $\mathbf{p}_i = (x_i, y_i)^T$. For any point $\mathbf{p} = (x, y)$ near $\mathbf{p}_i$, let $\mathbf{p}' = \mathbf{p} - \mathbf{p}_i = (x', y')^T$ be the re-centered point. Then, after some manipulation

$$\begin{aligned}
\hat{\mathbf{\Theta}}\mathbf{X}(\mathbf{p}) &= \hat{\mathbf{\Theta}}\mathbf{X}(\mathbf{p}_i + \mathbf{p}'); \\
&= \hat{\mathbf{\Theta}}\mathbf{X}(\mathbf{p}_i) + \left[\mathbf{A} + \mathbf{B}\begin{pmatrix} 2x_i & 0 \\ y_i & x_i \\ 0 & 2y_i \end{pmatrix}\right]\begin{pmatrix} x' \\ y' \end{pmatrix} + \mathbf{B}\begin{pmatrix} (x')^2 \\ x'y' \\ (y')^2 \end{pmatrix}; \\
&\approx \mathbf{q}_i^* + \mathbf{A}'\mathbf{p}'.
\end{aligned}$$

where

$$\mathbf{A}' = \mathbf{A} + \mathbf{B}\begin{pmatrix} 2x_i & 0 \\ y_i & x_i \\ 0 & 2y_i \end{pmatrix}.$$

This affine approximation is accurate because $\|\mathbf{p}'\|$ is small. The inverse transformation for $\mathbf{q}$ near $\mathbf{q}_i^*$ is $\mathbf{A}'^{-1}(\mathbf{q} - \mathbf{q}_i^*) + \mathbf{p}_i$.

Using this affine approximation and its inverse, the details of estimating a refined feature position in I_q are as follows. A square $W \times W$ window, denoted by R_q, centered at $\mathbf{q}_i^*$ is defined in image I_q. W is locally set to twice the pixel width of the vessels intersecting to form the landmark in I_p. For each pixel location $\mathbf{q}$ in R_q, $\mathbf{q}$ is inverse mapped into I_p, and bilinear interpolation is used to estimate the intensity from the surrounding pixels. This becomes the intensity at $\mathbf{q}$ in R_q. Once all pixel locations have been mapped, the intensity values in R_q are centered and normalized. Then, R_q is shifted over a $2U \times 2U$ region centered at $\mathbf{q}_i^*$ to minimize the normalized SSD. (U is set to 3 times the error standard deviation, $\hat{\sigma}_2$, of the current transformation estimate.) Parabolic interpolation gives subpixel position accuracy in both dimensions. This becomes the new location for $\mathbf{q}_j$.

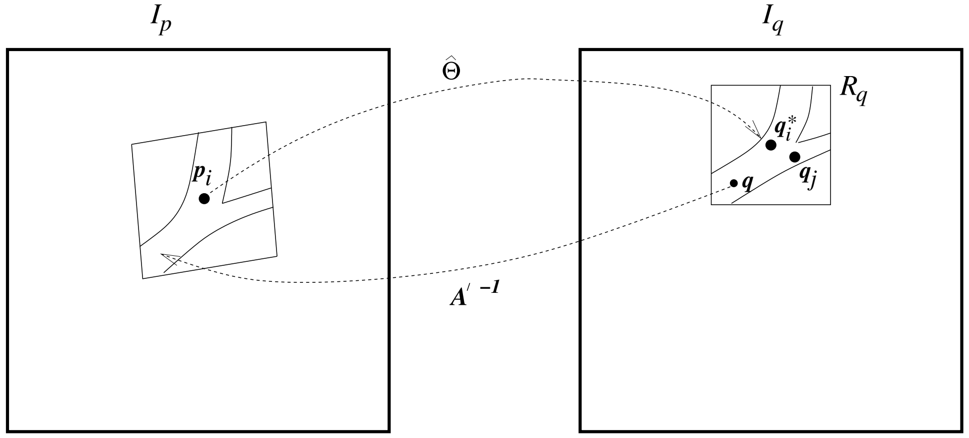


Figure 7: Use of the affine approximation and its inverse in feature position refinement. Each pixel $\mathbf{q}$ in window R_q of image I_q is inverse mapped into I_p using the inverse of affine approximation, $\mathbf{A}'^{-1}$, to obtain a pixel value. Since the inverse mapping will generally not place $\mathbf{q}$ on a discrete location in I_p, bilinear interpolation is used to calculate the value.

5.2 Adding Correspondences for Unmatched Landmarks

Following a re-estimate of $\hat{\mathbf{\Theta}}$ based on the refined landmark positions, correspondences are added for unmatched landmarks that fall into the region of overlap between transformed images. The detection procedure for unmatched landmarks in I_p is exactly the same as the refinement procedure. For unmatched

