# Automated Classification of Severity of Age-Related Macular Degeneration from Fundus Photographs

*Srihari Kankanahalli,*[1] *Philippe M. Burlina,*[1–3] *Yulia Wolfson,*[3] *David E. Freund,*[1] *and Neil M. Bressler*[3]

**Purpose.** To evaluate an automated analysis of retinal fundus photographs to detect and classify severity of age-related macular degeneration compared with grading by the Age-Related Eye Disease Study (AREDS) protocol.

**Methods.** Following approval by the Johns Hopkins University School of Medicine's Institution Review Board, digitized images (downloaded at http://www.ncbi.nlm.nih.gov/gap/) of field 2 (macular) fundus photographs from AREDS obtained over a 12-year longitudinal study were classified automatically using a visual words method to compare with severity by expert graders.

**Results.** Sensitivities and specificities, respectively, of automated imaging, when compared with expert fundus grading of 468 patients and 2145 fundus images are: 98.6% and 96.3% when classifying categories 1 and 2 versus categories 3 and 4; 96.1% and 96.1% when classifying categories 1 and 2 versus category 3; 98.6% and 95.7% when classifying category 1 versus category 3; and 96.0% and 94.7% when classifying category 1 versus categories 3 and 4;

**Conclusions.** Development of an automated analysis for classification of age-related macular degeneration from digitized fundus photographs has high sensitivity and specificity when compared with expert graders and may have a role in screening or monitoring. (*Invest Ophthalmol Vis Sci.* 2013; 54:1789–1796) DOI:10.1167/iovs.12-10928

Age-related macular degeneration (AMD) is the leading cause of blindness throughout much of the Western world for individuals older than 50 years of age.[1] Vision loss can occur from the advanced stage, which includes choroidal neovascularization (CNV) or geographic atrophy involving the center of macula. Left untreated, the advanced stage can lead to severely impaired central vision, influencing everyday activities.[2] In the United States, approximately 200,000 individuals older than 50 years of age develop the advanced stage of AMD each year in at least one eye.[3] Left untreated, approximately 70% of these cases develop substantial vision loss in the affected eye within 2 years. Furthermore, of those patients who developed advanced AMD in only one eye, approximately half will develop the advanced stage in the other eye within 5 years, resulting in a high risk of developing legal blindness if left untreated.[1]

Although there is no definitive cure for AMD, the Age-Related Eye Disease Study (AREDS) has suggested benefits of certain dietary supplements for slowing the progression of the disease from the intermediate stage to the advanced stage.[4] In addition, recent clinical trials of anti–vascular endothelial growth factor (VEGF) for treating CNV can eliminate a substantial proportion of cases that otherwise would progress to the advanced stage.[5] The better the visual acuity at the onset of anti-VEGF therapy, the greater is the chance of avoiding substantial visual acuity impairment or blindness.[2] Thus, it is critical to identify in a timely manner those individuals most at risk for developing advanced AMD, specifically individuals with the intermediate stage of AMD.

The following drusen classification method was adopted by the AREDS Coordinating Centers[6]: large drusen are defined as those that exceed 125 microns in diameter (the average size of a retinal vein at the optic disk margin), small drusen are defined as those with diameters less than 63 microns, and medium-sized drusen are defined as those with diameters in the range between 63 and 125 microns. The intermediate stage of AMD is characterized by the presence of numerous medium-sized drusen, or at least one large druse within 3000 microns of the center of the macula (Fig. 1). Although a dilated ophthalmoscopic examination at least every 2 years to detect asymptomatic conditions potentially requiring intervention, such as the intermediate stage of AMD, is recommended by the American Academy of Ophthalmology, the presence of drusen often causes no symptoms and therefore no motivation for an individual to seek examination by an ophthalmologist.

Currently, ophthalmoscopy of the retina by trained health care providers or evaluation of fundus photographs by trained graders remains the most effective method to identify the intermediate stage of AMD.[1] However, grading fundus images manually by a grader can be a tedious process, requiring the expertise of an adequately trained health care provider or extensively trained fundus photograph grader to understand the varying patterns recognized by an ophthalmologist.[7] Furthermore, access to an ophthalmology health care provider at least every 2 years to detect the intermediate stage of AMD after 50 years of age can be challenging for many health care environments. Therefore, there is a need for automated visual diagnostic tools that allow the detection of the intermediate stage AMD among a large pool of the at-risk population. As an example of the potential health care burden of this issue, in 2010, in the United States, there were approximately 98 million individuals older than 50 years of age and this number is projected to increase to approximately 109 million by 2015.[8]

From the [1]Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland; the [2]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland; and the [3]Wilmer Eye Institute, Johns Hopkins University School of Medicine, Retina Division, Baltimore, Maryland.

Supported in part by the Johns Hopkins Applied Physics Laboratory under internal research and development funds and unrestricted donations to the Johns Hopkins University School of Medicine.

Submitted for publication September 7, 2012; revised December 18, 2012; accepted January 21, 2013.

Disclosure: **S. Kankanahalli**, None; **P.M. Burlina**, P; **Y. Wolfson**, None; **D.E. Freund**, P; **N.M. Bressler**, P

Corresponding author: Philippe M. Burlina, 11100 Johns Hopkins Road, Laurel MD, 20723-6099; philippe.burlina@jhuapl.edu.

