ARTIFICIAL INTELLIGENCE TECHNIQUES FOR AUTOMATIC SCREENING OF AMBLYOGENIC FACTORS

BY Jonathan Van Eenwyk ME, Arvin Agah PhD, Joseph Giangiacomo MD, and **Gerhard Cibis MD***

ABSTRACT

Purpose: To develop a low-cost automated video system to effectively screen children aged 6 months to 6 years for amblyogenic factors.

Methods: In 1994 one of the authors (G.C.) described video vision development assessment, a digitizable analog video-based system combining Brückner pupil red reflex imaging and eccentric photorefraction to screen young children for amblyogenic factors. The images were analyzed manually with this system. We automated the capture of digital video frames and pupil images and applied computer vision and artificial intelligence to analyze and interpret results. The artificial intelligence systems were evaluated by a tenfold testing method.

Results: The best system was the decision tree learning approach, which had an accuracy of 77%, compared to the "gold standard" specialist examination with a "refer/do not refer" decision. Criteria for referral were strabismus, including microtropia, and refractive errors and anisometropia considered to be amblyogenic. Eighty-two percent of strabismic individuals were correctly identified. High refractive errors were also correctly identified and referred 90% of the time, as well as significant anisometropia. The program was less correct in identifying more moderate refractive errors, below +5 and less than −7.

Conclusions: Although we are pursuing a variety of avenues to improve the accuracy of the automated analysis, the program in its present form provides acceptable cost benefits for detecting ambylogenic factors in children aged 6 months to 6 years.

Trans Am Ophthalmol Soc 2008;106:64-74

INTRODUCTION

In the 1994 *Transactions of the American Ophthalmological Society*, Cibis[1] described a digitizable analog video system combining Brückner pupil red reflex imaging[2,3] with eccentric photorefraction[4] for video screening of young children for amblyogenic factors. In collaboration with computer scientists at the University of Missouri, Columbia,[5-8] an automated pupil imaging capture and identification program has been developed (Figures 1 and 2). The Hirschberg reflex is identified as shown in Figure 3.

This automated program has different levels of confidence for iris and limbus outline. The denser dots (Figure 1) mean a higher confidence in the localization. Missing frames indicate that the program has insufficient confidence or the image is too blurry or fixation is too eccentric as determined by the program.

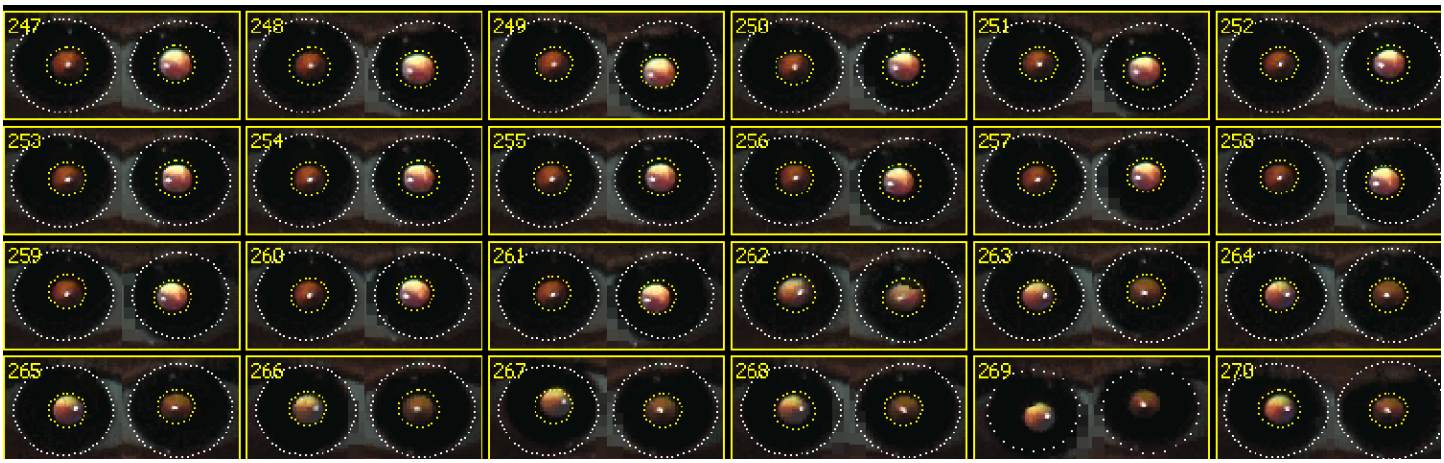

FIGURE 1

Patient with intermittent alternating exotropia. Right eye fixes frames 247 through 261, and left eye fixes frames 263 through 270. The shift takes place in 1/30th of a second (one frame).

Figure 4 demonstrates the principles of eccentric (off-axis) photorefraction. Our camera system mimics the direct opthalmoscope. The light source is eccentric (decentered off axis) below the observer pupil or camera lens aperture. In that case a crescent appears in the patient's pupil superiorly (above) with hyperopia and inferiorly with myopia. Because of chromatic aberration (a two diopter

From the School of Engineering, University of Kansas, Lawrence, (Mr Van Eenwyck and Dr Agah); the Department of Ophthalmology, University of Missouri, Columbia (Dr Giangiacomo); and the Department of Ophthalmology, University of Kansas, Kansas City (Dr Cibis).
*Presenter.
Bold type indicates AOS member.

spread between red and blue wavelength), the inferior myopic crescent begins in the blue spectrum and the superior hyperopic crescent in the red spectrum. If the light is decentered (above the observer pupil or camera lens, as when the ophthalmoscope is turned upside down), the opposite would be true.