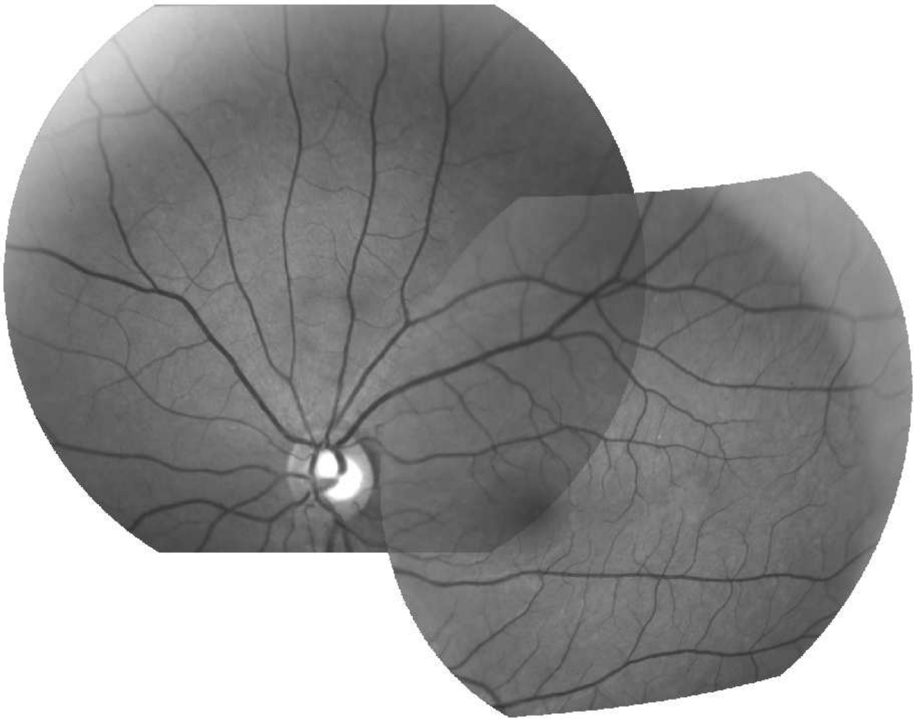


Figure 8: Successful registration for extremely low overlap between images — 22% overlap. (See point 'A' in Figure 12.) The measured centerline error is 1.01 pixel.

feature locations in I_q, the first step is to calculate a quadratic transformation estimate $\hat{\mathbf{\Theta}}'$ mapping I_q to I_p by reversing the roles of the correspondences. Then, feature matching proceeds as just described (except with reversal of the roles of I_p and I_q). The resulting new matches are added to the correspondence set and then the final estimate, $\hat{\mathbf{\Theta}}$, is calculated using the M-estimator technique.

6 Results

The experiments are driven by several goals. First and foremost, the effectiveness of the complete system must be determined. Second, the significance of each step of the algorithm must be analyzed. Third, the quadratic transformation must be validated.

6.1 Data Sets and Metrics

Ten to twenty 1024×1024 images were acquired from each of 16 different eyes. Pairs of images within each data set were chosen to cover a range of possible overlaps and fundus regions — 3000 image pairs in total.

Figure 9: Successful registration for an image pair having 38% overlap. The measured centerline error is 1.5 pixels. (See point 'B' in Figure 12.)

Registration error for any pair of images is determined from the alignment of the vasculature following application of the estimated transformation. Recall that the feature extraction method [14] summarized in Section 3 extracts both the landmarks — branching and cross-over points — and the centerline pixels of the blood vessels. The landmarks are used for matching, the centerlines for validation. Let image I_p be mapped onto I_q using estimated transformation parameters $\hat{\mathbf{\Theta}}$. To calculate the registration error, sample the centerline pixels $\mathbf{p}_c$ in I_p, and map them into I_q using $\hat{\mathbf{\Theta}}$. For each mapped centerline location that falls within the retina image area of I_q, find the distance to the nearest centerline point in I_q. This distance is easily calculated using a digital distance map [11]. The median of these distances is taken as the centerline mapping error. Median statistics are used to discard error measures from missing or spurious centerlines.

The other quantity used in evaluating the registration error is the percentage of overlap between images. For I_p and I_q, this is defined as the fraction of I_p appearing in I_q. Determining this requires knowing the inter-image transformation. For successful registrations, we can visually judge the rough correctness

Figure 10: Successful registration for an image pair having 50% overlap. The measured centerline error is 0.92 pixels. (See point 'C' in Figure 12.)

of the results and therefore use the estimated transformation to calculate overlap. For unsuccessful registrations, we resort to using the joint registration algorithm described in the companion paper [16], which simultaneously combines all images from each eye. When this joint registration succeeds, which we can also judge visually, we use the resulting inter-image transformation from I_p to I_q to calculate overlap.

6.2 Performance of the Complete Algorithm

The performance of the complete pairwise registration algorithm can be judged both visually and quantitatively. Figures 8 through 11 show several pairwise image mosaics formed using successful registrations. A range of overlaps between image pairs is shown. Non-linear distortions caused by the quadratic model are illustrated in Figures 8 and 9 where the overlap is low.

Quantitative analysis requires more care. The primary issue is determining a criterion for registration failure. Two measures are used. The first is the number of correspondences: if the number of landmarks features in I_p that have possible correspondences in I_q ever drops below 6 — the minimum number needed to estimate the quadratic transform parameters — registration immediately fails. These failures are detected at the end of the 0^{th} and 1^{st} levels of the estimation hierarchy. When the estimation process

