PENNSTATE

TransactionNumber: 1984103

ILL Number:	162739728	**Call #: Q334 .J68**
MaxCost:	**70.00IFM**	**Branch/Location: EG**
Billing Category:	**1984103**	

Penn State Interlibrary Loan – Article Request

Article Information

Journal Title: Journal of intelligent systems. **Need a clean, high resolution scan**

Volume: 20 **Issue:** 2
Month/Year: 2011 **Pages:** 101-127
Article Author:

Article Title: Arvin Agah: Applied artificial intelligence techniques for identifying the lazy eye vision disorder

Loan Information

Loan Title:

Loan Author:

Publisher:
Place:
Date:
Imprint:

Borrower Information

Interlibrary Loan - Univ. of Kansas
Rm 210 L Watson Library
1425 Jayhawk Blvd
Lawrence, Kansas 66045-7544
United States

Odyssey: 129.237.24.105
Email: illborr@ku.edu

NOTES

/ 1/11/2016 12:32:13 PM (System) Borrowing Notes: **Need a clean, high resolution scan (300-600 DPI)**; fein #486029925 (maxCost: $70)

J. Intell. Syst. **20** (2011), 101–127
DOI 10.1515/JISYS.2011.007 © de Gruyter 2011

Applied Artificial Intelligence Techniques for Identifying the Lazy Eye Vision Disorder

Patrick G. Clark, Arvin Agah and Gerhard W. Cibis

Abstract. Amblyopia, or lazy eye, is a neurological vision disorder that studies have shown to affect two to five percent of the population. Current methods of treatment produce the best visual outcome, if the condition is identified early in the patient's life. Several early screening procedures are aimed at finding the condition while the patient is a child, including an automated vision screening system. This paper aims to use artificial intelligence techniques to automatically identify children who are at risk for developing the amblyopic condition and should therefore be referred to a specialist, i.e., pediatric ophthalmologist. Three techniques, namely, decision tree learning, random forest, and artificial neural network, are studied in this paper in terms of their effectiveness, using metrics of sensitivity, specificity, and accuracy. The features used by the techniques are extracted from images of patient eyes and are based on the color information. The efficacy of pixel color data is investigated with respect to the measurement of the rate of change of the color in the iris and pupil, i.e., color slope features. A 10-fold stratified cross validation procedure is used to compare the effectiveness of the three AI techniques in this medical application domain.

Keywords. Applied artificial intelligence, pattern analysis, decision tree, random forest, artificial neural network, lazy eye, amblyopia.

2010 Mathematics Subject Classification. 68T10, 68T35, 68U10.

1 Introduction

Eye trouble of organic origin must be diagnosed and treated before the onset of an irreversible amblyopia, or what is commonly referred to as the "lazy eye". This condition is a developmental disorder of the visual system caused by ocular abnormalities early in life. While surgery or optical correction of refractive errors can often address the initial cause of amblyopia, once amblyopia has developed, such interventions cannot restore visual function since amblyopia itself is a cortical deficit, a neurological disorder and not a physical one [1]. Amblyopia has two primary causes, namely, strabismus and anisometropia. Strabismus is a misalignment between the two eyes. Anisometropia is when the refractive error between the two eyes is different [19]. The reason that these two conditions can lead to amblyopia is because they cause the brain to begin ignoring the signals from the weaker or

blurrier eye. Corrective action after amblyopia has developed becomes problematic as the brain will not be able to regenerate the neural pathways. Thus, early detection is essential for the patient to have a healthy visual outcome. Fortunately, amblyopia can be successfully treated if identified when the patient's brain is still in the developmental stages, generally when the patient is fewer than six years old, with the non-controversial methods of glasses and patching therapy achieving 80 to 90% effectiveness [4].