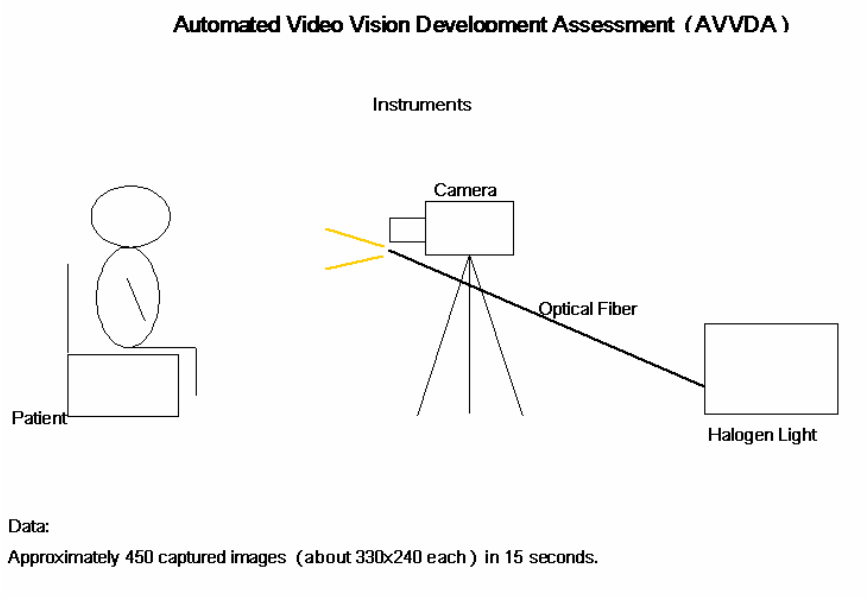

FIGURE 2

Schematic of the instrument/camera−patient relationship.

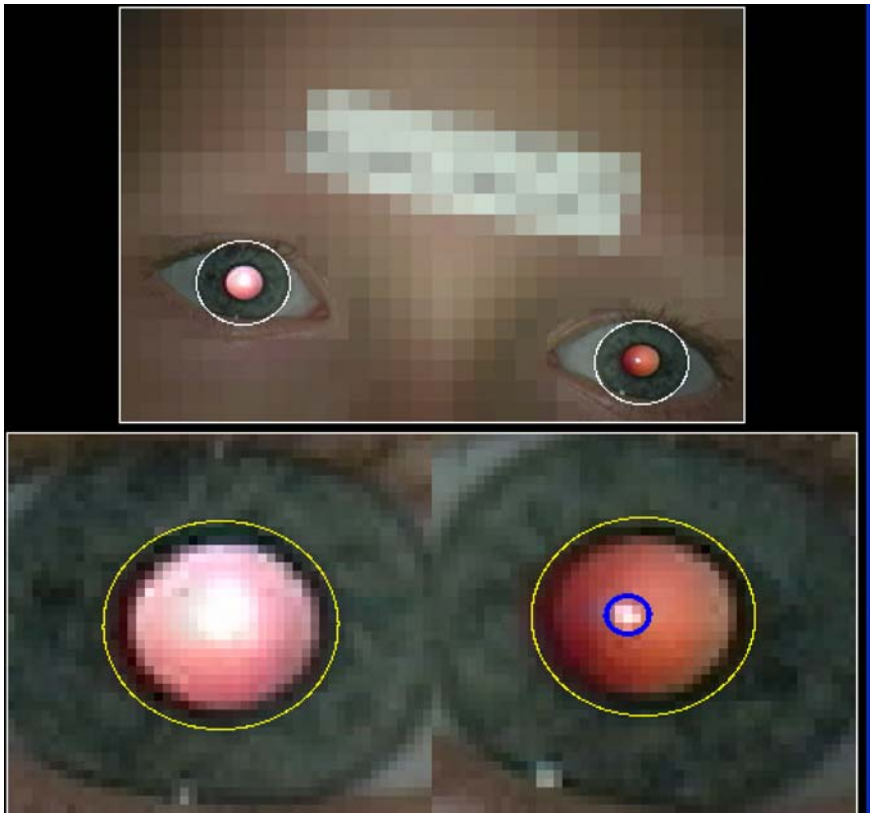

FIGURE 3

Left eye fixing, as evidenced by the darker reflex and proper Hirschberg location. Right esotropic eye shows the brighter reflex. Note identification of Hirschberg reflex OS.