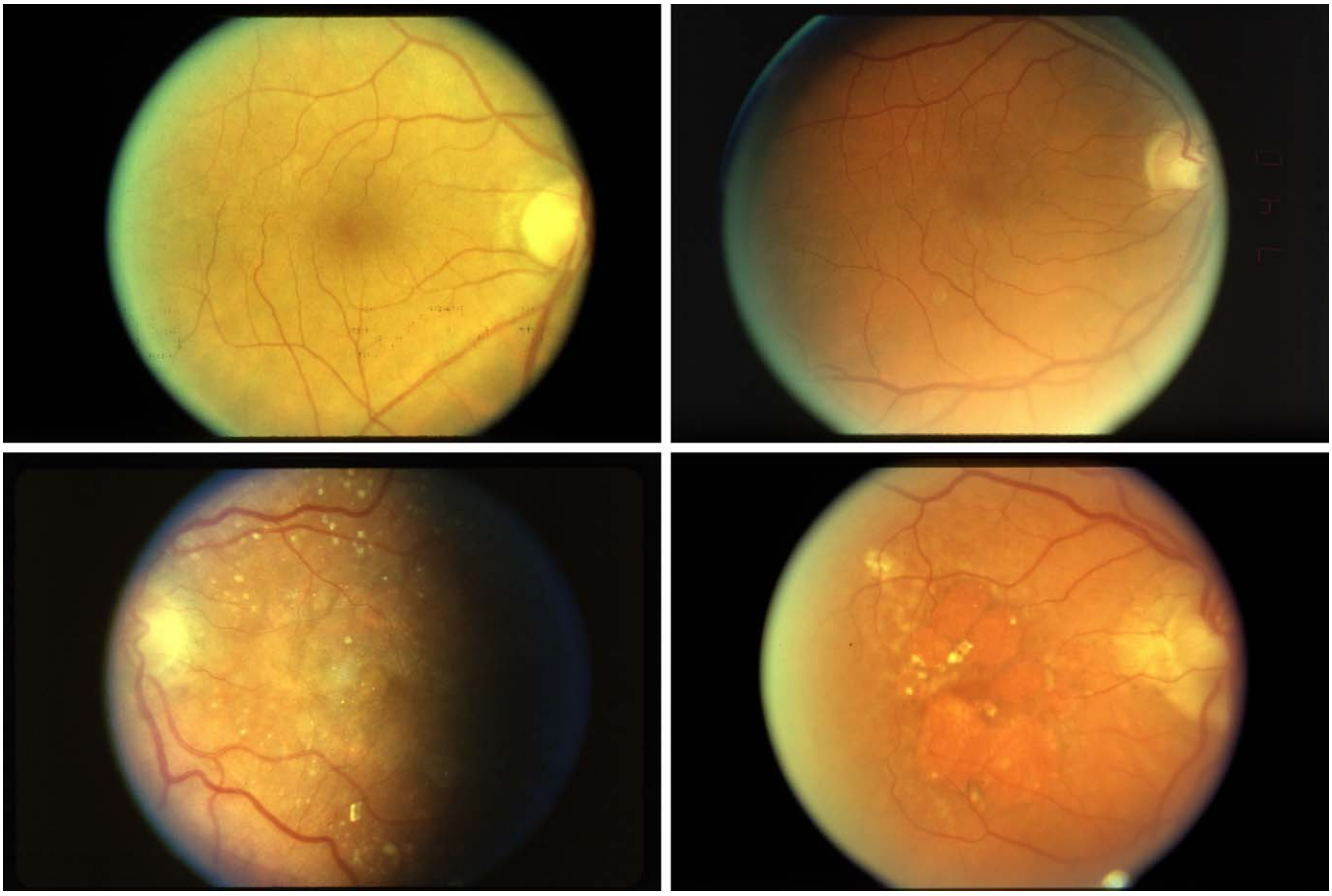

**Figure 1.** Examples of four fundus images with increasing AMD severity. *Top left*: An AMD category 1 (no AMD). *Top right*: An AMD category 2 (early AMD). *Bottom left*: An AMD category 3 (intermediate AMD) with geographic atrophy. *Bottom right*: An AMD category 4 with evidence of both neovascularization and geographic atrophy (advanced AMD). As seen in these images, the background retina can show variations in hue and the retinal fundus images may have various artifacts.

A substantial body of work has been devoted to the design of automated retinal image analysis (ARIA) algorithms. Although ARIA algorithms for diabetic retinopathy or glaucoma are showing promise,[9] less progress, in the opinion of the authors, has been made in the area of AMD. Some AMD detection methods require user intervention.[10] Recently, researchers have emphasized automated approaches by using adaptive equalization and wavelets[11]; applying mathematical morphology[12] on angiographic images; using adaptive thresholding[13]; exploiting probabilistic boosting approaches for the classification of nonhomogeneous drusen textures[14]; using probabilistic modeling and fuzzy logic[15]; applying histogram normalization and adaptive segmentation[16]; exploiting texture discrimination and the intensity topographical profile[17]; utilizing morphologic reconstruction[18]; employing a histogram-based segmentation method[19]; or, finally, using basic feature clustering to find bright lesions.[20] The interested reader is also referred to a recent review[9] of ARIA techniques.

The objective of our study was to develop and assess methods to automatically process fundus images based on a "visual words" approach, in order to reliably detect evidence of AMD as well as accurately categorize its severity. Because the key factor in mitigating the worsening of AMD as it progresses from the intermediate stage to the neovascular form (and potentially, in the future, the geographic atrophic form) is early intervention, the ultimate goal is to implement these algorithms in a public monitoring or screening system that is convenient and easily accessible to the general public. In essence, the system would analyze fundus images of an individual and quickly provide results including a grade of AMD severity and, if necessary, a recommendation to see an ophthalmologist for further evaluation, while avoiding false-positive referrals.

A natural approach for finding and classifying AMD patients consists of automatically finding drusen in fundus images (which is the aim of most of the above-cited studies) and then using this to detect and classify the severity of AMD. This task may be difficult due to variations in patient-specific appearance (variability in pigmentation of the choroid as well as drusen appearance within and across subjects), and it may be challenging to identify stable image features that are characteristic of drusen that can be used to build a robust classifier that will perform reliably over a large data set. Because of this, the current study uses an alternate strategy that focuses on classifying the entire fundus image, as a whole, as opposed to looking only for specific drusen or other lesions.

## Methods

### Algorithm Development

In this section, we first discuss the AREDS fundus images data set (dbGAP) and how the images ultimately were selected for this study. Next, the main steps and techniques of our AMD severity classification algorithm are described in detail.