The most pervasive challenge to early diagnosis is the accessibility to specialists to screen patients at an early age to identify the condition. Because the current methods of identifying the problem require well-trained operators or even medical doctors, the number of personnel available to evaluate even a small fraction of the worldwide population is not sufficient. It has been reported that amblyopia affects from two to five percent of the population [16, 23]. Unfortunately a large percentage of the population lacks access to proper vision care to facilitate treatment of the problem in the optimal years (prior to six years of age) [23]. In addition, the study by Weber and Woods shows that populations that undergo early intervention have a lower prevalence of amblyopia than those that do not [23]. This implies the condition does not improve on its own accord and further supports the need for early accurate detection and treatment [23]. The socio-economic factors of amblyopia are similar to other medical issues that arise when a population does not have reliable access to medical care.

The optimal solution for vision diagnosis would be a self-contained, low-cost, completely automated system that could accurately identify disorders with minimum operator training and low patient cooperation. The solution would need to enable wide spread use with small children, including infants. One idea for this sort of solution is based on [4, 21], pioneering the science of analyzing images for identifying features that may indicate the development of amblyopia, and using artificial intelligence techniques to automate the process [20].

This paper investigates the use of AI techniques, comparing their effectiveness in analyzing additional salient features extracted from the images. [6]. The approach to investigating this problem involves four steps. The first step is to capture the patient information that will be analyzed for any vision problems. The process involves an operator positioning the patient 52 inches from a video camera with a fiber optic light source mounted just below the optical lens. The second step is to record the patient looking at the light source for approximately two minutes. The purpose of the light source is to reflect the light off the retina at the back of the patient's eyes so that the refraction through the patient's lens is captured. The third step is to process the video with a specialized software package that will analyze the patient video. The analysis of the video processes it into a set of frames that meet certain criteria for further image processing and feature extraction techniques. These features attempt to capture the rate of change of the pupil color along several axes to determine if they can accurately identify the amblyopic condition. The fourth and final step is to use the features that were extracted by three AI techniques in order to make the refer or do not refer decision (to the specialist) that the operator can use. The overall goal of a system such as this is to provide a very accurate referral mechanism so that only the patients who are at risk of developing amblyopia vision disorders would be sent to a specialist, i.e., pediatric ophthalmologist.

2 Background

2.1 Human Vision and Vision Disorders

This section provides a brief description of pertinent information on human vision and related visions disorders.

Human Vision

A diopter is a unit of measure used for describing the optical power of a lens. When used in the context of the human eye, the measure is used for describing the degree of focusing error, or refractive error, in the eye. For a perfect eye, the refractive error is 0.00 diopters (D), while a hyperopic eye has a positive refractive error (e.g. $+3.00$ D) and a myopic eye will have negative refractive error (e.g. -1.50 D). Typically the refractive error is measured to the nearest quarter diopter [20].

Hyperopia is the medical term for a person who is farsighted. Medically speaking, a farsighted eye will focus the light through the lens behind the retina and will cause the image to appear blurry. The hyperopic condition can be dealt with using a positive diopter lens, or a convex lens. Myopia is the medical term for a person who is nearsighted. Medically speaking, a nearsighted eye will focus the light through the lens in front of the retina and will cause the image to appear blurry. The myopic condition can be dealt with using a negative diopter lens, or a concave lens [20].

Vision Disorders

Amblyopia is the primary vision disorder the research presented attempts to accurately identify. It has two primary physical causes, anisometropia and strabismus. Anisometropia is a condition where the refractive error in one eye is significantly different than the other. The difference in the refractive errors is difficult to overcome for a developing visual cortex primarily due to the very different images being presented. Studies have shown that anisometropia is the predisposing condition that leads to amblyopia 50% of the time, and that an undiagnosed

anisometropia will lead to strabismus [19]. Strabismus can be identified as a misalignment of the focal point between the two eyes, but the condition is not always identifiable with the naked eye and may require a thorough screening. Strabismus typically involves a lack of coordination between the two eyes and the extra-ocular muscles where the patient is unable to bring both eyes into focus on the same point in space, thus preventing proper binocular vision. One screening test is the Hirschberg test, where a light is reflected off the patient's eyes. If the reflection is at the same place on both eyes, then the eyes are properly aligned [19].

Both of these physical abnormalities in the eye have the potential to cause the development of a patient's visual function to be impaired and cause the image from the amblyopic eye to be disregarded by the visual cortex. If the condition is allowed to persist, the neural pathways become permanently formed and the use of the amblyopic eye is diminished. The degree to which it is diminished varies based on how early the condition presented in the patient's life and if any remediation treatment was used [19].