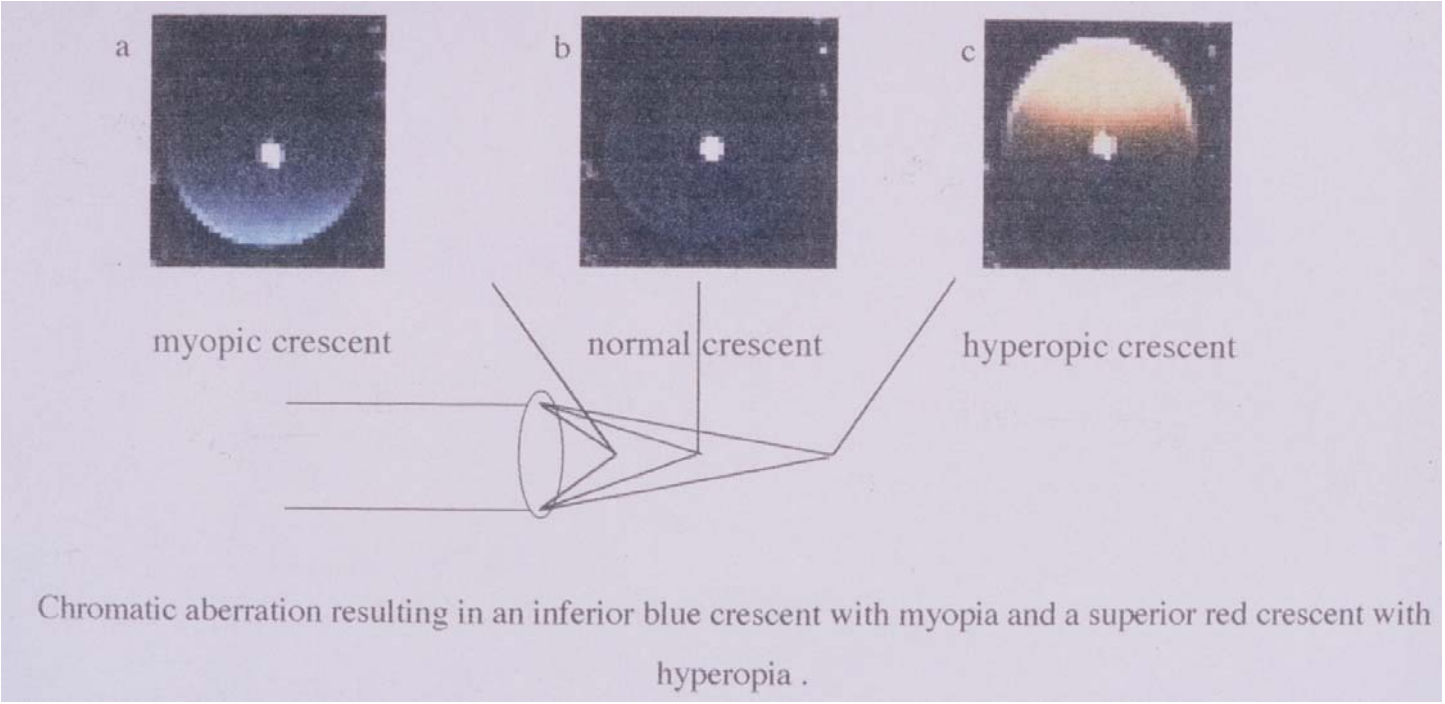

FIGURE 4

Hyperopic crescents appear red from above and myopic crescents appear blue from below. Absence of crescent implies no significant refractive error for that pupil size (dead space).

METHODS

Automated video vision development assessment (AVVDA) is a video system that records multiple frames (30 per second) and therefore captures the images of the pupils both when foveating (ie, on-axis) and when slightly off-axis. Slightly off-axis fixation is what creates the image difference between the two eyes in microtropia. The fixing eye is foveating, and the microtropic eye is fixating slightly off-axis. The deviated eye has a brighter red Brückner reflex than the foveating eye. The differences in brightness and in Hirschberg reflex in strabismic individuals are shown in Figures 1, 3, and 5. Figure 6 is an aligned myope without differences in brightness or Hirschberg reflex location.

FIGURE 5

Patient with intermittent exotropia aligned (left) and deviating left eye (right). Note brightness difference in the deviated left eye (right).

FIGURE 6

Myopic crescent appears inferiorly. Patient is aligned. Note identification of Hirschberg reflex (blue circles).

Because artificial intelligence imaging systems do not measure the crescents directly, they incorporate pupil red reflex information in their analysis in both situations when fixing the light source and when off-axis. Thus when foveating vs nonfoveating frames were separately identified, there was no difference in the various artificial intelligence programs in reaching a correct "refer/do not refer" decision (Table 1).

TABLE 1. ACCURACY WITH ALL FRAMES VERSUS FOVEATING FRAMES*

SYSTEM	ALL FRAMES	FOVEATING FRAMES ONLY
Case-based static	77.9%	77.1%
Case-based fuzzy	75.6%	74.8%
C4.5 eye features	76.9%	74.8%

*Difference was not statistically significant according to McNemar test, presumably because the improvement with better images was offset by the reduced number of images analyzed.

Because a light is not an accommodative target, accommodative esotropic eyes often do not cross on AVVDA examination; instead the accommodative esotropia is identified on the basis of the high plus refractive error. In contrast, eyes with intermittent exotropia may align and fuse with accommodation (Figure 5).

Van Eenwyk,[9] under the direction of Prof Dr Arvin Agah, with clinical input from Dr Cibis, explored other artificial intelligence techniques, besides template matching, to identify children at risk for amblyopia. Patient images were supplied by Dr Cibis from his private practice. These patients had undergone a strabismus, external, and fundus examination performed by Dr Cibis, as well as a cycloplegic refractive "gold standard" examination done by Dr Cibis or an associate.

Because we were developing a screening test, the age-group of interest was limited to patients entering the practice at age 6 months to no older than 6 years. This resulted in the refractive error distribution shown in Figure 7. Myopic patients are woefully underrepresented. The study protocol was approved by the University of Missouri, Columbia, institutional review board.

Four artificial intelligence techniques—case-based reasoning, case-based fuzzy logic, artificial neural networks, and decision tree—were evaluated using a tenfold testing procedure, whereby 90% of the patients were used for training and 10% for testing. The entire patient population (610 patients) was divided into 10 equal-sized groups. As Figure 8 shows, the distribution of hyperopic and myopic patients was not always equal for each group.

