Automated Assessment of Diabetic Retinopathy Severity Using Content-Based Image Retrieval in Multimodal Fundus Photographs

Gwénolé Quellec,1 Mathieu Lamard,2 Guy Cazuguel,1 Lynda Bekri,3 Wissam Daccache,3 Christian Roux,1 and Béatrice Cochener4

PURPOSE. Recent studies on diabetic retinopathy (DR) screening in fundus photographs suggest that disagreements between algorithms and clinicians are now comparable to disagreements among clinicians. The purpose of this study is to (1) determine whether this observation also holds for automated DR severity assessment algorithms, and (2) show the interest of such algorithms in clinical practice.

METHODS. A dataset of 85 consecutive DR examinations (168 eyes, 1176 multimodal eye fundus photographs) was collected at Brest University Hospital (Brest, France). Two clinicians with different experience levels determined DR severity in each eye, according to the International Clinical Diabetic Retinopathy Disease Severity (ICDRS) scale. Based on Cohen’s kappa (κ) measurements, the performance of clinicians at assessing DR severity was compared to the performance of state-of-the-art content-based image retrieval (CBIR) algorithms from our group.

RESULTS. At assessing DR severity in each patient, intraobserver agreement was κ = 0.769 for the most experienced clinician. Interobserver agreement between clinicians was κ = 0.526. Interobserver agreement between the most experienced clinicians and the most advanced algorithm was κ = 0.592. Besides, the most advanced algorithm was often able to predict agreements and disagreements between clinicians.

CONCLUSIONS. Automated DR severity assessment algorithms, trained to imitate experienced clinicians, can be used to predict when young clinicians would agree or disagree with their more experienced fellow members. Such algorithms may thus be used in clinical practice to help validate or invalidate their diagnoses. CBIR algorithms, in particular, may also be used for pooling diagnostic knowledge among peers, with applications in training and coordination of clinicians’ prescriptions.

D iabetic retinopathy (DR) is the leading cause of blindness in the working population of the United States and the European Union.1–2 Detecting and monitoring DR in the at-risk population (diabetics), generally using eye fundus photography, is crucial for providing timely treatment of DR and therefore preventing visual loss.3 With the likelihood of an increase in the at-risk population,4 many computer image analysis algorithms have been proposed in the literature to analyze fundus photographs automatically.5–6 Recently, several image-analysis groups compared the performance of their image analysis algorithms with the performance of clinicians in detecting DR in large-scale screening programs.7–10 Disagreements between algorithms and clinicians were found to be comparable to disagreements among clinicians.9

Screening DR mainly involves detecting microaneurysms, usually the first appearing signs of DR, although detecting hemorrhages and exudates has also been proven useful.11 In comparison, grading (i.e., monitoring) DR is much more complicated: according to the International Clinical Diabetic Retinopathy Disease Severity (ICDRS) scale, it involves detecting additional types of lesions (e.g., neovascularizations and intraretinal microvascular abnormalities) with a larger variability of scales and shapes.12 As a consequence, little work has been done so far to automate DR grading, in comparison with DR screening. However, several algorithms of increasing complexity have recently been proposed by our group for automated DR grading using fundus photography,13–14 together with demographic data in the most advanced algorithms.15–16 These algorithms are all based on the content-based image retrieval (CBIR) paradigm, which is explained hereafter.17,18 In CBIR, the content of images is characterized with low-level features (e.g., color, texture, and local shape) and these features are mapped to concepts (e.g., DR severity) using machine learning; in particular, it is not necessary to develop a dedicated algorithm for each lesion type. Once a query image has been automatically characterized, the most similar images, together with their medical interpretations, are searched for in a reference database. These most similar images are then used to infer an automated diagnosis for the query image. Note that a CBIR approach, relying on dedicated algorithms and limited manual inputs, has also been proposed by another image analysis group.19

The purpose of this study was twofold: first, to discover whether, as in a DR screening context, disagreements between automated DR grading algorithms and clinicians are comparable to disagreements among clinicians. For this purpose, a dataset of 85 consecutive DR examinations (1176 images) was graded, at eye and patient level, by two clinicians with different experience levels. The dataset was also automatically graded by three (one novel and two recently published) algorithms from our group. The first two automated algorithms analyze images independently to make a decision; the second...
algorithm, which is novel, allows additional flexibility in the way images are characterized. The third algorithm uses both image characterizations and demographic data to make a decision; note that analyzing both image characterizations and demographic data are in the DR monitoring protocol. Second, the possible uses of automated DR-grading algorithms in a clinical context were examined. The place of automated DR screening algorithms in clinical practice has been widely discussed.7,9—these algorithms may be used for triage, in that physicians are unable to screen the entire at-risk population—but those of automated DR grading have not. The specific advantages of CBIR in this context are emphasized.