Vision Screening

The Hirschberg testis a screening test used in the fields of ophthalmology to identify strabismus. The procedure was pioneered by Julius Hirschberg in 1886 when he used a candle to observe the reflection of light through the patient's corneas [24]. In a patient with normal eye function, the reflection of the light from the cornea results in the center of the eye. However, those patients with abnormal eye function will reflect the light off center and the degree in which the reflection is off center can be measured to determine the degree of misalignment [24]. The Hirschberg point is the point on the cornea that is the reflected light and a ratio is calculated based on that point and the edge of the pupil. The difference in the ratio is what determines the degree of misalignment between the two eyes [3]. Today, the procedure is done with more sophisticated and sensitive tools, but the general idea presented by Hirschberg remains the same.

The red reflex that occurs when a light source shines through the lens of the eye and reflects off the retina and then back to the observer is referred to as the Bruchner reflex [3]. It is named for its modern proponent who used an ophthalmoscope to view and measure the red reflex and compare the difference between the two eyes [3]. When the reflex is abnormal, the patient is considered to have positive Bruchner reflex with an abnormal Bruchner test. With a positive Bruchner test, the deviated eye is the brighter of the two [3]. Bruchner reflex is a key feature that is used to as a marker to identify a physical abnormality of the eyes that could lead to the amblyopic condition.

Foveation is used to identify the point of true fixation from a slightly off-axis fixation. Patients will sometimes experience an intermittent deviation of six degrees or less from true fixation [3]. An examination of recorded patient video reveals foveation when the focus goes from fixation to slightly off-axis fixation. The change will typically happen in one frame and is approximately $1/30$ of a second [20]. Dr. Cibis argues that this sort of behavior may hold a large amount of information regarding the true vision of a patient.

Photorefractive Screening

Photorefractive screening refers to the procedure used by ophthalmologists involving a light source and the non-linear optical effect seen as a result of shining the light source in the eye. Since the eye is essentially a transparent sphere, most of which is dedicated to focusing light on the back of the retina, how light travels through the lens and eye is essential to determining ocular health [20]. The goal is to capture the reflection of the light source off the eye in order to better understand the function of the patient's eye. The two types of photo refraction are on-axis and off-axis. The off-axis photo refraction technique has been used in order to highlight the crescent effect [3]. The degree and size of the crescent is important in identifying refractive errors between the two eyes (anisometropia) and any misalignment (strabismus). Using a recording device to capture the procedure allows the material to be reviewed in more detail by a specialist other than the person administering the test or, as in the case of this research paper, by a software system.

2.2 Artificial Intelligence Techniques

Three artificial intelligence techniques were used in this work: decision tree learning, random forests, and artificial neural networks.

Decision Tree Learning

Decision tree learning uses a decision tree as a predictive model and maps observations about an item to conclusions about the item's target value. The idea behind a decision tree is similar to how a person might make a decision. For instance, if it is cold, then one must put on a coat, and no otherwise. In the tree structures the leaves represent classifications and branches represent the joining of features that lead to those classifications. It is possible to induce a decision tree by example with the use of a training set; thus more complex data is modeled more easily by presenting the training set and then building the tree based on that data set [17]. Decision tree learning is used in data mining and is a model of the data that encodes the distribution of the class label in terms of the predictor attributes [14].

The task of developing a decision tree that agrees with the training set involves constructing a tree that has one path to a leaf for each example. Then, when given the same example again, it will output the correct classification [17]. However, the problem with this trivial tree is that it just memorizes the observations. Since the goal of a classification algorithm is to work well on the general cases and to extrapolate to examples that were not part of the training set, a set of algorithms to prune the tree to a more generalized form are typically used [17].

C4.5 is a common algorithm used to generate a decision tree, using the concept of information entropy from a set of training data to build the decision tree [15]. The algorithm uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then uses recursion on the smaller sub-lists [14, 15].