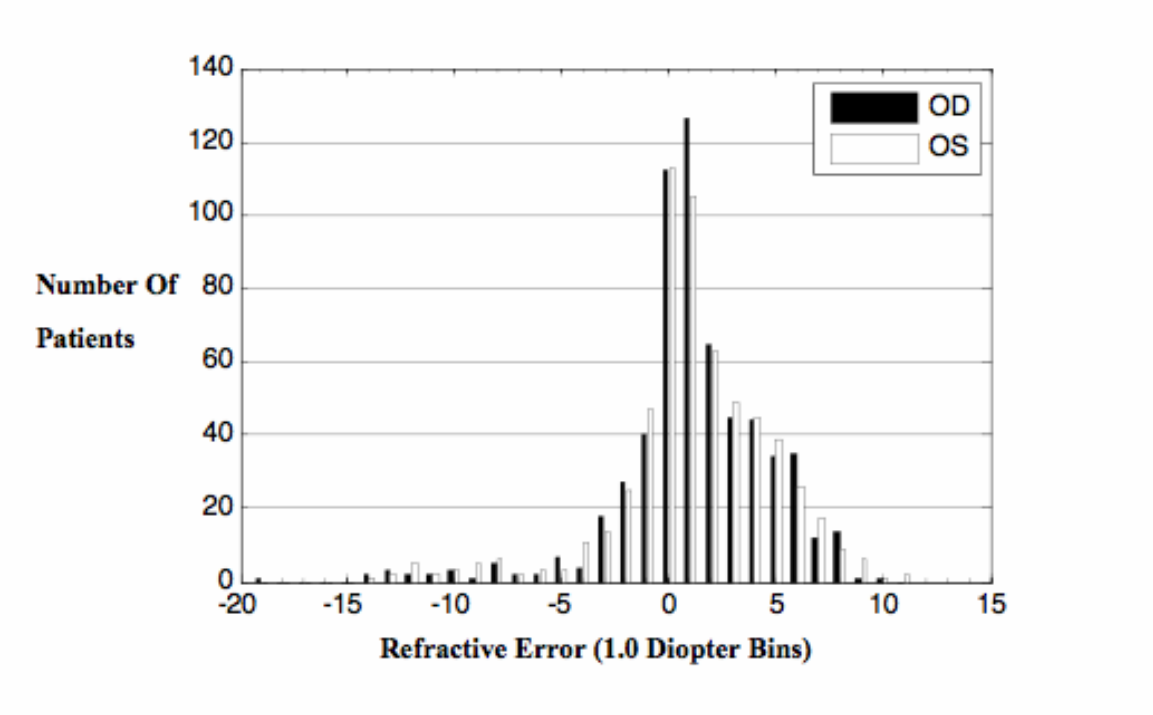

FIGURE 7
Refractive error of distribution.

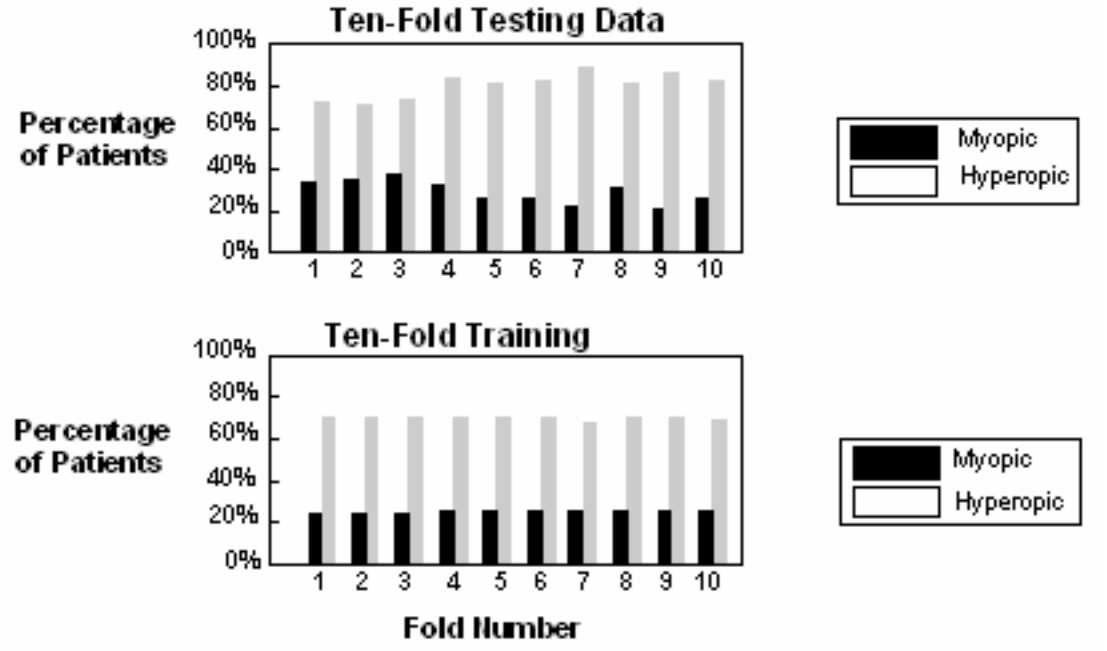

FIGURE 8
Tenfold testing and training. Unequal distribution of data.

Case-based reasoning uses template matching. It is a computer decision-making process that derives the solution to new problems by finding examples of similar problems for which there are solutions.[10] We applied the template images developed by Wang[7,8] (based on the template images of the first 200 of our known patients) to all the patients tested. With a template matching program, the features of any new unknown patient are "matched" by the computer to the top 20 stored images they most closely resemble. The program then determines refractive error and alignment for each eye (Figure 9).

An artificial neural network (NN) (Figure 10) is a model that attempts to simulate how the brain works. The network is composed

of layers of neurons that are connected, consisting of an input layer, hidden layers, and a final output layer.

A decision tree (Figure 11) is an approach to classifying cases by examining a set of associated attributes in a specific order.[11] The widely used C4.5 classifier system[12] works by calculating the importance of each attribute in the current set of attributes as determined by the gain in information as a result of the attribute.

For each system, the sensitivity, specificity, and accuracy were calculated, including 95% confidence intervals. In addition, the receiver operating characteristic curve was plotted for those systems with a numeric referral threshold. Finally, the McNemar test was used to find which systems are statistically better than the others.