METHODS
Diabetic Retinopathy Dataset (DRD)
A dataset of 85 consecutive DR follow-up examinations was collected at Brest University Hospital (Brest, France) from June 2003 through September 2007. All patients had known diabetes at the time of their examinations. Demographic data was collected during each examination (e.g., sex of the patient, age of the patient at the time of the examination, and personal and family medical history). Since two patients contributed one eye to the study, 168 eyes were photographed. Photographs were obtained by retinal photographers, using a retinal digital camera (Topcon TRC-50I A, Tokyo, Japan) connected to a computer. Seven photographs were obtained per eye: one red-free photograph, one blue-filtered photograph, and seven fluorescein angiograms (Fig. 1). After fluorescein injection, a temporal photograph sequence was obtained: one early angiograph, three intermediate angiographs (central, temporal and nasal fields), and one late angiograph. Overall, 1176 photographs were obtained. These images have a definition of 1280 pixels per row and 1008 pixels per column; they are stored in either TIFF or JPEG format.

The study protocol adhered to the tenets of the Declaration of Helsinki. The Institutional Review Board of Brest University Hospital approved the study protocol, and because only deidentified data were used, a waiver of written informed consent was granted.

Human Expert Standard
Two clinicians were involved in this study: one with 7 years’ (Clinician1) and the other with 2 years’ (Clinician2) experience. Each clinician was asked to grade disease severity in all 168 eyes according to a modified ICDRS scale. This scale consists of the five ICDRS levels: 0, no apparent DR; 1, mild nonproliferative DR; 2, moderate nonproliferative DR; 3, severe nonproliferative DR; and 4, proliferative DR,12 as well as an additional level (5, treated DR). The clinicians interpreted the 168 eyes in randomized order. When interpreting one eye, the clinicians had access to all seven photographs, as well as the available demographic data, but they were masked to all photographs and interpretations from the contralateral eye.

Two months later, Clinician1 interpreted the dataset a second time. Therefore, three interpretations are available at eye level: Clinician1 contributed EyeGrades1a and EyeGrades1b; Clinician2 contributed EyeGrades2.

Based on each of these three interpretations, disease severity was also determined (automatically) at patient level: When a patient contributed two eyes to the study, disease severity at patient level was defined as the maximum disease severity at eye level among those two eyes. Therefore, three interpretations are also available at patient level: PatientGrades1a, PatientGrades1b, and PatientGrades2.

Training and Testing Subsets
The proposed algorithms rely on the machine learning paradigm, and so examples are necessary for the training. In this purpose, DRD was divided between two subsets (A and B) with equal distribution of sex, diabetes type, and DR severity. All eyes from the same patient were assigned to the same subset. Except for the above-mentioned conditions, DRD was divided randomly between the two subsets.

EyeGrades1a and PatientGrades1a were used as the reference standard for algorithm supervision at eye level and patient level, respectively. Performance was assessed by two-fold cross-validation. At first, subset A was used for training (i.e., tuning the algorithms), and subset B was used for testing (i.e., comparing the outputs of the algorithms with the reference standard). Then, subset B was used for training, and subset A was used for testing.

Automated DR Severity Assessment Using CBIR
To automatically grade disease severity in a query eye (or patient) Q, the following procedure was applied:

1. The digital content of each image I_a in Q and of each image I_q in the training subset was automatically characterized by a feature vector.
2. The distance between the characterization of I_a and that of each image I_q in the training subset was computed.
3. The k nearest neighbors of I_a within the training subset, were sought with respect to the distance measure in step 2.
4. The most frequent diagnosis among the nearest neighbors of every image I_q in Q (according to the reference standard) was assigned to Q.