The work in this paper uses the J4.8 implementation of the C4.5 algorithm. This implementation was built by researchers are the University of Waikato in New Zealand (Weka) and is distributed for use in research. It is reported that the J4.8 implementation is modeled after C4.5 revision 8, the last published algorithm by Quinlan. The J4.8 is essentially the same as C4.5 but has performance improvements for space and complexity during tree generation [25].

Random Forests

Random forest classification is a supervised learning algorithm that builds on the decision tree algorithm. The original concept was first proposed by Tin Kam Ho of Bell Labs in 1995 where the random forest grows many classification trees. The values represented in each tree are a random vector sampled from the original dataset with the same distribution for all trees in the forest [2]. Once trained, the classifier takes a new input vector (pattern) and sends it down each of the trees in the forest. Each tree provides a classification, i.e., the tree votes for that class. The forest chooses the classification which has the largest number of votes [2, 9].

Each tree is grown in three steps. First, a training subset is defined from the original data set, and then a subset of the training set is sampled at random [2]. Second, if there are M input variables, a number $m \ll M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Third, each tree is grown to the largest extent possible. There is no pruning [2, 9].

In general, the random forest classifier is the most effective tree algorithm for problems involving highly dimensional data because of the existence of more subspaces [9]. Key benefits of using a random forest algorithm over a decision tree include the random forest classifier's ability to negate the effects of over-fitting. Over-fitting occurs when the model that results from a supervised training algorithm describes random error or noise instead of the underlying pattern. A classifier that has been over-fit to a dataset will generally have poor predictive performance and will exaggerate the minor fluctuation in data.

Artificial Neural Networks

An artificial neural network (ANN), also referred to as a multi-layer perceptron, is a technique that seeks to mimic the biological brain learning function. It is a mathematical or computation model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information from an input layer through one or many hidden layers and provides a response at the output layer [14]. The ANN is used as an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. It is a feed forward network that uses a back propagation supervised learning algorithm during the training phase. An input vector is presented to the network at a set of input nodes. The stimulus travels through each layer of the network and is transformed at each layer through a summation and an activation function. A common activation function is sigmoid [12, 14]. The primary purpose of the activation function is to amplify or attenuate the output based on a weighting factor that is adjusted at each node [14].

During the training phase, the output of the network measures the difference between the desired and the actual outputs. The algorithm then propagates weight changes back through each layer of nodes based on the difference between the desired and actual outputs. This is referred to as the generalized delta rule with back propagation of error [14]. The training algorithm iterates over the training data sets until a desired mean squared error is reached between the desired output and actual output, or after a certain number of iterations is reached.

2.3 Metrics

The metrics used in this work include sensitivity, specificity, and accuracy. Sensitivity, specificity and accuracy are statistical measures of binary classification tests and are analysis tools typically used by researchers in the medical community. In

		CLASSIFIER	
		REFER	DO NOT REFER
ACTUAL	REFER	292	146
	DO NOT REFER	155	130

Table 1. A confusion matrix.

this paper, sensitivity is defined as the proportion of the number of cases with vision disorders that were correctly classified as such by the classifier. Specificity is defined as the proportion of cases without vision disorders that were correctly classified as such. Accuracy is defined as the overall correctness of the classifier. Ideally, the system would have sensitivity and specificity of 100%; however, in reality, there is a trade off between those two values. For example, with a dataset which consists of 80% referral classes, a classifier can perform with 80% accuracy by classifying every input class as refer, resulting in sensitivity of 100% and specificity of 0%.

A confusion matrix is a method to represent classification results in a tabular format. Table 1 shows an example confusion matrix. The top row shows the summarized results of the "Refer" input class. In this case, it shows that 292 of the 438 "Refer" input class were correctly classified and 146 were not. The bottom row shows the summarized results for the "Do Not Refer" input class, where 130 were correctly classified, and 155 were incorrectly classified. The top-left and bottom-right cells are the correctly classified data, the bottom-left includes the false positives, and the top-right includes the false negatives. Using the data from Table 1, sensitivity is $66.67\% = 292/(292 + 146)$, specificity is $45.61\% = 130/(130+155)$, and accuracy is $58.37\% = (292+130)/(292+146+130+155)$.