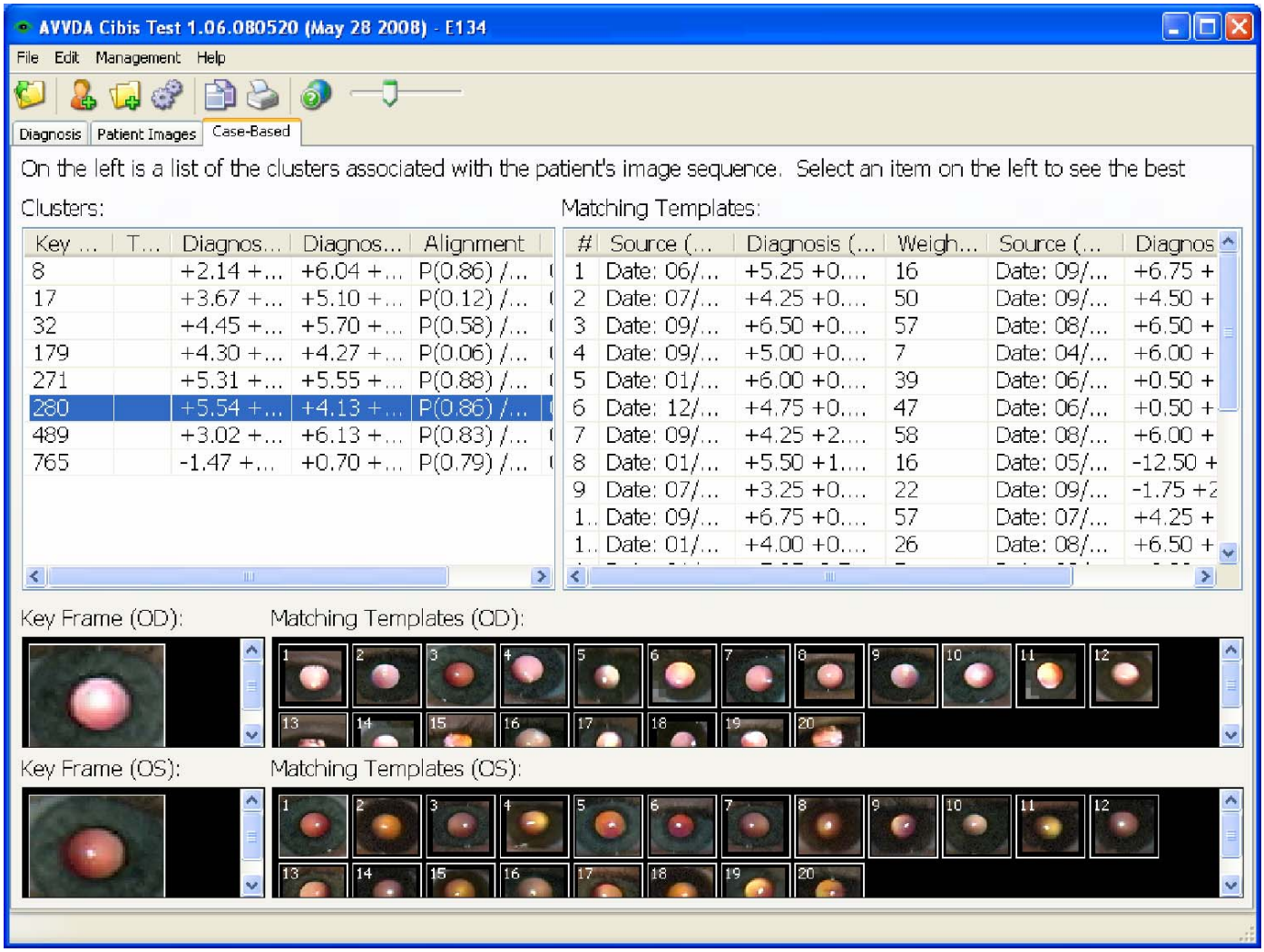

FIGURE 9

Template matching of fixing left eye vs the high refractive error in esotropic right eye. This illustrates how artificial intelligence looks at and uses all examples presented, not just the images showing fixation or crescents.

RESULTS

As seen in Figure 12, the C4.5 eye features system statistically performed better than the case-based fuzzy logic approach and NN hybrid system. Practically speaking, the difference is not great enough to justify exclusive use of the C4.5 eye features system. The C4.5 eye features system has significantly fewer steps, requiring less time to train and produce a referral decision. Time per case for analysis was 2.5 minutes for C4.5, 7 minutes for template matching, and 30 minutes for NNs. In addition, preparation and training time for some NN systems was over 100 hours, leaving C4.5 and template matching as the most practical methods for large-scale screenings Table 2.

Based on the logic that more examples yield better accuracy, we would expect better accuracy for the most numerous patient categories, namely, in the plano to +2 diopter range (Figure 13). Instead, accuracy is greater for the higher refractive errors, which are clinically more amblyogenic. This is because all example cases contribute to the classification of a single patient (Table 3).

From an artificial intelligence standpoint, extreme cases are easier to diagnose than borderline cases, because the difference from

normal cases is greater. In contrast, "correctly" diagnosing normal cases is a judgment call that is more difficult to automate.

Figure 13, showing the worst results for patients with low to moderate myopia, demonstrates the pressing need for more myopic test cases. Similarly, certain subgroups of patients had higher sensitivities than the overall numbers, more than 77% overall for astigmatism, 82% for strabismus, and 89% for anisometropia.