## AREDS Data Set of Images

Although the AREDS study involved thousands of participants, a subset of 600 subjects was selected by the National Institutes of Health (NIH) for genome studies. These consisted of 200 control patients, 200 neovascular AMD cases, and 200 geographic atrophy AMD cases.[21] These patients were followed over several years (median of 6.5 years and up to a total of 12 years), during which time a number of patients evolved to the more advanced stages. A data set consisting of additional information on these 600 patients, including fundus photographs, was made publicly available by the NIH. This data set is known as the AREDS dbGAP. A set of fundus photographs was digitized from 595 of these 600 patients forming a set of over 72,000 images graded for AMD severity.[21–24] For each patient and for each examination, several fundus photographs were taken of the left and right eyes. In particular, for each eye, left and right stereo pairs were taken for three fields of view: field 1M (centered on the temporal margin of the disc), field 2 (centered on the macula), and field 3M (centered temporal to the macula). Of these 595 participants with available fundus imagery, only 527 consented to reveal their AMD categories. Of all the retinal images available for these 527 participants, only field 2 images were used in our study because they are centered on the macula, resulting in 11,344 images in all. From all these images, an additional selection process (explained in the following text) was used to create a subset of good quality images for our study. This resulted in 4205 images, corresponding to 476 unique patients from the aforementioned 527 patients. From this set of images, when two stereo images of the same eye of the same patient on the same visit were present, we kept only the one classified as better quality (to remove what was for our purposes essentially redundant data). This resulted in a final number of 2772 images that were used in our study to train and test our AMD severity classifier algorithm. It should be noted that the good quality image selection step may have eliminated one right image, but not the corresponding left image of a stereo pair for some eyes, or vice versa (i.e., the right stereo image may have been classified as bad quality but the left was classified as good, or vice versa). Consequently, it is possible for the final number of images (2772) to be greater than half of the set from which it was derived (4205). The above-mentioned numbers of images and corresponding patients are summarized in Table 1.

In addition to the field 2 images, database tables provided by the NIH list the level or category of AMD severity associated with each image. Specifically, each image is assigned an AMD category from 1 to 4, with category 1 representing images showing minimal to no evidence of AMD, category 2 corresponding to the early stage of AMD,[25] category 3 corresponding to the intermediate stage of AMD, and category 4 representing images from patients with the advanced stage of AMD. Figure 1 shows a typical example for each of the four AMD severity categories.

## Description of the Classification (Algorithm) Approach

**Method Motivation and Summary.** Our approach to automatically classifying fundus images for AMD severity is built around the concept of visual words, also known as "bag of words."[26] This method was first used in the field of automated text classification. For example, suppose that the problem is to teach a computer to distinguish among newspaper articles on three news categories such as politics, sports, and business. The first step of this method is to determine what the salient words are—that is, the method automatically selects keywords such as "president," "congress," "stocks," "campaign," and "score," based on their importance. Next, a training phase is used in which the algorithm is provided example articles on the three news categories. During the training phase the algorithm is told under which category each article falls, and it infers the relative frequency (histograms) of all selected keywords in each article category. Given a corpus of new and uncategorized articles, the method would then categorize each article by looking at the frequency of each keyword it contains and selecting the category that has the closest histogram. This entire approach can be transposed to the problem of classifying retinal images into different categories of affected eyes, by substituting newspaper articles with fundus images and visual words with visual features computed in these fundus images.

**Table 1.** Number of Images and Corresponding Unique Patients Available in Our Study

| Field + Steps | Patients | Images |
|---|---|---|
| Field 2 | 595 | 12,401 |
| Field 2 (patients consenting to release categories) | 527 | 11,344 |
| Field 2 (postquality check) | 476 | 4,205 |
| Field 2 (postelimination of stereo pairs) | 476 | 2,772 |

Number of images and corresponding unique patients available in our study after (1) considering only patients who consented to release their AMD categories, (2) additional selecting for good quality images, and (3) removing redundant stereo image from possible stereo image pairs.

**Salient Visual Features.** Recently the visual words approach has been adapted by the computer vision community[26] to perform classification of images. As noted earlier, when used in this context, "salient visual features" take on the role of newspaper articles' keywords. Such features can be automatically characterized using robust feature detection methods such as Scale Invariant Feature Transform (SIFT) or Speeded Up Robust Feature (SURF). After all visual features in a set of training images have been detected, a K-means clustering approach is used to find centroids of the features. K-means clustering is a classical technique used to partition a data set into K different clusters and to find the clusters' centroids. The method is an iterative process that alternates between (1) ascribing data points to the clusters' centroids and relabeling them accordingly and (2) recomputing the clusters' centroids given the newly formed clusters.[27] Next, the method reduces the number of all selected visual features across all training images to a smaller, user-specified number of representative features, forming a data set of so-called visual words. The set of training images is used once again to find the relative frequency of each of the visual words from images of each AMD category, forming prototypical visual word histograms that are characteristic of each AMD image category. As described earlier, any new test image is then simply classified as follows: salient visual features are detected, a histogram is created, and the image is ascribed to the category whose visual word histogram most closely matches the visual word histogram of the test image. Other than selecting the number of visual words and providing images for training, the method does not need any additional input or supervision and is agnostic to the type of category or classification to which it is applied.

**Preprocessing to Obtain Region of Interest.** Retinal images almost always have a black border that needs either to be avoided or eliminated. Within the AREDS database there are many images where the macula is off center, the border is lighter than pure black from flash or other photographic artifacts, red timestamps are placed on the border, or other artifacts are present besides the desired retinal area (see Fig. 2). To accurately and consistently obtain the region of interest (ROI), the following steps are used: (1) the green channel of the red–green–blue (RGB) image is extracted and, to improve speed, resized to one-eighth size. (2) A $9 \times 9$ median filter is applied and then a binary image is created by thresholding (i.e., pixels above a prescribed value are set to 255 and those below to 0). (3) Next, a morphologic opening and closing is applied with a round $3 \times 3$ structuring element, to eliminate background noise, timestamps, and other artifacts that are sometimes present in the AREDS images. (4) The minimum enclosing circle around all of the remaining points is found, and the inscribed square within that circle becomes the ROI where the rest of the algorithm is applied. (5) The full image is then cropped to this square region and, to minimize processing time, resampled down to $700 \times 700$ resolution.

