Accuracy Assessment of Intra- and Intervisit Fundus Image Registration for Diabetic Retinopathy Screening

Kedir M. Adal,1,2 Peter G. van Etten,3 Jose P. Martinez,3 Lucas J. van Vliet,2 and Koenraad A. Vermeer1

1Rotterdam Ophthalmic Institute, Rotterdam, The Netherlands
2Quantitative Imaging Group, Delft University of Technology, Delft, The Netherlands
3Rotterdam Eye Hospital, Rotterdam, The Netherlands

Correspondence: Kedir M. Adal, Rotterdam Ophthalmic Institute, Schiedamse Vest 160d, 3011 BH Rotterdam, The Netherlands; KAdal@eyehospital.nl.
Submitted: October 27, 2014
Accepted: January 19, 2015

PURPOSE. We evaluated the accuracy of a recently developed fundus image registration method (Weighted Vasculature Registration, or WEVAR) and compared it to two top-ranked state-of-the-art commercial fundus mosaicking programs (i2k Retina, DualAlign LLC, and Merge Eye Care PACS, formerly named OIS AutoMontage) in the context of diabetic retinopathy (DR) screening.

METHODS. Fundus images of 70 diabetic patients who visited the Rotterdam Eye Hospital in 2012 and 2013 for a DR screening program were registered by all three programs. The registration results were used to produce mosaics from fundus photos that were normalized for luminance and contrast to improve the visibility of small details. These mosaics subsequently were evaluated and ranked by two expert graders to assess the registration accuracy.

RESULTS. Merge Eye Care PACS had high registration failure rates compared to WEVAR and i2k Retina ($P = 8 \times 10^{-6}$ and $P = 0.002$, respectively). WEVAR showed significantly higher registration accuracy than i2k Retina in intravisit ($P \leq 0.0036$) and intervisit ($P \leq 0.0002$) mosaics. Therefore, fundus mosaics processed by WEVAR were more likely to have a higher score (odds ratio [OR] = 2.5, $P = 10^{-3}$ for intravisit and OR = 2.2, $P = 0.006$ for intervisit mosaics). WEVAR was preferred more often by the graders than i2k Retina (OR = 6.1, $P = 7 \times 10^{-6}$).

CONCLUSIONS. WEVAR produced intra- and intervisit fundus mosaics with higher registration accuracy than Merge Eye Care PACS and i2k Retina. Merge Eye Care PACS had higher registration failures than the other two programs. Highly accurate registration methods, such as WEVAR, may potentially be used for more efficient human grading and in computer-aided screening systems for detecting DR progression.

Keywords: fundus mosaics, registration, diabetic retinopathy

Diabetic retinopathy (DR) is one of the most common complications of diabetes mellitus (DM), and results in vision loss and even blindness if not diagnosed and treated adequately. Currently, approximately 347 million people worldwide are reported to have diabetes,1 and DR accounts for 4.8% of the 37 million cases of blindness worldwide.2 The current practice of DR screening is based on regular examinations of a series of fundus images. A retinal specialist looks for pathognomonic abnormalities. In addition, manual grading is time-consuming, subjective, and limits the efficiency of the available DR screening facilities. Automated registration of fundus images can be instrumental to alleviate this problem and increases the efficiency of DR screening in two ways. Firstly, intravisit images that capture partially overlapping regions of the same retinal surface can be automatically registered to create a mosaic of the retina, enabling clinicians to do a comprehensive retinal examination at a single glance. Secondly, registration of intervisit image sets allows longitudinal analysis, facilitating retinal change detection to monitor DR development and progression.

In addition to preprocessing retinal images for more efficient human grading, fundus image registration is often used as part of computer-aided screening systems for detecting DR progression and longitudinal changes.3-6 Over the last decade, several computer-aided diagnosis (CAD) systems have been developed to analyze digital fundus images for symptoms of DR.4,7-17 The performance of these systems are comparable to expert readers in distinguishing fundus images of a normal retina from those with DR symptoms.4,10-18 Thus, CAD systems could be used in DR screening such that experts only have to evaluate suspicious or difficult cases.16-18 Moreover, registration of fundus images captured across multiple exams enables CAD systems to identify and analyze retinal surface changes due to disease progression.