FIGURE 1. Set of multimodal photographs from one eye diagnosed with mild nonproliferative DR. (a) Red-free photograph. (b) Blue-filtered photograph. (c) Early angiograph. (d) Intermediate angiograph (central retina). (e) Intermediate angiograph (temporal retina). (f) Intermediate angiograph (nasal retina). (g) Late angiograph.
shown. This approach was improved further in the present study, and the priority of this methodology over several alternatives has been established.

Results of CBIR algorithms from our group were compared to the performance of two clinicians with different experience levels. Agreement between each algorithm and clinicians was then assessed, in the testing subset, using the optimal value for $k$.

Use of a CBIR Algorithm as a Second Opinion

Finally, we tested the effectiveness of a CBIR algorithm at providing a second opinion. The proposed CBIR algorithms, trained against the reference standard provided by the most experienced clinician ($Clinician_1$), were used to automatically predict when the least experienced clinician ($Clinician_2$) disagrees with his more experienced fellow member. We compared the probability of agreement (or disagreement) between $Clinician_1$ and $Clinician_2$, knowing that $Clinician_2$ agrees (or disagrees) with a given algorithm, to the overall probability of agreement (or disagreement) between $Clinician_1$ and $Clinician_2$. A two-tailed test for difference between proportions was performed.

Results

Patients were 61 years old on average (SD 13 years). There were 41 women and 44 men. Of the patients, 32 had type I diabetes and 46 had type II; diabetes type was unknown for 7 patients.

Confusion matrices of clinician interpretations, in terms of DR severity at eye and patient level, are given in Tables 1 and 2, respectively.

In comparison, confusion matrices of the best performing algorithm ($Fusion$) versus the reference standard, in terms of DR severity at eye and patient level, are given in Table 3. The optimal value for $k$ was $k = 5$ on subset A and $k = 7$ on subset B. Agreement between all algorithms and clinicians, at eye and patient levels, are reported in Tables 4 and 5, respectively.

The effectiveness of CBIR at predicting when $Clinician_2$ agrees or disagrees with $Clinician_1$, at the eye and patient levels, is presented in Tables 6 and 7, respectively.

Discussion

In this article, the diabetic retinopathy (DR) grading performance of CBIR algorithms from our group was compared to the performance of two clinicians with different experience levels.

| Table 1. Confusion Matrices of Clinician Interpretations, in Terms of DR Severity, at Eye Level |

<table>
<thead>
<tr>
<th>EyeGrades1a</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyeGrades1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>23</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>36</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>41</td>
<td>35</td>
<td>38</td>
<td>14</td>
<td>17</td>
<td>168</td>
</tr>
</tbody>
</table>

Diagonal elements are indicated in bold.

The first three steps are the usual steps of a CBIR system. Should a clinician using the system disagree with the proposed automated diagnosis (step 4), the nearest neighbors can be displayed and used by the clinician to revise the diagnosis.

The use of the wavelet transform has been proposed in previous works to characterize the digital content of images, and the superiority of this methodology over several alternatives has been shown. This approach was improved further in the present study, in that a novel set of wavelet filters was introduced. Two wavelet adaptation algorithms of increasing complexity, referred to as the Global and Local algorithms, are presented in the Appendix.

A third algorithm, referred to as the Fusion algorithm, was evaluated: steps 1 and 2 were based on local wavelet adaptation and steps 3 and 4 were improved, as explained hereafter. A recently published information fusion algorithm from our group was compared to the performance of two clinicians with different experience levels.

A fourth algorithm, referred to as the NoAngiography algorithm, was evaluated: It is similar to the Fusion algorithm, except that it is masked to all angiograms. Should the proposed algorithm perform equally well without angiographs, we might recommend that no (or less) fluorescein injection be performed, which would allow the imaging session to be streamlined and less invasive.

Interobserver Agreement

Agreements between (1) the two clinicians, or (2) the same clinician at two different times, or (3) a clinician and an algorithm, were assessed by Cohen’s $\kappa$ and weighted $\kappa$ ($\kappa_w$). The equation for $\kappa$ is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where $p_o$ is the observed probability of agreement and $p_e$ is the probability of chance agreement. The simplest and more standard weighting scheme was used for $\kappa_w$, a disagreement of $n_{d}$ severity levels was weighted by $n_{d}$.