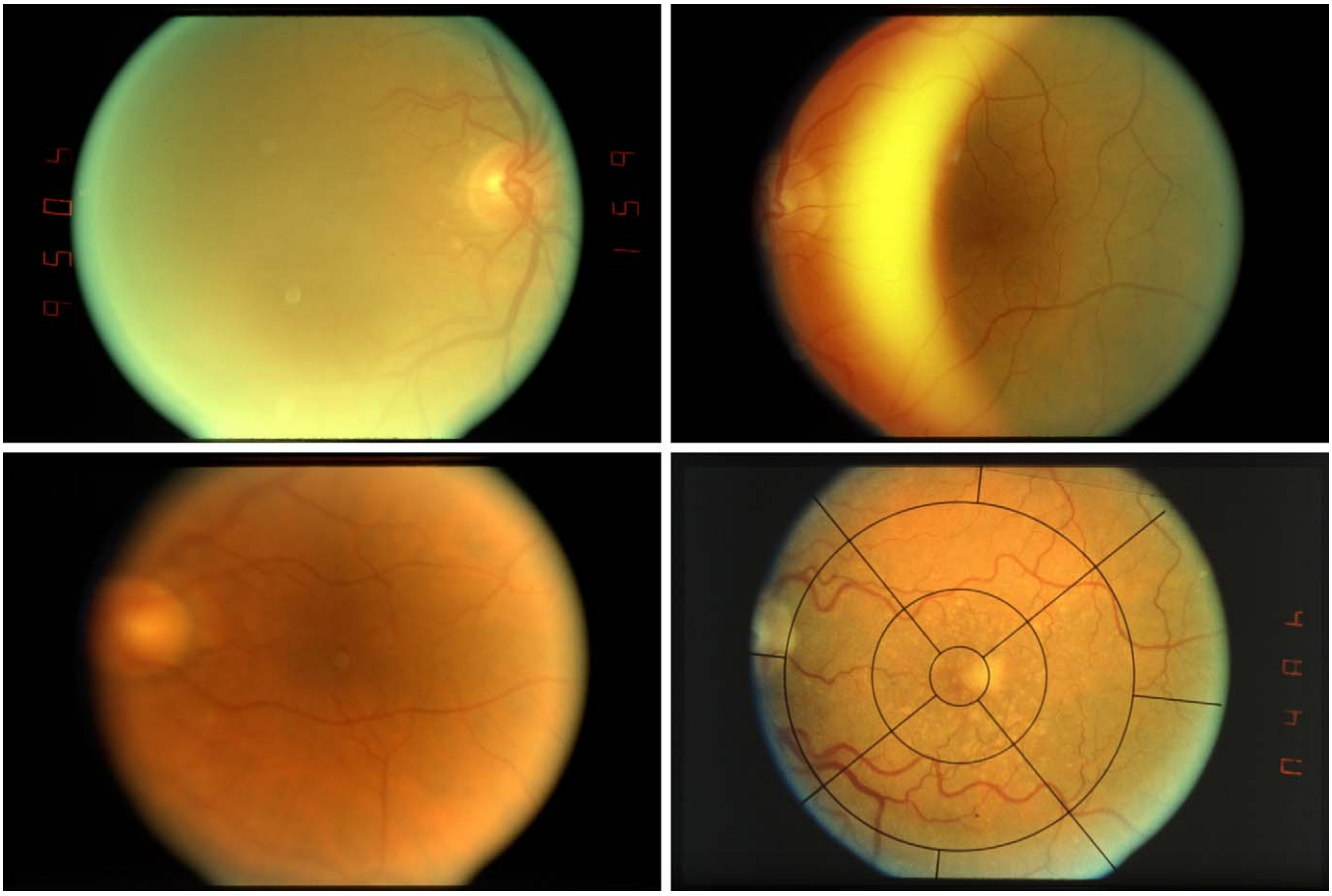

**Figure 2.** Examples of poor quality images: (*top left*) poor media, possible inadequate dilation, (*top right*) lateral misalignment, (*lower left*) poor focus, possible retinal layer separation (not the same focal plane for the lower arcade and the optic disc), and (*lower right*) grid still attached to fundus photograph.

**Preprocessing to Remove Large Background Intensity Gradient.** Images are affected by various degrees of intensity gradient variation that depends on the acquisition conditions. To remove this background intensity gradient, a new image is created by the following: (1) smoothing the green channel with a large median filter set to one-fourth of the image width, (2) subtracting the median filtered image from the original image's green channel, and (3) multiplying the result by 2 (to increase contrast) and adding 127.

**Keypoint Detection and Feature Vector Descriptor Extraction.** To find keypoints (i.e., visual salient features), a SURF detector is used on the image resulting from the previous step. The SURF algorithm exploits a technique known as integral images to quickly find the image second derivative (the Hessian) and apply approximate Gaussian filters at multiple scales. Each scale is known as an octave, and each application of the Gaussian filter forms one layer. Here, keypoints are detected using ten octaves, three layers per octave, and a Hessian threshold of 600. The original image is then converted from the RGB color space to the L*a*b* color space[28] and a SURF descriptor for every keypoint is then computed for each L*a*b* channel. These three descriptors are then concatenated into one. This aids in classification because color is an important visual cue in finding retinal anomalies. Briefly, we remark here that the L*a*b* color space is more representative of the way humans see and has several technical advantages over the more traditional RGB or HSV (i.e., hue, saturation, and value) spaces. In particular, metrics for measuring distance between colors are essentially Euclidian. Furthermore, tones (Lightness) and colors (the a channel is green or magenta hue and the b channel is blue or yellow hue) are held separately; thus, one can vary one without altering the other.[28]

**Vocabulary Creation.** A vocabulary of visual words is created from the keypoint descriptors of the set of AMD-positive (i.e., categories 3 and 4) images. The rationale for not including all images is that AMD-positive images contain all features (vessels, optical disk, artifacts) that are present in AMD-negative images plus drusen and other lesions (geographic atrophy, pigmentation, and other features). A vocabulary of 100 visual words for two-class classification problems is used, and 300 visual words for three-class problems. The visual words are selected as the centroids found using K-means clustering. We emphasize here that this needs to be done only once; the same vocabulary is used for each subsequent run.[26]