Tracking small retinal features, such as microaneurysms, over time needs a very high registration accuracy. This requires a thorough evaluation of image registration methods for DR screening. Evaluation can be done either by expert graders based on visual inspection of the registered image pairs or by objective, automatic computer algorithms that assess the registration accuracy between corresponding landmark points. Due to the sparse distribution of landmark points in the field-of-view and the difficulty to extract and match these points
accurately, an objective registration accuracy assessment may be limited to a few regions. On the other hand, visual inspection by expert graders permits qualitative accuracy assessment of the entire field-of-view. Moreover, clinicians are likely to focus on regions of clinical interest, thereby producing a more clinically relevant accuracy assessment.

In this study, the accuracy of a recently developed fundus image registration method (Weighted Vasculature Registration, or WEVAR) was systematically evaluated by clinical experts in the context of automated DR screening. The evaluation was performed on intravisit and intervisit fundus image sets acquired from diabetic patients who had annual retinal exams for DR. A comparison was made with state-of-the-art commercially available fundus mosaicking programs i2k Retina (Dual-Align LLC, Clifton Park, NY, USA) and OIS AutoMontage (OIS, Sacramento, CA, USA). These programs ranked first and second, respectively, in a recent comparative study that also included IMAGEnet Professional (ranked third; Topcon, Oakland, NJ, USA). A full evaluation was done for WEVAR and i2k Retina (version 2.1.6), while Merge Eye Care PACS (Version 4.2.0.4221), the successor of OIS AutoMontage, was only partially evaluated due to high registration failure rates.

METHODS

Data Description

This retrospective observational study was conducted on fundus images that were captured during annual retinal examinations of diabetic patients who were enrolled in the ongoing DR screening program of the Rotterdam Eye Hospital in The Netherlands. A representative sample of the screening population was gathered by including all patients who were examined in a 1-week period in June 2013. During this period, a total of 85 patients was screened for DR. Because repeated examinations were needed for our evaluation, first-time patients and those who were not examined in the year before were excluded. All fundus images were acquired after pupil dilation (one drop of tropicamide 0.5%) using a nonmydriatic digital fundus camera (TRC-NW6S; Topcon, Tokyo, Japan) with a 45° field-of-view. The fundus images were 2000 × 1312 pixels in size. Although clinical guidelines suggest two fields per eye for screening purposes, in this screening program four fields are acquired per examination (Fig. 1): images of macula-centered, optic nerve-centered, superior, and temporal regions of the retinal surface were acquired from both eyes.

This study was conducted in accordance with the Declaration of Helsinki and adhered to the applicable code of conduct for the re-use of data in health research. After exporting the fundus images from the clinical image storage system, all data was anonymized prior to further processing.

Fundus Image Normalization

Color fundus images often show highly variable luminosity and contrast due to nonuniform illumination of the retina during acquisition. Because of its higher contrast, the green channel of the digital fundus images (Figs. 2a, 2b), closely resembling red-free fundus photos, is used commonly in CAD of fundus images. However, the green channel images still show considerable variation in luminosity and contrast, within and between images. Foracchia et al. proposed a method to normalize retinal images based on estimates of the local luminosity and contrast from the intensity distribution of the so-called background retina (which excludes features, such as vessels, optic disc, and lesions) and subsequently correcting for their variation over the entire retinal image. However, this method does not compensate for all illumination variation, especially around the rim of fundus images. Recently, this limitation was addressed by applying a higher-order normalized convolution, resulting in a considerably larger area with discernible retinal features (Fig. 2c).

FIGURE 1. An example of a four-field fundus image set captured during a retinal examination. From left to right: macula-centered, optic nerve-centered, superior, and temporal fundus images of a left eye.

FIGURE 2. An example of a fundus image from our data set. (a) Color fundus image. (b) Green channel. (c) Normalized fundus image using the improved normalization method.
The enhanced visibility of retinal features in these normalized images not only is beneficial for further processing by computer algorithms, but also may be used by clinicians for a better evaluation of the fundus. The graders who participated in this study preferred the normalized images over the color and green channel fundus images. Therefore, all evaluations in this study were based on normalized image.