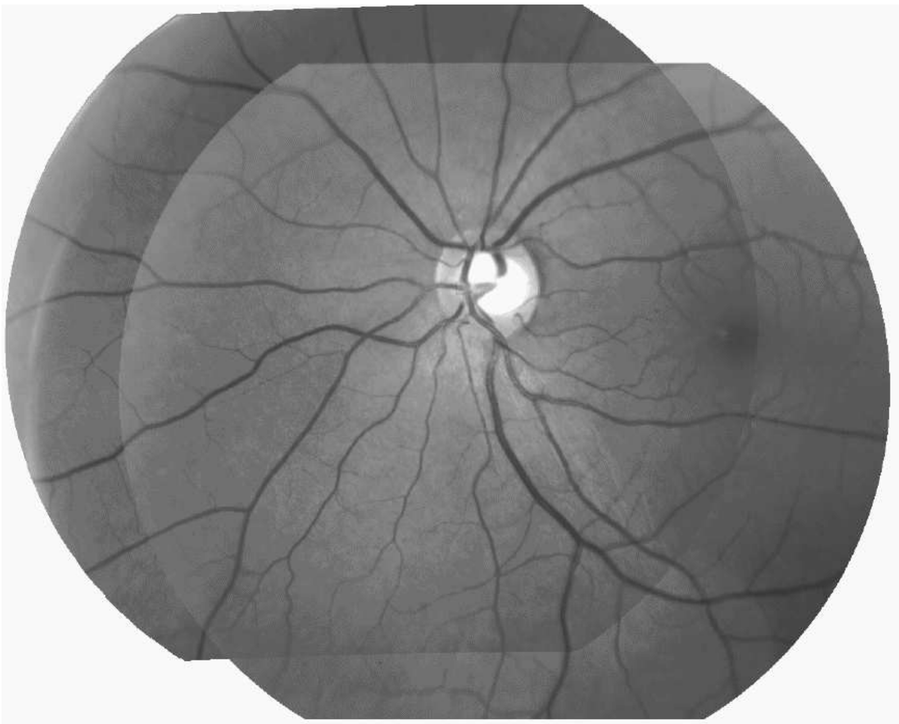


Figure 11: Successful registration for an image pair having 75% overlap. The measured centerline error is 0.76 pixels. (See point 'D' in Figure 12.)

reaches the 2^{nd} level, an estimate, $\hat{\mathbf{\Theta}}$, will always be obtained. The second measure, therefore, must be based on the magnitude of the median centerline error. The question then becomes establishing an error threshold.

We do so using a plot of the centerline error measure against the percentage of overlap between image pairs, as shown in Figure 12. Negative values on this graph show failures due an insufficient number of correspondences. Positive values show the centerline error. Clearly there is a tight cluster of errors around 1.0 pixel. We establish an error threshold by manually judging the success or failure of results at or above 1.0 pixels. Two potential choices of thresholds are of particular interest. First, no incorrect registrations have an error measure below 1.5. If the threshold is set here — which means we minimize "false positives" — 18% of the registrations rejected are in fact correct. Second, no correct registrations have an error measure above 2.3. If the threshold is set here, meaning we minimize "false negatives", 1.2% of the registrations accepted would be incorrect (false positives). In the application, false positives are highly undesirable. Hence, we fix the error threshold at 1.5 pixels for the remainder of the discussion.

Given the two rejection criteria — the minimum number of correspondences and the threshold on the median centerline error — we can evaluate the performance of the algorithm as a function of the

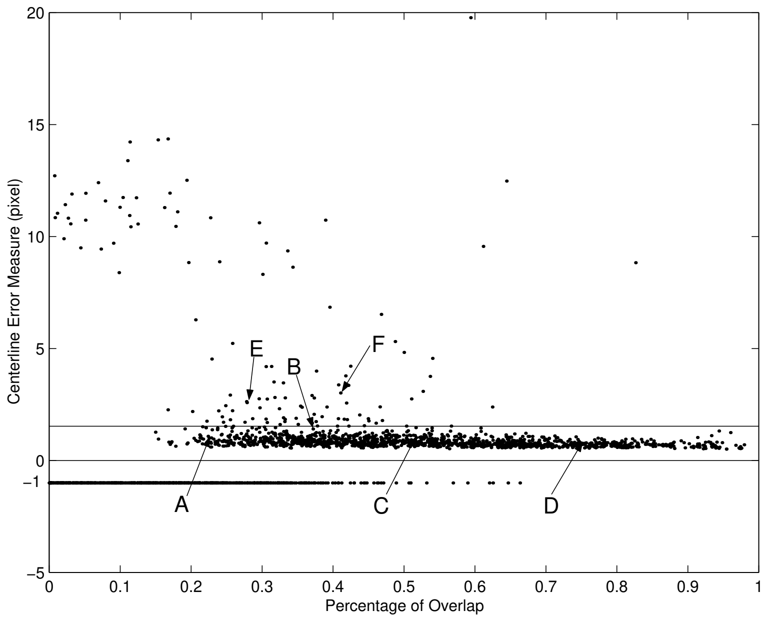


Figure 12: Percentage of overlap versus centerline error measure. The negative values on the graph indicate failures due to insufficient correspondences. The horizontal line at 1.5 pixels is the threshold to reject an estimated transformation. The letters A, B, C, D, E, F on the graph show the percentage of overlap and the error measures of the registered images shown in Figures 8, 9, 10, 11, 14, and 15, respectively.

percentage overlap and then use this to pinpoint causes of the registration failures. Figure 13 shows a plot of the percentage of registration failures as a function of the percentage of overlap. Clearly, an insufficient number of correspondences is the primary cause of failures, especially for overlap percentages below 40. Registration for small overlap percentages only succeeds when the overlap region is sufficiently rich in landmarks. These cases are illustrated in Figures 8 and 9.

Less frequently, registration fails despite having enough correspondences. There are two reasons for such failures. The first, illustrated in Figure 14, occurs when the landmarks are clustered in a small image region, making the estimated transformation insufficiently stable for accurate registration across the entire image. The second, illustrated in Figure 15, occurs when the overlap region contains nearly repeated structures, and estimation at the 0^{th} and 1^{st} order levels of the hierarchy initializes the quadratic transformation M-estimator in the wrong domain of convergence.