Training the Algorithms

Each algorithm (Local, Global, Fusion and NoAngiography) was tuned to maximize Cohen’s $\kappa$ between its outputs and the reference standard in the training subset (see Training and Testing Subsets). In particular, $k$, the number of nearest neighbors (see Automated DR Severity Assessment Using CBIR), was trained by leave-one-out cross-validation in the training subset. Agreement between each algorithm and clinicians was then assessed, in the testing subset, using the optimal value for $k$.

Use of a CBIR Algorithm as a Second Opinion

Finally, we tested the effectiveness of a CBIR algorithm at providing a second opinion. The proposed CBIR algorithms, trained against the reference standard provided by the most experienced clinician ($Clinician_1$), were used to automatically predict when the least experienced clinician ($Clinician_2$) disagrees with his more experienced fellow member. We compared the probability of agreement (or disagreement) between $Clinician_1$ and $Clinician_2$, knowing that $Clinician_2$ agrees (or disagrees) with a given algorithm, to the overall probability of agreement (or disagreement) between $Clinician_1$ and $Clinician_2$. A two-tailed test for difference between proportions was performed.
TABLE 3. Confusion Matrices of the Best Performing Algorithm (Fusion) versus the Reference Standard, in Terms of DR Severity

<table>
<thead>
<tr>
<th>Reference Standard</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion at eye level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>27</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>25</td>
<td>7</td>
<td>2</td>
<td>43</td>
<td>59</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>25</td>
<td>2</td>
<td>41</td>
<td>57</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>41</td>
<td>35</td>
<td>38</td>
<td>14</td>
<td>17</td>
<td>168</td>
</tr>
<tr>
<td>Fusion at patient level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>19</td>
<td>21</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>2</td>
<td>22</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>18</td>
<td>18</td>
<td>21</td>
<td>9</td>
<td>10</td>
<td>85</td>
</tr>
</tbody>
</table>

Diagonal elements are indicated in bold.

First, interclinician agreement ($\kappa = 0.493$ at eye level; $\kappa = 0.526$ at patient level) was much lower than intraclinician agreement ($\kappa = 0.809$ at eye level; $\kappa = 0.769$ at patient level), at least for clinicians with different experience levels. Note, however, that the diagnoses of Clinician 2 (the least experienced clinician) seldom differed from those of Clinician 1 (the most experienced clinician) by more than one severity level (Tables 1, 2); wider divergences were observed more often between algorithms and Clinician 1 (Table 3).

Second, the simplest CBIR algorithms (Global and Local), which combine image characterizations in a basic way, were less efficient than Clinician 2 in terms of Cohen’s $\kappa$ and weighted $\kappa$ (Tables 4, 5). On the other hand, the performance of the most advanced algorithm (Fusion), which elegantly combines image characterizations and demographic data, compared favorably to the performance of Clinician 2 ($\kappa = 0.573$ at eye level, $\kappa = 0.592$ at patient level).

Third, we found that masking the Fusion algorithm to all angiographs noticeably decreased diagnosis performance ($\kappa = 0.466$ at eye level; $\kappa = 0.457$ at patient level). This performance decrease may be due to the higher discrimination power of angiography over other image modes. However, it may also be because nasal and temporal fields were photographed only after fluorescein injection. Further analyses are therefore needed to draw conclusions about the usefulness of angiography for automated DR severity assessment.

Fourth, the potential usefulness of CBIR algorithms as a second opinion, to assist the least experienced clinicians, has been shown (Tables 6, 7). In particular, whenever Clinician 2 disagreed with the algorithm at patient level, there was a 73.81% probability that he also disagreed with Clinician 1, as opposed to 38.82% without prior knowledge ($P = 0.0002$). This result could serve as a warning that he should revise his diagnosis. Similarly, whenever Clinician 2 agreed with the algorithm, there was a 95.35% probability that he also agreed with Clinician 1, as opposed to 61.18% without prior knowledge ($P < 0.0001$). This should increase his confidence.

Note that each of the above observations was made both at eye level and at patient level.

We believe the fourth observation is of great practical value. We propose that an algorithm be used in the context of DR severity assessment, as a second opinion, to help validate or invalidate the diagnoses of young clinicians. Because such algorithms are able to provide a diagnosis in seconds, they may be embedded in clinicians’ workstations and display the proposed diagnosis on a screen. One advantage of the CBIR approach, over traditional computer-assisted diagnosis (CADx), is its interactivity: Should the clinician disagree with the proposed second opinion, he or she may visualize (also from a workstation) the $k$ nearest neighbors from the reference dataset that were used to infer the second opinion, together with their medical interpretations from more experienced clinicians. This feature would help the clinician (1) to see whether the algorithm obviously made an error and, if not, (2) to compare his or her interpretation with that of his or her peers on similar cases.