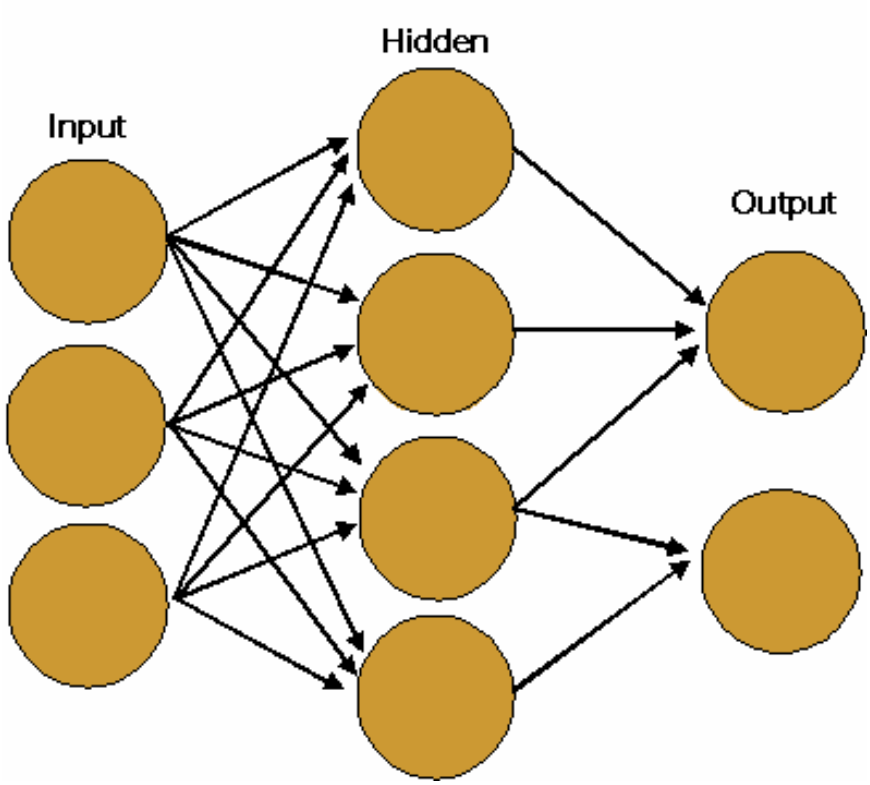

FIGURE 10

Neural network structure (image courtesy of Wikipedia Foundation Inc).

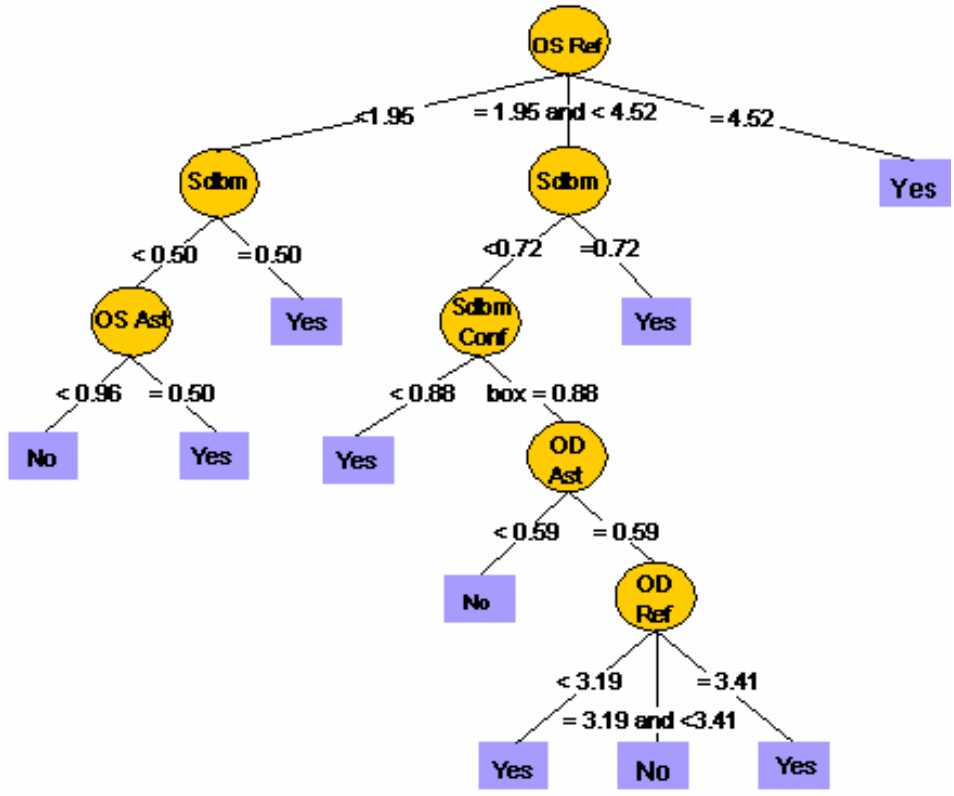

FIGURE 11

C4.5 hybrid decision tree.

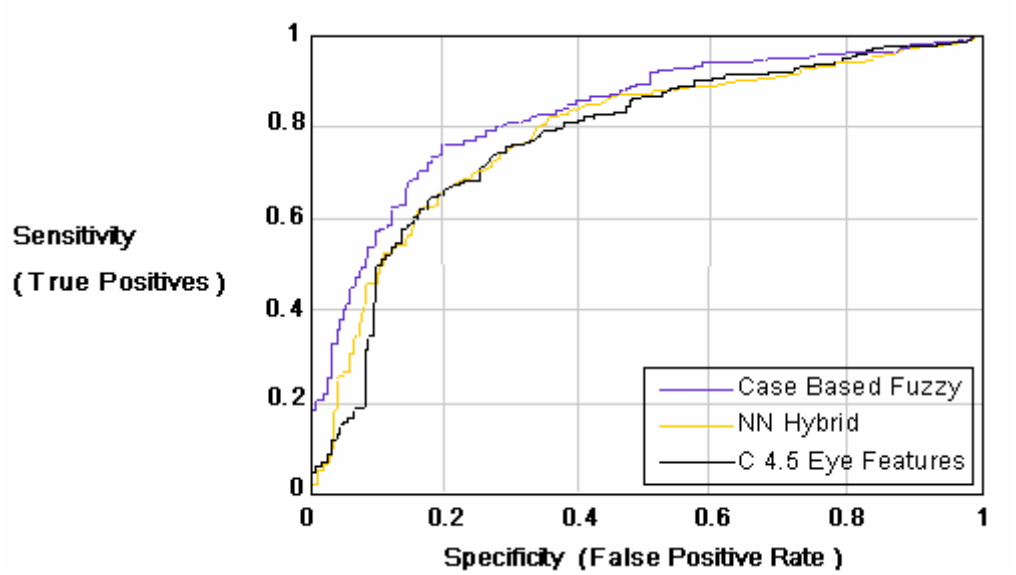

FIGURE 12

Comparision of receiver operating characteristic curves. C4.5 eye feature system performs consistently above other systems throughout entire range of cutoff values.