6.3 Importance of Each Stage in the Hierarchical Estimation

Now that the successes and failures of the overall algorithm have been established, it is important to consider the contributions of each stage of the algorithm. The plots in Figure 16 compare the centerline errors

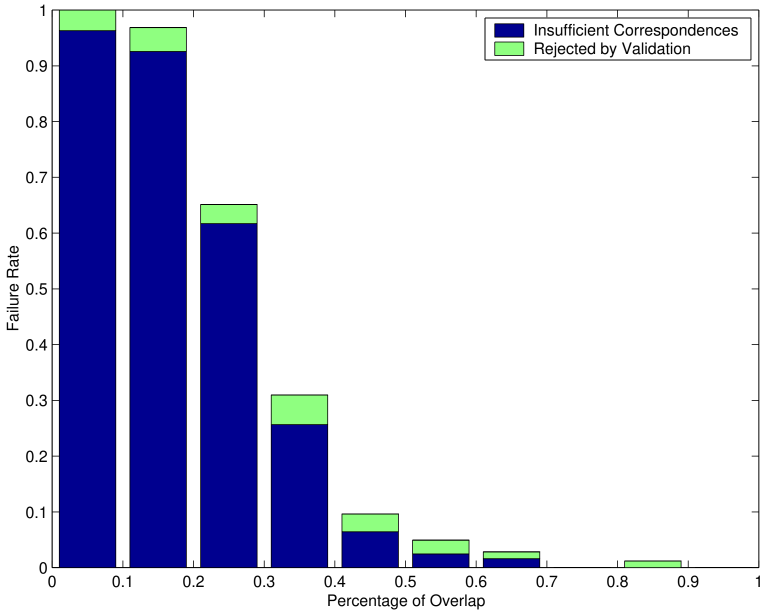


Figure 13: Percentage of failures as a function of the overlap percentage. Failures due to an insufficient number of correspondences and due to above-threshold centerline errors are both shown. The estimation algorithm nearly always fails when the overlap is less than 20%. With increasing overlapping percentage the failure rate decreases exponentially, and above 67% of overlap the algorithm only failed for one out of 330 image pairs.

resulting from the transformation estimated at each successive stage of the algorithm, starting from the translation estimate at the 0^{th} level of the hierarchy, and moving all the way through the landmark enrichment step. Plots (a) and (b) show that, as expected, increasing the order of the transformation increases the accuracy of the estimate. Plots (c) and (d) show the importance of the SSD-based landmark position refinement and the landmark enrichment. Overall, the average error after the translation estimation (0^{th} level) is 5.0 pixels, after the affine estimate (1^{st} level) is 3.0 pixels, after the initial quadratic M-estimate (2^{nd} level) is 1.56 pixels, after landmark position refinement is 0.97 pixels, and after landmark correspondence enrichment is 0.83 pixels. The improvements obtained using the SSD are especially striking since the order of the transformation remains the same. We conclude, therefore, that the locally-centered, affine SSD refinement is an important component of the algorithm, both to increase the accuracy of landmark positions and to increase the number of correspondences.

6.4 Validation of the Quadratic Image Transformation Model

The final part of the analysis is to validate the quadratic transformation model derived in Section 2. Important visual evidence has already been given in the example pairwise mosaics of Figures 8 through 11.

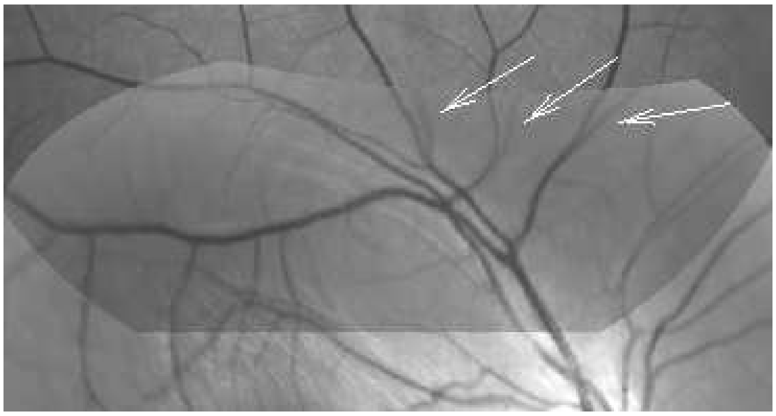


Figure 14: Overlapping region of incorrectly registered images where spatially localized landmarks caused the algorithm to fail (point 'E' in Figure 12.) Several of the "ghost vessels" due to misalignment are indicated with arrows. The overlap is 28% and the centerline error is 2.63 pixels.

Quantitative evidence has already been shown in the numerical accuracy of the final registrations — 0.83 pixels after landmark enrichment. More complete analysis may be obtained by estimating the parameters of the best fitting (least-squares) translational (0^{th} order), affine (1^{st} order) and quadratic (2^{nd} order) transformation models using the final correspondence sets for each successful registration. The final correspondence set is preferred because the landmark position refinement and enhancement substantially reduces noise and estimation error.

The errors are plotted in two ways in Figure 17. Plot (a) is a scatter plot of the median centerline error for each model as a function of the overlap percentage. The average centerline error for the translational, affine and quadratic models are 4.88, 2.47, and 0.83 pixels respectively. The error measures for each model that fall into the same 10% overlapping range are averaged and shown in plot(b). Notice that the error measure for the quadratic model decreases only slightly with the amount of overlap. In fact, this may not even be modeling error: it could due to increased stability of the estimate caused by an increase in the number of landmarks. By contrast, the translational and affine errors increase slightly with increasing overlap, reach a peak at 40-50% overlap, and then decrease steadily.[2] This gives strong evidence that the quadratic model closely approximates the true image-to-image transformation. We conclude that the quadratic model is the correct image-to-image transformation model to use.

[2]The apparently counter-intuitive result that error increases slightly with increasing overlaps (below 50%) can be understood as follows. The transformation model need only "explain" differences between regions that appear in both images. When the overlap is small, the region is small, and therefore lower-order models produce lower error. As the overlap increases higher-order models are needed to account for viewpoint changes and retinal surface curvature, but as the overlap approaches 100%, the correct transformation approaches the identity transformation, meaning that lower-order models are sufficient again.