**Spatially Dependent Histogram Generation.** To reflect the fact that AMD severity is graded by taking into account the location of the drusen and other AMD-related lesions, with the macular region taking a preponderant weight in the decision process (see Fig. 1 in AREDS Report No. 6[25]), the feature selection is region dependent. This is based on subdividing the fundus image in predefined concentric regions. Several options are considered and compared in the Results section (see Fig. 3). Consequently, based on their distance from the center of the image (which corresponds approximately to the macula in most images), feature descriptors are grouped into several different sets and importance weights are applied to the histograms of each region based on distance of the region from the center, to emphasize regions close to the macula. Regional histograms are then concatenated back into a single large histogram for the entire image. This concatenated vector forms the final "feature vector" used for classification.

**Training and Testing.** The entire corpus of available images and their associated category labels is then used for training and testing. For

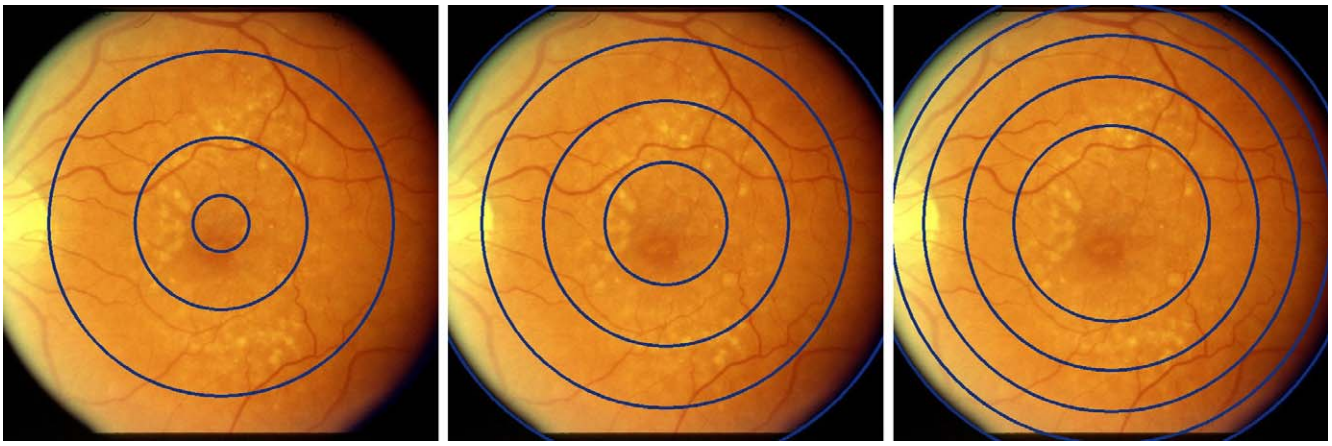

**Figure 3.** The three grids tested: (*left*) grid based on AREDS specifications, (*center*) grid with equally spaced circles, and (*right*) custom grid with a large central circle near the macula.

each image, a final feature vector (visual word histogram) is generated once and for all. As is standard in machine learning applications, an N-fold cross-validation approach is used. This consists of subdividing the data set into N equally sized folds (i.e., subsets), using $N - 1$ folds for training, and the remaining Nth fold for testing. Then, a random forest classifier is trained using the training data set. The random forest algorithm uses the consensus of a large number of *weak* (only slightly better than chance) binary decision trees to classify the testing images into different severity classes.[29] For the two-class problems the random forest consisted of 1000 decision trees, whereas for the three-class problem it consisted of 2500.

## Results

### Classifications Problems

For the purpose of this study, we evaluated the performance of the algorithm based on several two-class problems among the various AMD categories: (1) {1 & 2} vs. {3 & 4}; (2) {1 & 2} vs. {3}; (3) {1} vs. {3}; (4) {1} vs. {3 & 4}, and one three-class classification problem: {1 & 2} vs. {3} vs. {4}. These problems were structured to discriminate between intermediate stage individuals, for whom treatment would help to maintain vision at a useful level, and other individuals that were either not at risk or too advanced. It was judged appropriate to cluster category 2 with category 1 since both categories have little clinical relevance to the risk of the need for monitoring for the advanced stage of AMD or the need to consider dietary supplements compared with either category 3 or category 4.

### Selection of Good Images

As is customary in ARIA processing for detection of AMD or diabetic retinopathy, a first step is applied to select good quality images.[30,31] From the initial data set, only "good quality" images are retained. In our study, this was performed in two ways: manually and automatically. The automated selection of good quality images is essentially a problem of classifying images into *good* versus *poor* images. For this classification, we have also used a visual words approach that is essentially identical to the approach we reported earlier to classify fundus images into different levels of AMD severity, minus the spatial histogram generation step (since, unlike the AMD classification problem, the location of certain image features does not generally correlate with its quality). As should be noted, *good* versus *poor* image classification was also used in the study reported by Niemeijer et al.[31] We evaluated the performance of our *good* versus *poor* image classifier on a data set of 400 AREDS images that were manually ground-truthed as either "good" or "bad" using a 5-fold performance cross-validation. This approach resulted in a specificity of 93.0%, a sensitivity of 94.5%, a positive predictive value (PPV) of 93.1% and a negative predictive value (NPV) of 94.4%, and an accuracy of 93.8%.