Registration Methods for Fundus Image Mosaicking

Fundus image registration is the process of spatially mapping two or more images of the retina into a common coordinate system. The resulting spatial correspondence allows for combining the images into a single mosaic of the retinal surface to facilitate comprehensive retinal examination at a single glance. 26 Different attributes of fundus images, such as the raw intensity, the vasculature tree, and its bifurcations, may be used to determine the optimal spatial mapping parameters. In this study, two fundus image registration methods were extensively evaluated: WEVAR19 and i2k Retina, the latter representing the state-of-the-art in fundus image registration methods. 20 The main difference between the two methods lies in the fundus image attributes they use for the registration.

In brief, WEVAR aligns fundus images based on intensity and structural information derived from the retinal vasculature. 19 The method starts by normalizing the green channel of the fundus images for luminosity and contrast. The optimal alignment of the normalized images then is determined using a multiresolution matching strategy coupled with a deformation model of progressive complexity. For each intra- and intervisit image set of each eye, the method automatically selects the image having the largest overlap with the other images as the anchor image. Then, the image with the largest overlap to the anchor image is mapped sequentially to the coordinate system of this anchor image. This result becomes the new anchor image and the procedure is repeated for the remaining images until all images have been registered. This yields a set of normalized images that were transformed into a common coordinate system. These outputs then are combined into a mosaic for grading (Fig. 3a).

The i2k Retina program finds similar corresponding regions between a pair of images based on the information extracted from landmark points of the retinal vasculature. 27, 28 The program initializes the alignment by matching features extracted from vessel bifurcations and crossover points in the image pairs. The results then are refined based on vessel centerlines. Hence, the method does not make use of most of the other intensity and structural information within fundus images. Each complete set of color fundus images that must be registered were loaded into the i2k Retina program and aligned to one coordinate system using the default program settings. No preprocessing, such as normalization, was performed on the color fundus images before processing by i2k Retina, because the software may have its own internal preprocessing algorithms. To compare the registration produced by both methods, the green channel of the individual color images were normalized for variations in luminosity and contrast as before, and then combined into a mosaic using the spatial mapping that was determined during registration (Fig. 3b).

Registration Accuracy Assessment

In this study, two experienced graders who are involved in DR care, including screening and diagnosis of DR, independently assessed the registration accuracy of WEVAR and i2k Retina by scoring intra- and intervisit fundus mosaics. The graders also ranked the mosaics produced by both methods in a side-by-side comparison.

Grading mosaics is a time-consuming task and, therefore, each grader did not evaluate all data. However, to be able to compare the scores between graders, half of the available data were assessed by both graders. The remaining half were divided equally between the two graders. Note that the two mosaics of each eye by both methods were scored by the same grader. Since the side-by-side comparison was less time-consuming, both graders scored all data.

In the intervisit evaluation, the accuracy of the fundus mosaics constructed from registered fundus images that were captured during one examination was assessed. Conventionally, when combining multiple fundus images into one mosaic, overlapping areas are averaged. Although averaging or more advanced blending methods produce visually appealing results, it conceals misalignment of retinal features and thereby hinders the quality assessment. In this study intervisit mosaics were created by stacking the four registered images on top of each other. By changing the order of the images in the stack, each image appeared in the top layer once, resulting in four mosaics that differed in the regions where the images overlapped. These mosaics were put together in a video and played repeatedly for grading (see intervisit Supplementary Material for an example). The graders evaluated each mosaic by visually inspecting the vasculature alignment in the overlapping regions and assigned one of the following grades to it: “off,” at least one image is fully misplaced; “not acceptable,” misalignment larger than the width of the misaligned vessel
Graders were instructed to base their score on the region with the worst alignment. Hence, a mosaic was graded as “not acceptable” even if the misalignment occurred only in a small region of the mosaic. The i2k Retina program sometimes discarded one or more images that could not be registered into a mosaic; these mosaics were given the score “off.”