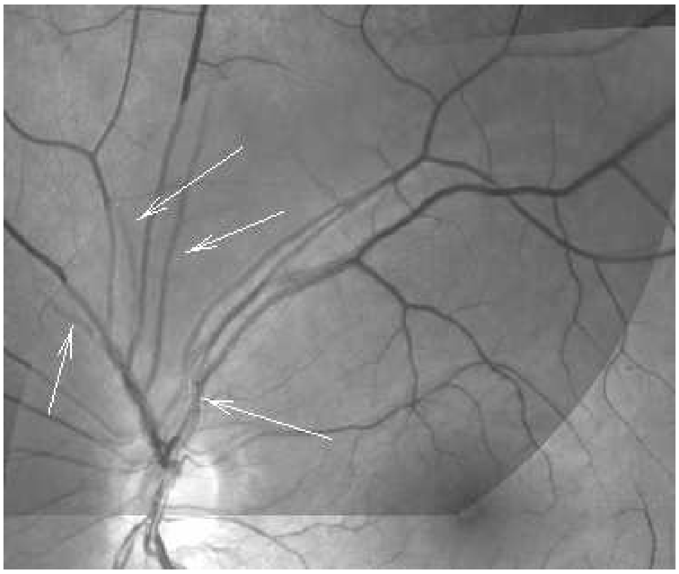
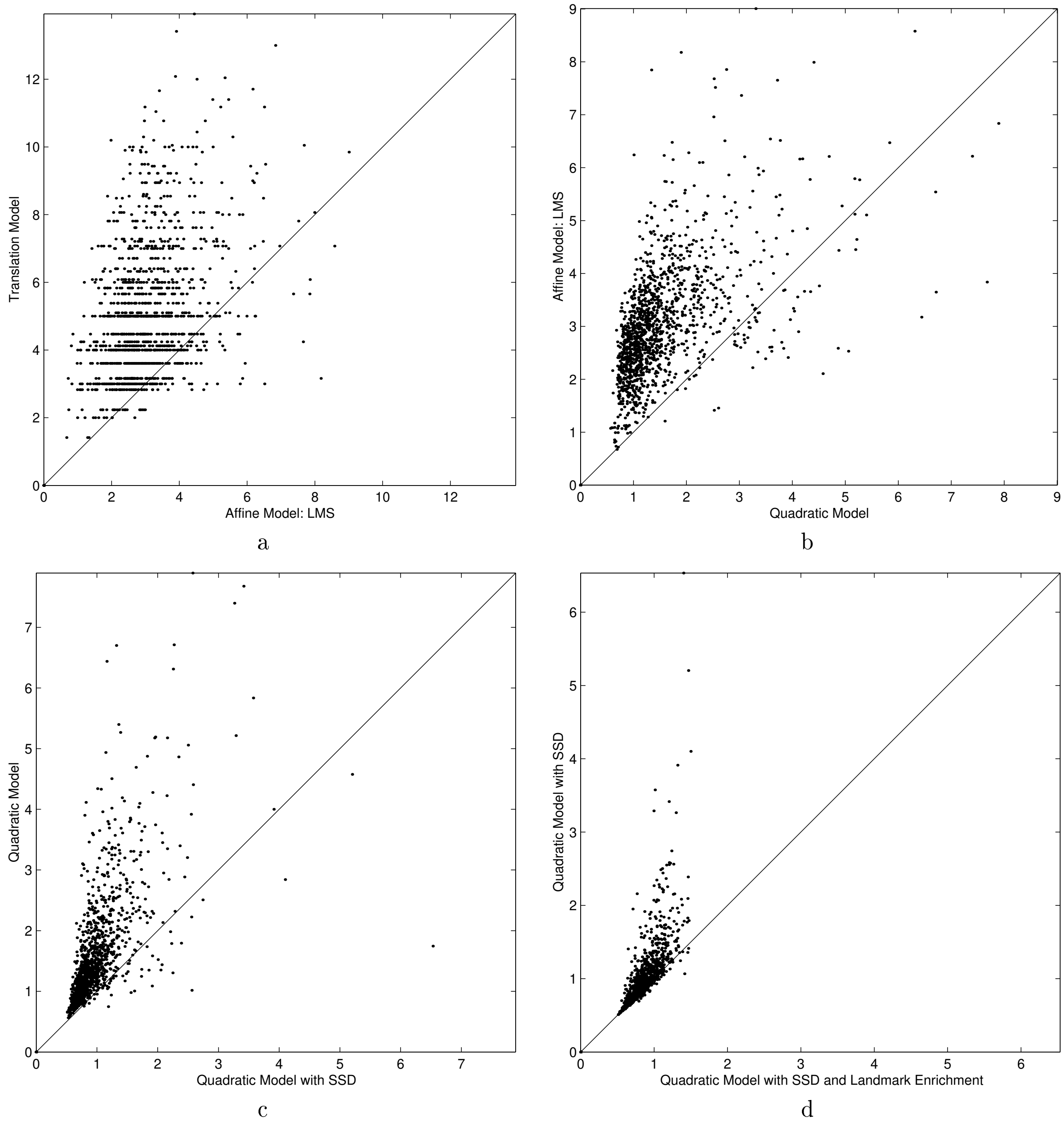


Figure 15: Overlapping region of incorrectly registered images where close and similar landmarks caused the algorithm to fail (point 'F' in Figure 12). Several of the "ghost vessels" due to misalignment are indicated with arrows. The overlap is overlap=41% and the centerline error is 3.02 pixels).