As shown in Table 2, the number of patients and images available for training and testing purposes is unequal among the four AMD categories. Because of this, the following cohorts were considered:

**Data Set with Maximum Number of Images per Class (Denoted MIPC).** This is a subset of automatically selected

**Table 2.** Number of Images and Corresponding Unique Patients in Each Data Set

| Set | Number of Unique Patients | Number of Total Images | Images in Category 1 | Images in Category 2 | Images in Category 3 | Images in Category 4 |
|---|---|---|---|---|---|---|
| EIPC† | 468 | 2145 | 626 | 89 | 715* | 715* |
| MIPC‡ | 476 | 2772 | 626 | 89 | 1107 | 950 |
| MS§ | 236 | 384 | 180 | 13 | 113 | 78 |

* Depending on the test, this number may be lower to keep each "total class" equal. For example, in the problem of classifying categories {1} vs. {3} only 626 images would be used (selected randomly) from category 3, since the maximum number of images for category 1 is 626. Similarly, in the test {1 & 2} vs. {3 & 4}, 715 images would be selected randomly from categories 3 and 4, since the maximum number of images for categories 1 and 2 combined is 715 (a number lower than the maximum number of images for categories 3 and 4 combined, 2057).

† Data sets with equal number of images per class.

‡ Data sets with maximum number of images per class.

§ Manually selected data sets.

**Table 3.** Comparing Performance Results for the Three Grid Schemes Shown in Figure 3 for the Problem of Classifying Categories {1 & 2} vs. {3 & 4}, and Using EIPC

| Grid | Specificity | Sensitivity | PPV | NPV | Accuracy |
|---|---|---|---|---|---|
| AREDS | 92.3% | 91.3% | 92.3% | 91.4% | 91.8% |
| Regular | 92.9% | 91.3% | 92.8% | 91.5% | 92.1% |
| Broad | 95.4% | 95.5% | 95.4% | 95.5% | 95.5% |

images of good quality where the number of images in each AREDS category was kept as large as possible.

**Data Set with Equal Number of Images per Class (Denoted EIPC).** This is a subset of automatically selected images of good quality where the numbers of images in each AREDS category was kept equal. Depending on the test, this number may be lower to keep the "total number in each class" equal. For example, in the test {1} vs. {3} only 626 images would be used (selected randomly) from category 3, since the maximum number of images for category 1 is 626. Similarly, in the test {1 & 2} vs. {3 & 4} 715 images would be selected randomly from categories 3 and 4, since the maximum number of images for categories 1 and 2 combined is 715 (a lower number than the maximum number of images for categories 3 and 4 combined, 2057.)

**Data Set with Manually Selected Images (Denoted MS).** This is a much reduced subset of images that were selected manually, deemed to be of "good quality," without an attempt at a full search or optimization.

## Sensitivity and Specificity of Automated Classification Compared with Expert Fundus Grading

The number of true positives (denoted TP), false positives (FP), true negatives (TN), and false negatives (FN) using our automated AMD severity classification method was compared with the expert fundus grading provided in the NIH database with respect to the following:

*Sensitivity* (also called probability of detection) is defined as TP/(TP + FN) (i.e., percentage of retinas correctly identified as having the AMD category of interest as determined by the expert fundus grading); *specificity* is defined as TN/(TN + FP) (i.e., percentage of retinas correctly identified as not having the AMD category of interest as determined by the expert fundus grading); *positive predictive value* (PPV), the probability that a retina identified as having the AMD category of interest actually has that classification as determined by the expert grading, is defined as TP/(TP + FP), and *negative predictive value* (NPV), the probability that a retina identified as not having the AMD category of interest is indeed not that category as determined by the expert fundus grading, is defined as TN/(TN + FN); and *accuracy*, the total percentage of retinas correctly categorized by the automatic algorithm as categorized by the expert fundus grading, is defined as (TP + TN)/(TP + FP + TN + FN).

**Table 4.** Performance Results for the Various Two-Class AMD Severity Classification Problems

| AMD Category Test | Set | Specificity | Sensitivity | PPV | NPV | Accuracy |
|---|---|---|---|---|---|---|
| {1 & 2} vs. {3 & 4} | EIPC | 95.4% | 95.5% | 95.4% | 95.5% | 95.5% |
| | MIPC | 91.6% | 97.2% | 97.1% | 91.9% | 95.7% |
| | MS | 98.4% | 99.5% | 98.4% | 99.5% | 98.9% |
| {1 & 2} vs. {3} | EIPC | 96.1% | 96.1% | 96.1% | 96.1% | 96.1% |
| | MIPC | 95.7% | 96.0% | 97.2% | 94.0% | 95.9% |
| {1} vs. {3} | EIPC | 98.6% | 95.7% | 98.5% | 95.8% | 97.1% |
| | MIPC | 96.3% | 96.8% | 97.9% | 94.5% | 96.7% |
| {1} vs. {3 & 4} | EIPC | 96.0% | 94.7% | 96.0% | 94.8% | 95.4% |
| | MIPC | 95.4% | 97.7% | 98.6% | 92.3% | 97.1% |

**Table 5.** Performance Results for the Three-Class AMD Severity Classification Problem

| AMD Category Test | Set | Accuracy | Confusion Matrix | | |
|---|---|---|---|---|---|
| {1 & 2} vs. {3} vs. {4} | EIPC | 91.8% | 89.23% | 7.27% | 3.50% |
| | | | 11.33% | 86.43% | 2.24% |
| | | | 0.28% | 0.00% | 99.72% |
| | MIPC | 90.2% | 83.78% | 3.36% | 12.86% |
| | | | 2.80% | 90.69% | 6.51% |
| | | | 1.47% | 3.79% | 94.74% |

Results obtained for the different regional retinal division schemes (as discussed in the histogram generation step of the algorithm) were compared. The AREDS regions are based on the scheme developed by the AREDS group.[25] It contains regions with radii equal to 1/3, 1, and 2 times the disk diameter. Alternatively, an equally spaced regional division was tested along with an unequal (termed "broad") division that emphasized the size of the central region (as opposed to the smaller and more focused central region used in the AREDS division scheme). As seen in Table 3, the best performance was obtained for the "broad" subdivision. By design, the field 2 images are centered near or on the macular region. The "broad" approach guarantees the consistent inclusion of the macular region without having to determine exactly the macular location.