2.4 Analysis Methods

The analysis methods used in this work are holdout testing and k-fold cross-validation testing. Holdout testing, sometimes called test sample validation, is a technique where the dataset is divided into two mutually exclusive subsets, where one subset is the training set and the other is the test set, or holdout set [11]. The holdout data set is randomly selected from the entire dataset. Designating less than third of the dataset for the holdout is common, and a classifier's accuracy can potentially increase with the more instances it sees during training. The holdout method is considered a pessimistic estimator because only a portion of the information is given to the classifier [11]. The holdout method is a common technique for testing classifier performance.

k-fold cross-validation, sometimes called rotation estimation, is a classifier testing methodology that attempts to overcome some of the weaknesses of the holdout testing methodology [11]. In this case, the dataset is randomly divided into k mutually exclusive data subsets of equal size. The classifier is then trained on $k-1$ of the datasets and tested with the remaining dataset. This train/test cycle is repeated with the classifier k times, and on each iteration, a different dataset is used the test. The cross- validation estimation of accuracy is calculated by averaging the overall number of correct classifications divided by the number of those instances in the dataset [11]. This method can be further improved by using a stratification of the original dataset. This means that the data is selectively stratified to have roughly the same proportions of the classes in each dataset as are represented in the overall dataset [11].

3 Related Work

The methods which are currently used to identify amblyopia include traditional vision screening, photorefractive screening [10], and an automated vision screening system [20].

Traditional Vision Screening

Traditional vision screening is based on the Snellen E Chart or the Stycar Graded Balls [10]. These involve the patient sitting with a trained specialist (optometrist or ophthalmologist). The Snellen E Chart is printed with eleven lines of block letters, where the first line consists of one very large letter and the subsequent rows have increasing numbers of letters that decrease in size. A patient taking the test is located 20 feet from the chart, covers one eye, and reads aloud the letters of each row, with the smallest row that can be read accurately indicating the patient's visual acuity in that eye. For example, a patient who needs to stand 20 feet away from a target that could be seen at 40 feet by a standard patient is said to have "20/40" vision. The Stycar Graded Ball test involves using a white ball before a dark background with a specialist observing the patient. The specialist will hold the white ball approximately 1.2 feet from the patient and measure the time needed for a patient to fixate on the ball. In addition, measurements of peripheral vision are also taken using this method [18]. The specialist uses these two tests to identify if the patient is at risk for developing amblyopia.

3.1 Photorefractive Screening

Photorefractive screening is based on a system to interpret the images of the eyes. It does not directly identify amblyopia, but it looks for defects in the eyes that may lead to amblyopia. Examples of these systems include Photoscreener, RetinoMax K-Plus 2, SureSight Vision Screener, Pediatric Vision Screener, MTI Photoscreener and Visiscreen [7, 20].

Photoscreener system is based on single images of the patient's eyes. The system is a hand held camera with instructions for the operator on how to identify potential problems. The operator takes a picture of each eye and then analyzes the resulting frame according to the instructions. If the frame shows markers of a vision disorder, the patient is identified for referral. RetinoMax K-Plus 2 and SureSight Vision Screener are two automated screening systems that are commercially available. The benefit of this sort of system is that the subjective nature of an operator's analysis is eliminated, and a more consistent result should be evident [20]. However, the drawback s that it operates on single frame analysis and not on a video of the patient looking at a light source. In addition, it is more focused on identifying general vision problems than those related to amblyopia. Pediatric Vision Screener, instead of taking pictures with a recording device, measures the frequency of the polarized light off the retina as a light source circles the eye. By comparing the results of the test on both eyes, the system is able to identify strabismus. The system requires the cooperation of the patient, which can be a problem when dealing with children younger than six [20]. MTI Photoscreener and Visiscreen systems are not automated and require the systems to be operated and analyzed by skilled professionals [7].

While all of these systems hold promise, they all fall short in one of three ways. The first is they require patient cooperation that would typically be beyond the demographic that would be helped by the amblyopic screening. The second is they operate only on single frame analysis and can miss the subtle foveation that may help more accurately identify a vision problem. Finally, these systems require a well trained operator or a specialist to analyze the results.