7 Discussion

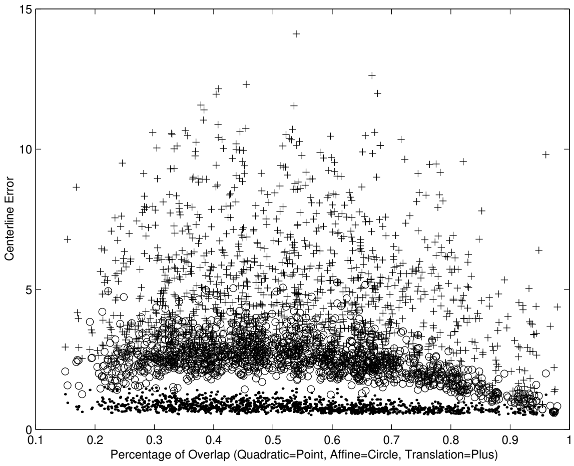
The effectiveness of the three major components of our algorithm for retinal image registration — the quadratic transformation model, the hierarchical, robust estimation procedure, and the feature refinement technique — have now been demonstrated experimentally. Further understanding of the significance and novelty of these innovations can be seen in comparison to related techniques in the registration literature.

A variety of transformation models have been used for registration, ranging from low-order models that may include just rotation, translation and scale to high-order, non-linear models that include global transformations and local deformations [6, 51, 52]. Our quadratic model fits somewhere in between, but is more sophisticated than models used in most global motion estimation and mosaic construction algorithms. Similarly, most previous retinal image registration methods have used simpler models, especially multi-point warping, affine and bilinear transformations [4, 35, 40, 56]. The few exceptions where higher-order models have been used have not been fully automatic. Mathematically, the most closely related model is the quadric reference surface of Sha'shua and others [46, 54]. Indeed, our transformation model is a *quadratic* reference surface (based on a weak perspective camera). Our experimental analysis on a large number of datasets has shown that our model accurately describes the retinal surface, sufficient to less than one pixel of registration error on 1024×1024 images. More complex models may ultimately be necessary for some clinical conditions, however. For example, a combination of our quadratic transformation model



a b c d

Figure 16: Comparison of the centerline error measures (in pixels) at the consecutive steps of the hierarchical estimator. The x, y coordinates of a point represent the error for the data pair of images for two consecutive stages of the algorithm: the earlier stage is the y value and the later stage is the x value. The $45°$ line corresponds to equally accurate results. Points above the line correspond to the later stage giving more accurate results.



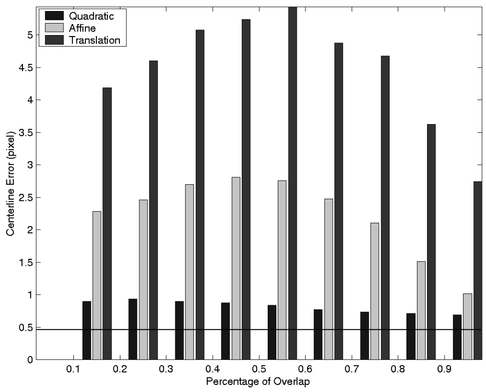


Figure 17: (a) shows the scatter plot of the centerline error measures of all pairs that are successfully registered, while (b) shows the average errors in 10 percent ranges. The centerline error measure for the quadratic model changes very little as the percentage of overlap changes, showing that the quadratic model closely represents the true image-to-image transformation. The horizontal line at 0.4766 in both plots is the minimum error due to quantization [13].

with local deformations may be necessary when registering images of a partially detached retina or images of a retina with raised surfaces caused by glaucoma or underlying tumors.

Turning now to the estimation technique, the first design decision made was to use vascular landmarks as the basis for registration. This choice is consistent with many other retinal image registration algorithms [4, 35, 56]. Reasons for the choice include viewpoint-dependent illumination, the existence of stable features, and large interframe motions. By contrast, many recent motion estimation and mosaic construction techniques in the computer vision literature are intensity-based [6, 29, 43]. These algorithms are justified by quite different assumptions: slowly varying illuminations, no assurance of stable features, and small interframe motions.

Perhaps the most significant distinguishing property of our algorithm is the hierarchy of transformation models and associated robust estimation techniques. A number of other motion and registration algorithms in the computer vision and medical image analysis literatures have used a similar model hierarchy [17, 29, 32, 43, 44, 51]; doing so is a natural approach. Of particular interest here, Cham and Cipolla [17] have developed a hierarchical, Bayesian method combining Kalman filters and RANSAC methods to establish feature correspondences over large image distances. By contrast, we use the estimated transformation itself at each stage of the hierarchy to cull the correspondence set and initiate estimation at the next level. Overall, the danger of using model hierarchies is that the estimation of lower level transformations may provide insufficient or incorrect results for upper levels of the hierarchy. For example, our translation estimation technique would fail on large inter-image scaling or rotation about the optical axis. Addressing this issue is a major direction of our ongoing research.

Robust estimation techniques have been used in a wide variety of applications within computer vision. The most closely related are those of Torr [53] and Zhang [57] for fundamental matrix estimation and related problems. These algorithms match corners detected in two different images of the same scene and use robust techniques — RANSAC [23] and M-estimators [27] — to estimate the fundamental matrix. Our estimation techniques go beyond the algorithms of Torr and of Zhang in two ways. The first is in the hierarchy of models and robust estimation. The second and more important way is that our algorithms do not assume unique correspondences, but instead mix the selection of correct correspondences with robust parameter estimation. Very few algorithms in the literature do this. See [19] for one recent exception.

The last step of our estimation algorithm is landmark refinement and correspondence enhancement based on the estimated transformation and a normalized SSD measure. This type of coupling between transformation estimation and feature detection has been used several times before. See, for example, [24, 49]. In our application, we showed that it reduces the average centerline error measure from 1.56 pixels to 0.83 pixels, clearly demonstrating its significance.