More generally, the proposed CBIR-based approach may be used to pool diagnostic knowledge among peers (either hospitalwide or nationwide), which has several possible applications. First, it may be used for training: Interns may now be able to compare their interpretations of real-life cases with those of renowned experts. Second, it may help in reducing interclinician variability and therefore help clinicians to coordinate their clinical decisions and prescriptions in a DR-grading program (e.g., to determine which patients should undergo a...

TABLE 4. Agreement between Observers at Eye Level

<table>
<thead>
<tr>
<th></th>
<th>EyeGrades1a</th>
<th>EyeGrades1b</th>
<th>EyeGrades2</th>
<th>Global</th>
<th>Local</th>
<th>Fusion</th>
<th>NoAngiography</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyeGrades1a</td>
<td>1</td>
<td>0.809</td>
<td>0.493</td>
<td>0.387</td>
<td>0.456</td>
<td>0.573</td>
<td>0.466</td>
</tr>
<tr>
<td>EyeGrades1b</td>
<td>1</td>
<td>0.410</td>
<td>0.346</td>
<td>0.422</td>
<td>0.509</td>
<td>0.391</td>
<td>0.446</td>
</tr>
<tr>
<td>EyeGrades2</td>
<td>1</td>
<td>0.411</td>
<td>0.380</td>
<td>0.391</td>
<td>0.414</td>
<td>0.516</td>
<td>0.507</td>
</tr>
<tr>
<td>Global</td>
<td>1</td>
<td>0.683</td>
<td>0.516</td>
<td>0.516</td>
<td>0.683</td>
<td>0.696</td>
<td>0.614</td>
</tr>
<tr>
<td>Local</td>
<td>1</td>
<td>0.689</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>Fusion</td>
<td>1</td>
<td>0.738</td>
<td>0.738</td>
<td>0.738</td>
<td>0.738</td>
<td>0.738</td>
<td>0.738</td>
</tr>
<tr>
<td>NoAngiography</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Agreement with the reference standard is indicated in bold.
The convolution of information from following set (referred to as a subband):

\[ I = \text{CHARACTERIZATION FOR CBIR} \]

Algorithm: test of agreement

\[ P(\text{EyeGrades2} = \text{EyeGrades1a}) \]

The coefficients of filter \( w \) were tuned (as described in §a or §b) to increase the performance of DR severity assessment in the training subset.

To characterize the digital content of image \( I \), the distribution of the \( x_{s,K,I}(i,j) \) coefficients is modeled in several subbands. Because the \( x_{s,K,I}(i,j) \) coefficients in each subband have a 0-mean generalized Gaussian distribution (for small values of \( s \)), their distribution can be efficiently modeled by their standard deviation \( \sigma_{s,K,I}(i,j) \) and kurtosis \( \kappa_{s,K,I}(I) \):

\[
\begin{align*}
\sigma_{s,K,I}(I) &= \sqrt{\frac{m_{s,K,I}(I)}{\sigma_{s,K,I}(I)}} - 3 \\
\kappa_{s,K,I}(I) &= \frac{\text{third order moment of the distribution}}{\text{square of the second order moment}} \\
m_{s,K,I}(I) &= \frac{s^4}{MN} \sum_{i,j} x_{s,K,I}(i,j)
\end{align*}
\]

where \( m_{s,K,I}(I) \) is the \( d \)th order moment of the distribution. The proposed image characterization is a feature vector consisting of the \( (\sigma_{s,K,I}(I), \kappa_{s,K,I}(I)) \) couples extracted in several subbands. To characterize the lowest frequencies (horizontal, vertical, and nondirectional).

\[ \text{Algorithm: test of disagreement} \]

Distance \( D(I,J) \) between the characterization of image \( I \) and that of image \( J \) was defined as follows:

\[ D(I,J) = \sum_{i,j} w_{i,j} |x_{s,K,I}(i,j) - x_{s,K,J}(i,j)| \]

\[ \text{Algorithm: test of agreement} \]