In the intervisit accuracy evaluation, all images were registered to a common coordinate system and a mosaic was produced for each visit. The two mosaics then alternated in a video and played repeatedly for grading (see intervisit Supplementary Material for an example), using the same grading scheme as for the intravisit evaluation.

In the third evaluation, the registration methods were ranked in a side-by-side comparison for each pair of intravisit mosaics. The mosaics of both methods were produced from the registered intravisit fundus images by averaging overlapping areas and each grader ranked all 140 resulting intravisit mosaic pairs that were displayed simultaneously on two identical monitors (1920 × 1080 pixels resolution). The possible grades were “slightly better” or “much better” for either mosaic, or “equal” if both were of the same quality. To avoid bias, the monitor that presented the result of each method was selected randomly for each mosaic pair.

In all three evaluations, the graders were blinded with respect to the method that was used for registration in each mosaic. Moreover, in all accuracy assessments, the mosaics of all eyes from both methods were presented in random order to the graders to avoid any bias.

Data Availability

All data that were used in this study are made publicly available through the Rotterdam Ophthalmic Data Repository (available in the public domain at http://rod-rep.com). This includes the source data (1120 fundus images), processed data (1120 normalized fundus images, all intra- and intervisit mosaic movies and images used for grading), and all grading results (intra- and intervisit mosaic grading and ranking).

Statistical Analysis

For each evaluation, two types of analyses were performed: First, the grades for both methods were evaluated for each grader separately by conventional nonparametric statistical analyses. Second, a comprehensive statistical model was defined to simultaneously evaluate all grades of both graders.

In the evaluation per grader, the grades assigned to each method were compared. To assess the difference between grades assigned to i2k Retina and WEVAR, a Wilcoxon signed-rank test was applied to the intra- and intervisit grades. Then, to quantify the preference of a grader for either method, the odds ratio (OR) of the methods was computed from the ranking grades and its significance was tested by Fisher’s exact test. The OR for each method was defined as the ratio of the number of cases that the method was preferred over all other cases. To determine the intergrader agreement and consistency, the intraclass correlation coefficient (ICC(3,1)) was calculated. The ICC values were interpreted as follows: <0.4 corresponds to poor, 0.4 to 0.75 was fair to good, and >0.75 was excellent agreement or consistency.

For a comprehensive statistical analysis of each evaluation, proportional odds mixed models were used. Here, all mosaics are modeled as random effects, whereas the methods and graders (and their interaction) are modeled as fixed effects. The ORs resulting from this model were used to quantify the influence of the aforementioned effects on the grade. Such an OR is defined as the ratio of the odds that an image gets a better grade including a certain effect over the same odds excluding that effect. The analysis for the side-by-side method ranking was based on a proportional odds model with only the graders as fixed effects. The results of this model were used to compute the OR for the methods, which then was used to determine the preference of one method over the other. The odds of each method was defined as the ratio of the probability that a method is preferred over all other cases.

RESULTS

During the 1-week screening period, 85 patients were examined for DR; among these patients, 4 were first-time patients and 11 were not examined the year before, resulting in 70 patients who had consecutive retinal examinations. A total of 1120 fundus images was acquired from 70 patients. At the time of the examination in 2012, the average age of the patients was 63 years (SD, 12 years), 33 (47.1%) were male, and 37 (52.9%) were female. From 70 patients, 140 intra- and 140 intervisit fundus photo sets were processed by WEVAR and i2k-Retina to produce mosaic movies. The 140 intravisit image sets also were processed by Merge Eye Care PACS; however, the results (described later) did not warrant further evaluation by the expert graders. The mosaic movies from WEVAR and i2k-Retina were independently assessed by two expert graders. Of the mosaic movies from each method, 70 were graded by both graders, the other mosaics were graded by a single grader. The resulting grades are summarized in Tables 1 and 2. The results
showed that WiVAr produced more “acceptable” or “perfect” mosaics and fewer “off” cases than i2k Retina according to both graders. Each grader assigned significantly more often a higher grade to the WiVAr than to i2k Retina in intravisit (Wilcoxon signed-rank test, $P = 0.0036$ and $P = 0.006$ for graders 1 and 2, respectively) and intervisit ($P = 0.0002$ and $P = 0.0001$ for graders 1 and 2, respectively) mosaic evaluations. A partial evaluation of Merge Eye Care PACS revealed that it failed to register one or more images into a mosaic, that is, “off” cases, in 19 (of 140) intravisit image sets. This was significantly higher compared to i2k Retina and WiVAr (McNemar’s test, $P = 0.002$ and $P = 8 \times 10^{-6}$, respectively). Therefore, Merge Eye Care PACS was excluded from further evaluation.