Finally, our complete algorithm offers a substantial improvements over current methods published for retinal image registration. Several early algorithms used correlation-based tracking for alignment of images within a sequence [34, 55]. Clearly, these techniques do not handle large changes in viewpoint. Our earlier method estimated a similarity transformation from feature locations detected with an interest operator [4]. Berger et. al. used a Hausdorf metric based on edges for affine estimation [8]. In practice, they currently use a higher-order transformation based on manual correspondences. Recently, Matsopoulos, et al. [35] published a retinal image registration method that used a genetic algorithm and simulated annealing to search for a globally optimum transformation. They obtained better results with the genetic algorithm. Ritter, et al. [40], by contrast, used simulated annealing, and found a strong trade-off between accuracy and speed. Zana and Klein [56] used a landmark feature-based technique and a Bayesian Hough transform to search for the best similarity transformation, which they then refined to obtain an affine transformation. Overall, these methods tend to be much slower than our hierarchical estimation technique. We exploit the blood vessel structure to obtain a more efficient search algorithm, which runs on the order of seconds rather than minutes. We have also clearly demonstrated the effectiveness of our algorithm as a function of image overlap. Finally, as discussed previously, we use a higher-order transformation for more precise registration on larger images.

8 Ongoing Work

Our ongoing work is proceeding in several major directions:

- As reported in the companion paper [16], the pairwise algorithm we have described here forms the basis for a technique that combines multiple retinal images into a seamless mosaic, giving a nearly complete, high-resolution view of the retinal surface. Together the pairwise registration and mosaic construction algorithms will form the basis for a toolset that assists in diagnosing diseases of the retina. As a preliminary step toward building this we will be providing a web-based interface to our algorithms. Our goal will be to expose the retina community to our algorithms and to obtain experimental data on a wider variety of images.
- There are two important areas where the pairwise registration algorithm can be improved. The first is in handling extremely low overlap between images. The plot in Figure 13 showing the failure rate as a function of overlap clearly establishes a baseline for measuring future progress. The second is in extending the initial translation estimation to handle rotations and scale changes. Both of these are being addressed in ongoing work by matching the detected blood vessel centerline points in addition to the landmarks used here.
- Finally, motivated by the need for computer vision techniques to assist in retinal laser surgery and related procedures, we are working toward developing real-time registration algorithms. These methods must be opportunistic, computing the most reliable registration estimate in the minimum amount of time. Progress toward this objective for the exploratory tracing and landmark feature detection algorithm is reported in [47].

Our ultimate goal is to develop a complete set of computer vision algorithms to assist in the diagnosis and treatment of diseases of the human retina. The success of the pairwise registration algorithm described here represents an important milestone in this project.

Acknowledgments

This work was supported by NIH grant 1R21 RR14038-01A1, the NSF-supported Engineering Research Center for Sub-surface Imaging and Sensing (CenSSiS), and Rensselaer Polytechnic Institute. A portion of Dr. Can's time was supported by Inonu University, Turkey. The authors gratefully acknowledge the help from the professional staff at the Center for Sight, especially the professional photographers Gary Howe and Marc Fish. The retinal images were of unnamed volunteers whose assistance is greatly appreciated. Thanks are also due to Matthew Freshman for assistance with image capture. Finally, the continuing guidance provided by Dr. James N. Turner at the Wadsworth Center in Albany is greatly appreciated.

References

[1] E. Balduf. Real-time algorithms for a laser retinal surgery system - implementation on the Silicon Graphics Octane system. Master's thesis, Rensselaer Polytechnic Institute, 1998.

[2] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16:147–185, 1974.

[3] D. E. Becker, H. Ancin, D. H. Szarowski, J. N. Turner, and B. Roysam. Automated 3-D montage synthesis from laser-scanning confocal images: Application to quantitative tissue-level cytological analysis. *Cytometry*, 25(3):235–245, November 1996.

[4] D. E. Becker, A. Can, H. L. Tanenbaum, J. N. Turner, and B. Roysam. Image processing algorithms for retinal montage synthesis, mapping, and real-time location determination. *IEEE Transactions on Biomedical Engineering*, 45(1), 1998.

[5] D. E. Becker, A. Can, H. L. Tanenbaum, J. N. Turner, and B. Roysam. Image processing algorithms for retinal montage synthesis, mapping, and real-time location determination. *IMIA Yearbook of Medical Informatics, International Medical Informatics Association, Schattauer Press, Germany (Bemmel et al., eds)*, 1999.

[6] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings Second European Conference on Computer Vision*, pages 237–252, 1992.

[7] J. W. Berger. Quantitative image sequence analysis of fundus fluorescein angiography. *Ophthalmic Surg Lasers*, 30(1):72–3, Jan 1999.

[8] J. W. Berger, M. E. Leventon, N. Hata, W. Wells, and R. Kikinis. Design considerations for a computer-vision-enabled ophthalmic augmented reality environment. In *Lecture Notes in Computer Science: 1205*, pages 399–408, 1997.

[9] J. W. Berger and D. S. Shin. Computer vision enabled augmented reality fundus biomicroscopy. *Ophthalmology*, 106(10), Oct 1999.

[10] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[11] G. Borgefors. Distance transformations in digital images. *CVGIP*, 34(3):344–371, June 1986.

[12] L. Brown. A survey of image registration techniques. *Computing Surveys*, 24(4):325–376, 1992.

[13] A. Can. *Robust Computer Vision Algorithms for Registering Images from the Curved Human Retina.* PhD thesis, Rensselaer Polytechnic Institute, 2000.

[14] 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.

[15] A. Can, C. Stewart, and B. Roysam. Robust hierarchical algorithm for constructing a mosaic from images of the curved human retina. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 286–292, 1999.