TABLE 6. Agreement between Observers at Patient Level

<table>
<thead>
<tr>
<th>PatientGrades1a</th>
<th>PatientGrades1b</th>
<th>PatientGrades2</th>
<th>Global</th>
<th>Local</th>
<th>Fusion</th>
<th>NoAngiography</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>( 0.769 )</td>
<td>( 0.526 )</td>
<td>( 0.384 )</td>
<td>( 0.450 )</td>
<td>( 0.592 )</td>
<td>( 0.457 )</td>
</tr>
<tr>
<td>( I )</td>
<td>( 0.485 )</td>
<td>( 0.375 )</td>
<td>( 0.410 )</td>
<td>( 0.578 )</td>
<td>( 0.391 )</td>
<td>( 0.431 )</td>
</tr>
<tr>
<td>( I )</td>
<td>( 1 )</td>
<td>( 0.397 )</td>
<td>( 0.382 )</td>
<td>( 0.391 )</td>
<td>( 0.436 )</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>( 1 )</td>
<td>( 0.778 )</td>
<td>( 0.556 )</td>
<td>( 0.757 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local</td>
<td>( 1 )</td>
<td>( 0.723 )</td>
<td>( 0.734 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>( 1 )</td>
<td></td>
<td>( 0.673 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoAngiography</td>
<td>( 1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Agreement with the reference standard is indicated in bold.

The best prediction rate is indicated in bold.
The best prediction rate is indicated in bold.

\[ D(I, f) = D_0(I, f) + \sqrt{\sum_{k,l} \alpha_{s,k,l} [\sigma_{s,k,l}(I) - \sigma_{s,k,l}(f)]^2 + \beta_{s,k,l} [\kappa_{s,k,l}(I) - \kappa_{s,k,l}(f)]^2} \quad (4) \]

where \( D_0(I, f) \) denotes the Euclidean distance between the intensity histograms of \( I \) and \( f \); \( \alpha_{s,k,l} \) and \( \beta_{s,k,l} \) are subband weights.

### a. Global Adaptation of Image Characterizations and Distance Measures

In this first scenario, one wavelet filter was tuned for each filter size (i.e., each couple: \( K, L \)). Small filter sizes proved sufficient in previous works,\(^{13,14}\) and so the following couples were used:

- \( K = 2, L = 0 \) \( \Rightarrow \) filter size: \( 5 \times 1 \) (vertical filter \( V_5 \)).
- \( K = 0, L = 2 \) \( \Rightarrow \) filter size: \( 1 \times 5 \) (horizontal filter \( H_1 \)), and
- \( K = 1, L = 1 \) \( \Rightarrow \) filter size: \( 3 \times 3 \) (nondirectional filter \( N_3 \)).

The sum of all coefficients in a wavelet filter must be \( 0 \); consequently, there were four undetermined coefficients in the first two filters and eight in the last one. As in previous work, three analysis scales were used in this study: \( s = 1 \), \( s = 2 \), and \( s = 3 \); overall, there were 18 undetermined weights in equation 4.

The undetermined coefficients (16 wavelet coefficients and 18 weights) were tuned to maximize Cohen’s \( \kappa \) between the outputs of the algorithm and the reference standard, in the training subset. A genetic algorithm was used to find the optimal set of coefficients.\(^{13,24}\)

### b. Local Adaptation of Image Characterizations and Distance Measures

Since several types of lesions play a role in the definition of DR severity,\(^{12}\) it is likely that the boundaries between severity levels in the area of image characterizations are complex. Therefore, in this second scenario, the wavelet filters and distance measures were allowed to vary continuously in image characterization space. A continuous function was used to map \( d_s(I) \), the initial characterization of an image \( I \), obtained by the initial set of filters \( (V_5, H_1, N_3) \) (see §a), to a new set of filters \( (V_5^d, H^d, N_3^d) \) and a new set of weights \( W(I) \). Continuity means that two images with similar initial characterizations are mapped to similar sets of filters and weights; the following equation holds for \( V(I) \) [similar equations hold for \( H(I), N_3(I) \), and \( W(I) \)]:

\[ \forall \ v > 0, \exists \ \eta > 0 \ \forall I, f / d_s(I) - d_s(f) / \| \eta \Rightarrow \| V(I) - V(f) \| / \leq \varepsilon \quad (5) \]