In Table 3, the ranks assigned to each of the methods in the side-by-side comparison of intravisit mosaic pairs are summarized. WiVAr produced mosaics that were preferred more often by both graders than i2k Retina. Examples of pairs of mosaics that were compared and ranked are shown in Figure 5. For grader 1, the odds of preferring WiVAr, expressed by the ratio of grades D and E over grades A to C was 0.14; the odds of preferring i2k Retina (A and B over C to E) was 0.02. The resulting OR was 6.3 ($P = 0.002$). For grader 2, the OR was 6.0 ($P = 0.0007$), showing that a higher rank was assigned significantly more frequently to mosaics produced by WiVAr than to mosaics produced by i2k Retina.

Although the results indicated that WiVAr yields significantly higher accuracy in intra- and intervisit registration than i2k Retina, the intergrader agreement and consistency between the grades assigned by both graders ranged from poor to moderate levels ($I_{CC}(3,1)$ agreement and consistency of 0.52 and 0.65, respectively, for the intravisit grades and 0.35 and 0.71 for intervisit grades). Thus, the data cannot simply be pooled for analysis by ignoring the grader. Instead, a comprehensive statistical analysis based on the proportional odds mixed model was applied to the grades of both graders altogether.

The fitted parameters of a proportional odds mixed model for the intravisit grades considering the method and grader as fixed effects and the fundus image set as a random effect are summarized in Table 4. Interaction between the methods and graders was not included in the final model as its effect on the grades was insignificant ($P = 0.17$). The resulting coefficient of the method indicates that mosaics processed by WiVAr were significantly more likely to receive a higher grade compared to i2k Retina ($P = 10^{-5}$). The coefficient of 0.94 corresponds to an OR of $e^{0.94} = 2.5$. Thus, the odds of receiving a higher score were 2.5 times more likely for WiVAr than for i2k Retina. The model corrected for the fact that the odds of receiving a higher grade from grader 2 were far higher (OR of 13.0).

Similarly, Table 5 shows the parameters of a proportional odds mixed model fit for the intervisit grades. In this case, the interaction between the methods and graders was significant. The estimated coefficient associated with the method’s effect...
TABLE 5. The Estimated Coefficients of the Proportional Odds Mixed Model Fit for the Intervisit Grades

<table>
<thead>
<tr>
<th>Effects</th>
<th>Estimate</th>
<th>SE</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>0.79</td>
<td>0.29</td>
<td>0.006</td>
</tr>
<tr>
<td>Grader</td>
<td>5.06</td>
<td>0.55</td>
<td>&lt;2 × 10⁻¹⁶</td>
</tr>
<tr>
<td>Method × grader</td>
<td>1.23</td>
<td>0.60</td>
<td>0.037</td>
</tr>
</tbody>
</table>

indicates that for both graders the intervisit mosaics processed by WeVaR are significantly more likely to receive a higher grade compared to i2k Retina ($P = 0.006$). In addition, grader 2 assigned a higher grade compared to grader 1. The resulting OR of WeVaR versus i2k Retina is $e^{0.79} = 2.2$ and $e^{5.06+1.23} = 72.5$ for graders 1 and 2, respectively.

To quantify the difference in accuracy and quality between the intravisit mosaics of the two methods, the ranks of each method from the side-by-side comparison also were modeled using the proportional odds model. As expected, due to the relative nature of the scores, the grader effect was insignificant ($P = 0.68$). The resulting model parameters are shown in Table 6. The odds of preferring WeVaR were $e^{1.85} = 0.16$, while the odds of preferring i2k Retina were $e^{-3.66} = 0.03$. The OR, therefore, was 6.1 ($P = 7 \times 10^{-5}$), indicating that the graders were 6.1 times more likely to prefer the results of WeVaR.