Results of the two-class classification problems are shown in Table 4, whereas Table 5 provides additional results in the form of a confusion matrix for the {1 & 2} vs. {3} vs. {4} classification problem. As can be seen from Table 4, the best results were obtained when image quality was determined manually (i.e., MS) as opposed to automatically. This underscores the importance of obtaining high quality fundus images. Comparison between EIPC and MIPC results did not show either approach as being clearly superior. Nevertheless, overall, EIPC did somewhat better than MIPC. Table 6 shows the results for the three-class classification test. Again, the EIPC approach performed slightly better than MIPC, although it is worth noting that MIPC outperformed EIPC in classifying category 3.

## Discussion

Our method for automatic colored fundus photography AMD severity classification was developed and tested on the available pool of digitized fundus photographs. Overall, the results are very promising across different classification

**Table 6.** Individual Class Performance Results for the Three-Class AMD Severity Classification Problem

| Set | Class | Specificity | Sensitivity | PPV | NPV | Accuracy |
|---|---|---|---|---|---|---|
| EIPC | 1 & 2 | 94.2% | 89.2% | 88.5% | 94.6% | 92.5% |
| | 3 | 96.4% | 86.4% | 92.2% | 93.4% | 93.1% |
| | 4 | 97.1% | 99.7% | 94.6% | 99.9% | 98.0% |
| MIPC | 1 & 2 | 97.8% | 83.8% | 93.0% | 94.5% | 94.2% |
| | 3 | 96.4% | 90.7% | 94.4% | 94.0% | 94.1% |
| | 4 | 91.0% | 94.7% | 84.6% | 97.1% | 92.3% |

problems, consisting from four different categories of AMD either unmixed or grouped together, with some details and variations that are discussed in the following text.

As observed in the previous section, and as should be expected, the algorithm performs best when applied to *good quality* images (see Table 4). As is often the case in ARIA methods, a good quality image selection step was first used prior to retinal image analysis. Although image quality was rated by the AREDS reading centers, this information was not made available through the NIH publicly accessible database. Consequently, in our study, an automated algorithm was developed to select good quality images. In addition, a manual selection process was also considered, which allowed us to assess the performance of the automated *good quality* selection algorithm. Since the number of manually selected images was sufficient for the problem of classifying categories {1 & 2} vs. {3 & 4}, the performance results were also computed and reported for that problem in Table 4. These results reveal that, although the best performance was obtained when image quality was determined manually (see Table 4, where MS has 98.9% accuracy for classifying categories {1 & 2} vs. {3 & 4}), performance for automatically selected images was also very good (accuracy equal to 95.5% for classifying categories {1 & 2} vs. {3 & 4} and 97.1% for {1} vs. {3}, with the automated quality selection using EIPC). As it stands, our automatic good quality image selection algorithm makes some errors, and with a specificity of 93.0%, 7% of the images automatically classified as *good quality* may actually be *poor quality* images. We would therefore expect our performance to improve if the quality selection is improved as well, which is a possible area of future work.

It should be noted that the digitized images in the AREDS study were not created directly in a digitized format but instead were obtained by digitizing a Kodachrome or Ektachrome (both Kodak, Rochester, NY, but now discontinued), or an approved equivalent slide transparency film. As a result, even images that were deemed of good quality are probably not as good as the original film photographs, or as can be obtained using fundus images that are captured in native digitized format.

The increased use of native digital fundus images raises the issue of whether differences between the method of image capture might alter the performance of our automated AMD classification approach. As noted in the study reported by Hubbard et al.,[32] the quality of the best digital images matches that of the best digitized film images. Nevertheless, it was also noted that natively digitized images tend to exhibit more variability in brightness, contrast, and color balance when compared with images that are originally captured on film. Some of the fundamental reasons noted include the following: (1) digital cameras may have a stronger response to red than to green, which can lead to images that have a strong red component compared with green; (2) digital cameras may have a narrower dynamic range than film cameras, which makes it more difficult to find the best illumination level; or (3) the response profile of a digital camera is linear, whereas film is nonlinear (i.e., curved). As a result, in film, the red response is damped as illumination increases, whereas the digital image saturates. In spite of this, we believe native digital imagery would not present problems that could not be addressed with minor variations in the preprocessing phase of our algorithm (e.g., dealing with possible oversaturation and histogram equalization). Because of the inherent agnostic nature of the underlying visual words technique we have used, and the fact that it adapts to the underlying corpus of training images, we believe our method would perform well with natively digitized images. This belief, however, remains to be tested and is planned as future work. The newer AREDS 2 data set (not yet available to the public) is made up of both film and native digital imagery in approximately a 1:3 ratio[32] and clearly presents an opportunity to test the robustness of our algorithm in the future.

Table 3 demonstrates the importance of grid selection to algorithm performance. The grids considered are shown explicitly in Figure 3. The summary of the results in Table 3 shows that grids with larger central regions yield better algorithm performance. Indeed, these grids are more immune to issues associated with images not being exactly centered. The grids shown in Figure 3 are the only ones we have considered at this time. Nevertheless, additional future efforts could be directed at finding an optimum grid design that captures the clinically relevant aspects of AMD severity as closely as possible. However, it should be kept in mind the data for grading have been previously well established for the standard AREDS grid in multiple clinical trials, possibly making this specific grid more relevant for future possible clinical applications.

Overall, results show good performance for the three-class problem classification with some small differences, depending on the image partitioning, which may be due to the fact that a bias may arise when the maximum number of exemplars is used. As noted previously, the EIPC data selection did somewhat better than did MIPC. Table 5 (confusion matrix) and Table 6 (metrics) show individual results for the three-class problem {1 & 2} vs. {3} vs. {4}. In the future, more efforts will be directed toward automated AMD severity classification for three or more classes.

Finally, we believe that, to the best of our knowledge, the tests we have conducted have been done on the largest data set used thus far with respect to automated classification of AMD severity. Because of this, the results obtained are promising in regard to applying such an algorithm over a larger population data set of the same type as the AREDS data set, specifically for a population composed of individuals of different ethnicities between 50 and 80 years of age.

In conclusion, we developed a method based on visual words to automatically classify the severity of AMD from fundus retinal images and focused on discriminating category 3 AMD (the intermediate stage of AMD) versus the early stage of AMD or no AMD or both. The algorithm was tested on the AREDS image data set. The resulting performance metrics show good promise for the application of the algorithm over fairly large data sets, which is important from a health policy perspective.