The mapping functions above were tuned on the training subset: For each image \( I_u \) in the training subset \( (V_5^d(I_u), H^d(I_u), N_3^d(I_u)) \) and \( W(I_u) \) were tuned to maximize severity assessment performance for \( I_u \), while respecting the continuity constraint (see §c). The continuity constraint prevented the algorithm from overfitting the training data. Moreover, it was necessary to define \( (V(I_u), H(I_u), N_3(I_u)) \) and \( W(I_u) \), where \( I_u \) is an image in the testing subset. These coefficients were obtained by multivariate interpolation:\(^{25}\)

\[ V(I_u) = \sum_{I_0} e^{-\lambda_{i}d(I_0) - d_s(i) / \| \lambda > 0 \quad (6) \]

Similar equations hold for \( H(I_u), N_3(I_u), \) and \( W(I_u) \). These new filters and weights were used to find the \( k \) most similar neighbors of \( I \).

### c. Training the Mapping Functions for Local Filter and Weight Adaptation

Let \( I_u \) be an image from the training subset. A gradient descent was used to find \( (V(I_u), H(I_u), N_3(I_u), W(I_u)) \), using \( (V_5^d(I_u), H_1(I_u), N_3(I_u)) \) as the initial point.

At each step of the descent \( (j) \), let \( I_s, y_i \) be the nearest neighbor of \( I_u \) with a different severity level from \( I_u \), and let \( K_s, y_i \) be the nearest neighbor of \( I_u \) with the same severity level as \( I_u \), such that \( D^d(I_u, K_s, y_i) \leq D^d(I_u, I_s, y_i) \). One way to increase severity assessment performance for \( I_u \) is to minimize the ratio between \( S^d(I_u, K_s, y_i) \) and \( O^d(I_u, I_s, y_i) \). To respect the continuity constraint, each coefficient \( x \) in \( (V(I_u), H(I_u), N_3(I_u), W(I_u)) \) was increased by \( \Delta x^d \) at each step \( (j) \):

\[ \Delta x^d(I_u) = - \frac{\partial S^d(I_u)}{\partial x} \bigg|_{I_u} - \sum_{I_0} e^{-\lambda_{i}d(I_0) - d_s(i) / \| \partial x O^d(I_u) \bigg|_{I_u}} \quad (7) \]
where \(\lambda > 0\), which controls the regularity of the mapping functions, was obtained by cross-validation on the training subset; the same value for \(\lambda\) was used in equation 6. If \(x\) is a weight (respectively a wavelet coefficient), then equation 7 can be computed using equation 4 (respectively, equation 4 and \(|d|\)).

### d. Derivatives of the Proposed Image Characterizations with Respect to Wavelet Filter Coefficients

Let \((k,l)\) be the coordinates of one coefficient in \(w(k \in \{-K,\ldots,K\}, l \in \{-L,\ldots,L\})\) with respect to which we would like to compute the derivative of \(I\)'s characterization (see equation 3).

Let's first compute the derivative of the \(d\)th order moment \(m_{s,k,l,d}(I)\) of the distribution (see equation 3) with respect to \(w_{k,l}\) (see equation 2):

\[
\frac{\partial m_{s,k,l,d}(I)}{\partial w_{k,l}} = \frac{d}{MN} \sum_{i,k,j} (l_{i,k,j} - \mu_s) \frac{d}{I_l} \left( \frac{d}{\sigma_{s,k,l}} \right)
\]

The derivative of \(I\)'s characterization, with respect to \(w_{k,l}\), is given by:

\[
\begin{align*}
\frac{\partial \sigma_{s,k,l}}{\partial w_{k,l}}(I) &= \frac{1}{2 \sigma_{s,k,l}} \frac{\partial m_{s,k,l}}{\partial w_{k,l}}(I) \\
\frac{\partial K_{s,k,l}}{\partial w_{k,l}}(I) &= \frac{1}{\sigma_{s,k,l}^2(I)} \left( \frac{\partial m_{s,k,l}}{\partial w_{k,l}}(I) - 2 \frac{m_{s,k,l}}{\sigma_{s,k,l}^2(I)} \frac{\partial m_{s,k,l}}{\partial w_{k,l}}(I) \right)
\end{align*}
\]

### References