**DISCUSSIONS**

This study assessed the accuracy of vessel alignment in mosaics constructed from intra- and intervisit fundus image sets for a recently developed fundus registration method (WeVaR). An extensive accuracy assessment and comparison with two top-ranked state-of-the-art commercial fundus mosaicking programs (i2k Retina and Merge Eye Care PACS) shows that WeVaR yields a significantly higher registration accuracy in intravisit ($P \leq 0.0036$) and intervisit ($P \leq 0.0002$) mosaics. The likelihood of receiving a higher score was 2.5 ($P = 10^{-5}$) and 2.2 ($P = 0.006$) times higher for WeVaR than for i2k Retina on intra- and intervisit mosaics, respectively. Due to a very high registration failure rate, Merge Eye Care PACS was excluded from a full evaluation. Despite the generally higher scores from one grader, the results from graders show that WeVaR has a significantly higher registration accuracy and significantly fewer failures than i2k Retina. A comprehensive statistical analysis, taking into account the intergrader variability, also revealed a strong association between the grades assigned to each mosaic and the method used to produce it. In the side-by-side comparison, both graders preferred WeVaR over i2k Retina. The two graders are very experienced in grading DR, but not in assessing registration accuracy. Therefore, the difference between the scores of the two experts might be attributed to differences in evaluation strategies of the mosaics. WeVaR aligns fundus images based on intensity and structural information derived from the retinal vasculature. Therefore, the better the visibility of the retinal vasculature, the more accurate the registration results become. This might be due to the possibly very small number of available features that i2k Retina uses to match the temporal, superior, and macula-centered fundus images that have a (very) small overlap region. Such image deformation causes changes in shape and area of retinal pathology and, thus, may hinder a correct interpretation. WeVaR, on the other hand, introduces an example of an intravisit fundus mosaics of the same eye produced by using (a) WeVaR and (b) i2k Retina. Note the difference in deformation between the two mosaics despite the registration of the individual images being correct in both cases.
little deformation, and yet provides higher quality mosaics that allows clinical experts to make accurate diagnosis.

Intra- and intervisit fundus mosaics of WiVeAR can improve the efficiency of today’s DR screening practice. Intravisit mosaics provide a single large field of the retina for comprehensive analysis and more efficient grading. Intervisit mosaics can be used to analyze fundus photos of successive retinal exams to monitor DR progression through biomarkers, such as the microaneurysm turnover rate. Although our target application is DR screening, these mosaics also could be used in diagnosis and monitoring of other retinal diseases, such as age-related macular degeneration.

Current clinical guidelines on referral of DR patients are based on the presence and detection of lesions in fundus images and, therefore, exclude the dynamics of these lesions. This implies that longitudinal analysis, and, therefore, accurate registration, of fundus images does not have a significant role in current clinical care. However, recent studies suggest that the progression rate of microaneurysms over time may be a better biomarker for DR progression than the differences between the number of microaneurysms at successive examination. The studies also suggest a correlation between the microaneurysm turnover rate and the likelihood of developing clinically significant macular edema (CSME).

We showed that WiVeAR was significantly better in constructing intra- and intervisit mosaics, and obtained a significantly higher registration accuracy than Merge Eye Care PACS and i2k Retina. Merge Eye Care PACS had high registration failures compared to WiVeAR and i2k Retina. A higher registration accuracy is of clinical interest as it is an essential step toward having an automated and reliable objective disease progression measure for progressive eye disease, such as DR. An objective progression measure, such as microaneurysm turnover rate, aids clinicians in assessing disease progression for a proactive and effective screening and treatment planning, thereby improving the quality of service provided by eye care centers.

Acknowledgments
The authors thank Susan R. Bryan for her advice on the statistical modeling used in this study.

Disclosure: K.M. Adal, None; P.G. van Etten, None; J.P. Martinez, None; L.J. van Vliet, None; K.A. Vermeer, None

References