[16] 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 Transactions on Pattern Analysis and Machine Intelligence*, submitted, 2000.

[17] T. Cham and R. Cipolla. A statistical framework for long-range feature matching in uncalibrated image mosaicing. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 442–447, 1998.

[18] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.

[19] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages II:44–51, 2000.

[20] A. V. Cideciyan. Registration of ocular fundus images. *IEEE Engineering in Medicine and Biology*, 14(1):52–58, Jan. 1995.

[21] J. S. Duncan and N. Ayache. Medical image analysis: progress over two decades and the challenges ahead. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):85–105, 2000.

[22] J. L. Federman, editor. *Retina and Vitreous.* The C.V. Mosby Company, St. Louis, 1988.

[23] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24:381–395, 1981.

[24] M. Gleicher. Projective registration with difference decomposition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 331–337, 1997.

[25] M. Goldbaum, N. Katz, S. Chaudhuri, M. Nelson, and P. Kube. Digital image processing for ocular fundus images. *Ophthalmology Clinics of North America*, 3(3):447–466, September 1990.

[26] M. H. Goldbaum, V. Kouznetsova, B. L. Cot, W. E. Hart, and M. Nelson. Automated registration of digital ocular fundus images for comparison of lesions. *SPIE: Ophthalmic Technologies III*, 1877:94–99, 1993.

[27] F. R. Hampel, P. J. Rousseeuw, E. Ronchetti, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.

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

[29] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, 1994.

[30] P. K. Jensen and E. Scherfig. Resolution of retinal digital colour images. *Acta Ophthalmol Scand*, 77(5):526–529, Oct. 1999.

[31] M. Kashiwase, T. Sata, Y. Yamauchi, H. Minoda, N. Usui, T. Iwasaki, T. Kurata, and M. Usui. Progressive outer retinal necrosis caused by herpes simplex virus type 1 in a patient with acquired immunodeficiency syndrome. *Ophthalmology*, 107(4):790–794, Apr 2000.

[32] H. Lester and S. Arridge. A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149, 1999.

[33] A. A. Mahurkar, M. A. Vivino, B. L. Trus, E. M. Kuehl, M. B. Datiles, and M. I. Kaiser-Kupfer. Constructing retinal fundus photomontages. *Investigative Ophthalmology and Visual Science*, 37(8):1675–1683, July 1996.

[34] M. S. Markow, H. G. Rylander, and A. J. Welch. Real-time algorithm for retinal tracking. *IEEE Transactions on Biomedical Engineering*, 1993.

[35] G. K. Matsopoulos, N. A. Mouravliansky, K. K. Delibasis, and K. S. Nikita. Automatic retinal image registration scheme using global optimization techniques. *IEEE Transactions on Information Technology in Biomedicine*, 3(1):47–60, 1999.

[36] J. V. Miller and C. V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 300–306, 1996.

[37] F. Neumann, W. Schreiner, and M. Neumann. Computer simulation of cornorary arterial trees. *Advances in Engineering Software*, 28:353–357, 1997.

[38] T. Onuki and S. Nitta. Computer simulation of geometry and hemodynamics of canine pulmonary arteries. *Annals of Biomedical Engineering*, 21:107–115, 1993.

[39] K. E. Rayner, editor. *Eye Movements and Visual Cognition: Scene Perception and Reading*. Springer Series in Neuropsychology. Springer-Verlag, New York, 1992.

[40] N. Ritter, R. Owens, J. Cooper, R. Eikelboom, and P. van Saarloos. Registration of stereo and temporal images of the retina. *IEEE Transactions on Medical Imaging*, 18(5):404–418, 1999.

[41] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[42] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

[43] H. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *Proceedings 5th European Conference on Computer Vision*, volume II, pages 103–119, 1998.

[44] H. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):235–243, 1999.

[45] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Proceedings IEEE International Conference on Computer Vision*, pages 230–235, 1998.

[46] A. Shashua and S. Toelg. The quadric reference surface: Theory and applications. *International Journal of Computer Vision*, 23(2):185–198, 1997.

[47] H. Shen, B. Roysam, C. Stewart, J. Turner, and H. Tanenbaum. Optimal scheduling of tracing computations for real-time vascular landmark extraction from retinal fundus images. *IEEE Transactions on Information Technology in Biomedicine*, 5(1):77–91, Mar 2001.

[48] D. S. Shin, N. B. Javornik, and J. W. Berger. Computer-assisted interactive fundus image processing for macular drusen quantitation. *Ophthalmology*, 106(6):1119–1125, June 1999.

[49] H.-Y. Shum and R. Szeliski. Construction and refinement of panoramic mosaics with global and local alignment. In *Proceedings IEEE International Conference on Computer Vision*, pages 953–958, 1998.

[50] C. V. Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41(3), September 1999.

[51] R. Szeliski and S. Lavallee. Matching 3-d anatomical surfaces with non-rigid deformations using octree-splines. *International Journal of Computer Vision*, 18(2):171–186, 1996.

[52] D. Terzopoulos and D. Metaxas. Dynamic 3d models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.

[53] P. Torr and D. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.

[54] Y. Wexler and A. Shashua. Q-warping: Direct computation of quadratic reference surfaces. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages I:333–338, 1999.

[55] C. H. Wright, R. D. Ferguson, H. G. Rylander III, A. J. Welch, and S. F. Barrett. Hybrid approach to retinal tracking and laser aiming for photocoagulation. *Journal of Biomedical Optics*, 2(2), 195–203 1997.

[56] F. Zana and J. C. Klein. A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform. *IEEE Transactions on Medical Imaging*, 18(5):419–428, 1999.

[57] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.