## *References*

1. Bressler NM. Age-related macular degeneration is the leading cause of blindness. *JAMA*. 2004;291:1900–1901.
2. Bressler NM, Chang TS, Fine JT, Dolan CM, Ward J. Improved vision-related function after ranibizumab vs photodynamic therapy: a randomized clinical trial. *Arch Ophthalmol*. 2009; 127:13–21.
3. Chang TS, Bressler NM, Fine JT, Dolan CM, Ward J, Klesert TR. Improved vision-related function after ranibizumab treatment of neovascular age-related macular degeneration: results of a randomized clinical trial. *Arch Ophthalmol*. 2007;125:1460–1469.
4. Age-Related Eye Disease Study (AREDS) Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Arch Ophthalmol*. 2001;119:1417–1436.
5. Bressler NM, Bressler SB. Photodynamic therapy with verteporfin (Visudyne): impact on ophthalmology and visual sciences. *Invest Ophthalmol Vis Sci*. 2000;41:624–628.

6. Age-Related Eye Disease Study (AREDS) Research Group. Definitions of final age-related macular degeneration (AMD) phenotype categories. Available at: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd001138. Accessed December 12, 2012.
7. Scotland GS, McNamee P, Fleming AD, et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol*. 2010;94:712–719.
8. US Department of Commerce, US Census Bureau. *The 2012 Statistical Abstract: The National Data Book*. Available at: http://www.census.gov/compendia/statab/2012/tables/12s0009.pdf. Accessed December 14, 2012.
9. Abramoff MD, Garvin MK, Sonka M. Retinal imaging and image analysis. *IEEE Rev Biomed Eng*. 2010;3:169–208.
10. Smith RT, Nagasaki T, Sparrow JR, Barbazetto I, Klaver CC, Chan JK. A method of drusen measurement based on the geometry of fundus reflectance. *Biomed Eng Online*. 2003;2:10.
11. Brandon L, Hoover A. Drusen detection in a retinal image using multi-level analysis. In: Ellis RE, Peters TM, eds. *Medical Image Computing and Computer-Assisted Intervention: MICCAI* 2003. Proceedings of the 6th International Conference, Montreal, Canada, November 2003. Berlin: Springer Verlag; 2003:618–625.
12. Sbeh B, Cohen ZB, Mimoun LD, Coscas G, Soubrane G. An adaptive contrast method for segmentation of drusen. In: *Proceedings 1997 International Conference on Image Processing (ICIP 1997)*; October 26–29, 1997; Washington, DC.
13. Rapantzikos K, Zervakis M, Balas K. Detection and segmentation of drusen deposits on human retina: potential in the diagnosis of age-related macular degeneration. *Med Image Anal*. 2003;7:95–108.
14. Lee N, Laine AF, Smith TR. Learning non-homogenous textures and the unlearning problem with application to drusen detection in retinal images. In: *Proceedings 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*; May 14–17, 2008; Paris, France. 1215–1218.
15. Thdibaoui A, Rajn A, Bunel P. A fuzzy logic approach to drusen detection in retinal angiographic images. In: *Proceedings of the 15th International Conference on Pattern Recognition*; September 3–7, 2000; Barcelona, Spain, Vol. 4:748–751.
16. Checco P, Corinto F. CNN-based algorithm for drusen identification. In: *Proceedings IEEE International Symposium on Circuits and Systems*; May 21–24, 2006; Island of Kos, Greece.
17. Parvathi SS, Devi N. Automatic drusen detection from colour retinal images. In: *Proceedings International Conference on Computational Intelligence and Multimedia Applications*; December 13–15, 2007; Sivakasi, Tamil Nadu.
18. Karnowski TP, Govindasamy V, Tobin KW, Chaum E, Abramoff MD. Retina lesion and microaneurysm segmentation using morphological reconstruction methods with ground-truth data. *Conf Proc IEEE Eng Med Biol Soc*. 2008;2008:5433–5436.
19. Santos-Villalobos H, Karnowski TP, Aykac D, et al. Statistical characterization and segmentation of drusen in fundus images. *Conf Proc IEEE Eng Med Biol Soc*. 2011;2011:6236–6241.
20. Niemeijer M, van Ginneken B, Russell SR, Suttorp-Schulten MS, Abramoff MD. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Invest Ophthalmol Vis Sci*. 2007;48:2260–2267.
21. National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS). Available at: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1. Accessed December 12, 2012.
22. Ferris FL, Davis MD, Clemons TE, et al. A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. *Arch Ophthalmol*. 2005;123:1570–1574.
23. Davis MD, Gangnon RE, Lee LY, et al. The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17. *Arch Ophthalmol*. 2005;123:1484–1498.
24. The Age-Related Eye Disease Study (AREDS). Design implications. AREDS report no. 1. *Control Clin Trials*. 1999;20:573–600.
25. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. *Am J Ophthalmol*. 2001;132:668–681.
26. Fei-Fei L, Perona PA. Bayesian hierarchical model for learning natural scene categories. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; June 20–25, 2005; San Diego, CA. Vol. 2:524–531.
27. Duda RO, Hart PE, Stork DG. *Pattern Classification and Scene Analysis*. New York: Wiley; 1973.
28. Jain AK. *Fundamentals of Digital Image Processing*. Upper Saddle River, NJ: Prentice Hall; 1989.
29. Breiman L. Random forest. *Machine Learn*. 2001;45:5–32.
30. Giancardo L. *Quality Analysis of Retina Images for the Automatic Diagnosis of Diabetic Retinopathy. Vision and Robotics* (*VIBOT*). Bourgogne, France: Université de Bourgogne; 2008. Thesis.
31. Niemeijer M, Abramoff MD, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Med Image Anal*. 2006;10:888–898.
32. Hubbard LD, Danis RP, Neider MW, et al. Brightness, contrast, and color balance of digital versus film retinal images in the age-related eye disease study 2. *Invest Ophthalmol Vis Sci*. 2008;49:3269–3282.