TABLE 2. STATISTICAL COMPARISON OF ARTIFICIAL INTELLIGENCE SCREENING SYSTEMS

SYSTEM	SENSITIVITY	SPECIFICITY	ACCURACY
Case-based static	84.6%	58.6%	75.2%
Case-based fuzzy	76.4%	68.6%	73.6%
NN eye features	61.5%	63.6%	62.3%
NN hybrid	72.3%	72.3%	72.3%
C4.5 eye features (50%)	84.6%	61.4%	76.2%
C4.5 eye features (61%)	76.2%	79.5%	77.4%
C4.5 hybrid	80.8%	59.5%	73.1%

NN, neural network.

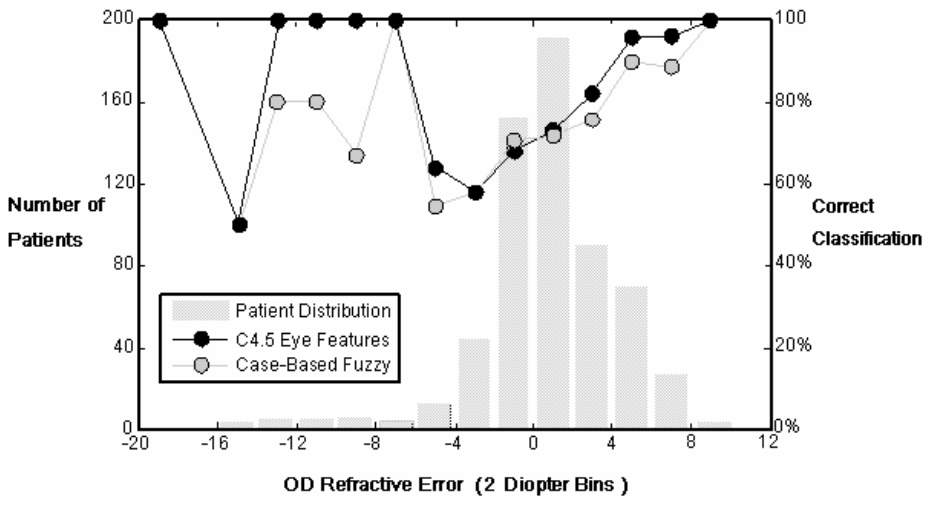

FIGURE 13

Classification by refractive error shows high accuracy for high refractive errors, both plus and minus. Modest refractive errors, plano to +4 and −4, have less accuracy, plano to −4 being the worst.

TABLE 3. RESULTS FOR HYPEROPIA VS MYOPIA

SYSTEM	MYOPIA			HYPEROPIA			ACCURACY DIFFERENCE
	Sens	Spec	Acc	Sens	Spec	Acc	
Case-based static rules	74.0	50.8	64.6	87.3	60.1	78.0	13.4 (5.4 to 21.8)
Case-based fuzzy	63.5	64.6	64.0	80.3	69.3	76.5	12.5 (4.5 to 21.0)
NN eye features*	27.3	66.7	45.0	75.0	64.3	71.4	26.4 (0.8 to 48.7)
NN hybrid	32.3	86.2	54.0	82.8	67.5	77.6	23.5 (15.0 to 32.0)
C4.5 eye features (50% cutoff)	75.0	52.3	65.8	87.6	64.4	79.7	13.8 (5.9 to 22.2)
C4.5 eye features (61% cutoff)	53.1	76.9	62.7	82.8	80.4	82.0	19.2 (11.2 to 27.6)
C4.5 hybrid	64.6	46.2	57.1	84.4	62.6	76.9	19.8 (11.4 to 28.3)

NN, neural network.
*Results are not from a complete tenfold test.

DISCUSSION

Artificial intelligence techniques to automatically analyze digitized video images for ambylogenic factors hold promise as a video screening method. The higher reliability in detecting extreme cases of amblyogenia suggests a strategy to assign a percentage after weighing the need for referral. The need to refer is based on the degree of ambylogenic potential rather than by just a "refer/do not refer" decision. For example, a child has a 70% chance of needing glasses for nearsightedness or a 50% chance of having a lazy right (or left) eye. If other factors, such as the child's visual behavior, seem to confirm this, a complete examination may be indicated.

To improve accuracy of the automated video system, more individuals with myopia need to be added to the study group. For that purpose, we have extended the age criterion for data collection to 18 years. Once that is accomplished, we will study the data to see what possible combinations of case-based and C4.5 techniques might improve accuracy.

Finally, there is a need to test a random population, such as presumably "normal" preschoolers, to see how well the program does in the real world, where the incidence of ambylogenic factors is presumed to be 5% to 10%. The population in our pediatric ophthalmology practice population invites bias. With information obtained from a nonbiased population, the cost-effectiveness for different categories of referral can be determined.

ACKNOWLEDGMENTS

Funding/Support: None.

Financial Disclosures: Dr Cibis holds a patent for automated video vision development assessment (AVVDA).

Author Contributions: All authors contributed to design management, collection of data, and analysis of results.

Conformity With Author Information: Approved by the University of Missouri, Columbia, institutional review board.

REFERENCES

1. Cibis GW. Video vision development assessment (VVDA): combining the Brückner test with eccentric photorefraction for dynamic identification of amblyogenic factors in infants and children. *Trans Am Ophthalmol Soc* 1994;92:644-685.
2. Cibis-Tongue A, Cibis GW. Brückner test. *Ophthalmology* 1981;88:1041-1044.
3. Cibis GW, Waeltermann JM. Rapid strabismus screening for the pediatrician. *Clin Pediatr* 1986;25:304-307.
4. Cibis GW. Video vision development assessment in diagnosis and documentation of microtropia. *Binoc Vis Strabismus Q* 2005;20:151-158.
5. Wang T, Giangiacomo J, Cibis GW. Computerized video-based photoscreening for amblyopia-causing factors, 2005. Available at: http://www.cs.missouri.edu/~csgsc/act/miniconf2005posters/wang_tsaipei-eyeball.ppt.
6. Wang T, Keller JM, Cibis GW. A fuzzy approach to find Hirschberg points and to determine fixation in digital images of infants. *Proceedings of 12th International IEEE Conference on Fuzzy Systems* 2003;2:955-960.
7. Wang T. *Eye Location and Fixation Estimation Techniques for Automated Video Vision Development Assessment* [master's thesis]. Columbia, Missouri: University of Missouri; 2002.
8. Wang T. *Investigation of Image Processing and Computer-assisted Diagnosis System for Automated Video Vision Development Assessment* [PhD dissertation]. Columbia, Missouri: University of Missouri; 2005.
9. Van Eenwyk J. *Using Artificial Intelligence Techniques to Automate Human Vision Screening* [master's thesis]. Lawrence, Kansas: University of Kansas; 2006.
10. Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications* 1994;7:39-59.
11. Berikov V, Litvinenko A. Methods for statistical data analysis with decision trees, 2003, Novosibirsk, Russian Federation: Sobolev Institute of Mathematics. Available at: http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm.

12. Quinlan JR. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann; 1993.

PEER DISCUSSION

DR. EDWARD G. BUCKLEY: The authors have described an innovative technique to digitize and analyze photorefractive videos in order to determine if ophthalmic referral is necessary for amblyogenic factors. The system developed relies on a complex image recognition computer program to determine which frames of a 30 second video are useful for analysis. These images are then processed by an algorithm used to generate decision trees (in the paper called "C4.5") to determine whether they have the characteristics judged to contain elements which are known to be associated with the development of amblyopia. In essence, the goal is to simplify the process of vision screening by automating the current manual review and interpretation task.

In order for photorefractive vision screening to be practical, it must eliminate the need for a skilled technician to take the photo and an expert reader to interpret the result. The described automatic technique is a step in the right direction and holds some promise. I must confess as an ophthalmologist with an engineering degree and some savvy around computers, I found the paper too technical to easily interpret. The details provided are sufficiently sketchy that an accurate understanding of just how this system was developed is difficult to understand. The basis for analysis of the images is rooted in artificial intelligence techniques that have statistical, pattern recognition, and information theory infrastructure.

Several different approaches were tested. The most intuitive to the lay reader was Case Based Reasoning. This approach uses templates of known conditions and new cases are compared until a reasonable match is found. Using this technique, images are compared to those which represent amblyogenic conditions and are then flagged if a sufficient likeness is found. The authors preferred approach is the modified decision tree program C4.5 that evaluates attributes of a case in a specific predetermined order, making decisions using those attributes which have the most discrimination information first and then selecting less helpful attributes in an orderly fashion. The trick is deciding which attributes contain the most information. The authors chose to develop this decision tree by a sophisticated data analysis of known cases using a theory called "information entropy", which is a measure of uncertainty associated with a random variable. The algorithm works by calculating the importance of each attribute in the known data pool as determined by the gain in information as a result of knowing the value of that attribute. Exactly why these specific techniques were selected is not described and there is no justification as to whether either is appropriate for this task. Using data obtained from a referral practice, the various strategies were analyzed using the above models. Each model performed reasonably well, and there was no statistical difference in the results. However, from a computational standpoint, the C4.5 decision tree model required far less calculation time than the others.

It is unclear from the paper exactly what criteria were used for possible referral and no data was presented on how it was determined whether the various programs were accurate in detecting an abnormal result. This makes it difficult to compare the sensitivity and specificity results with other studies. The recent published reports using photorefractive screening techniques combined with expert readers have sensitivities and specificities around 85 -90% [1-3]. The current study falls short of that goal and raises concerns over the ultimate usefulness as a screening tool.

A second concern is the use of a population with high ocular pathology to validate a technique which will be used to screen a population with a much lower prevalence of the disorder. As the authors correctly note, this may well result in a much different outcome thereby raising concerns of how cost effective this methodology is in detecting problems in the general pediatric population. The high sensitivities obtained may be biased by the high prevalence of the condition in the study population and the attributes that the decision tree analysis selected as having the highest information may not be useful in sorting out a less homogenous group.

Even with the above concerns, this represents a potential major step forward in the usefulness of photorefractive techniques in the screening of children for eye disorders. Replacing a cumbersome camera and reading center with a short digital video and computer assisted analysis system will address a major obstacle to the universal adoption of the photorefractive approach to vision screening. The authors are encouraged to further pursue this innovative strategy.

ACKNOWLEDGMENTS

Funding/Support: None

Financial Disclosures: None

REFERENCES

1. Group Vip S. Comparison of preschool vision screening tests as administered by licensed eye care professionals in the Vision in Preschoolers Study. *Ophthalmology.* 2004;111:637-650.
2. Kennedy R, Thomas D. Evaluation of the iScreen digital screening system for amblyogenic factors. *Can J Ophthalmol.* 2000;35:258-262.
3. Ottar W, Scott W, Holgado S. Photoscreening for amblyogenic factors. *J Pediatr Ophthalmol Strabismus.*1995;32:289-295.

DR. GERHARD W. CIBIS: What Dr. Buckley means by "real life testing" is the high percentage of pathology in our test population. I agree with everything he has said, including the shortcomings of where we are at this point. We screened around 600 children, and 50% of them should be referred based on the pathology in this population. When we achieve 90% sensitivity/specificity, we must test

this in the field where the incidence is not 50%, as it is in a referral population, but 10% or even less in the normal population. We will then see how well it does.

How does it work? Neither the computer scientist nor I know exactly. It is a black box and has its own mathematical decisions and simply makes them. As Dr Buckley says, "Yes, play golf", or "No, do not play golf"; however, the outcome is reasonable with real possibilities of improving on those answers. Thank you very much.