A system for identifying strabismus and amblyopia using video images has been developed [4]. The method involves a consumer grade video camera with a light source attached to the base of the camera. The patient sits approximately 52 inches from the camera and looks at the light source while approximately two minutes of video is recorded. The system is called Video Vision Development Assessment (VVDA). The video is then digitized and analyzed by specialists or trained technicians to determine if the patient should be referred to a specialist. This processing is generally divided into frame selection and feature extraction.

Frames are selected using the following criteria: the patient's head is in the primary position; the frame is in focus, with the Hirschberg point at its smallest; at least one pupil should be at its largest diameter; and the Hirschberg point should be just nasal to the center of the pupil. The selected frames are then measured, and the following measurements are taken: pupil width in pixels; the distance between the center of the Bruchner reflex and the medial edge; any observable crescent length and position; and the luminance of each pupil with a circular histogram. The resulting linear measurements are converted from pixels to millimeters and analyzed by the system. Finally, using the measurements, the technician is able to make a refer or do not refer decision for the patient. While this process begins the steps of developing an automated system, it still suffers the same issue as the traditional methods: it requires a trained technician or specialist to analyze the data. In addition, it adds a delay to the processing since the recorded data would need to be sent to a processing center before analysis can begin.

3.2 Automated Photorefractive Screening

More recently, research has been performed in order to automate the analysis of the frames using artificial intelligence techniques [20–22]. These works focused on implementing the image processing and case based reasoning algorithms that constituted the first version of the Automated Video Vision Development Assessment (AVVDA) [5]. In a completely automated fashion, they were able to identify the key frames of the video, isolate the pupils, and locate the Hirschberg point [21,22]. Figure 1 shows the image output from AVVDA system, where the Hirschberg reflex and iris diameter are highlighted. The automated photorefractive screening system works by having an operator take a short video of the patient, which is then analyzed automatically by the software in the following manner. Initially, the software identifies the frames where both eyes are open and looking at the light source. These frames are identified as key frames. Next, the software isolates the location of the eyes and pupils in the key frames, and finally, the software uses various techniques to extract the distinguishing features that may be indicators for amblyopia.

In order to further enhance AVVDA, Researchers have investigated using the same feature set with a different set of classifiers, with the overall goal of the AVVDA system being to allow an unskilled technician to operate the system and accurately get a decision about patient referral to an optometrist or ophthalmologist [20]. In the current form, AVVDA uses cased based reasoning, C4.5 decision tree, and artificial neural network classifiers to assist in making the decision. Training the classifiers involves 54 features that are extracted from the images.

At this point, various artificial intelligence techniques have been utilized to automatically produce the referral to a specialist or not [4, 20]. The work reported

Figure 1. Key frame output from the AVVDA system [20].

in this paper expands upon the automated photorefractive screening method, with the goal to investigate a different set of features, i.e., color slope analysis.

4 Video Processing and Feature Extraction

In this paper, the utilization of eye features based on the color slopes that are extracted from the key frames is investigated. These features are used by three artificial intelligence techniques and the accuracies of the classifiers are studied. The three classifiers are a C4.5 decision tree, a random forest, and an artificial neural network. The Weka software implementation is used for the C4.5 decision tree and random forest [25], and Matlab Neural Network Toolbox is used for the artificial neural network [12].

The experiments involve processing the patient videos and prepping the data for feature extraction. The available data include 723 patient videos.

Feature extraction and efforts to reduce the number of features in the dataset (dimension) is an important aspect of this project. The research is divided into three phases, reflecting how the features are extracted. The overall goal of feature extraction is to represent the rate of change of the colors from one side of the pupil to the other. In order to get this color data in the affected area of the eye, the raw image data must go through a series of steps to extract the information that will be submitted to the classifiers. Figure 2 illustrates the general procedure.

4.1 Extracting Key Frame

A major step involves selecting a key frame from which to extract the pixel data. The goal is to use a frame in which the patient is exhibiting the amblyopic condition. The process focuses on ensuring that the patient's head is in a primary position, the image is in focus, the Hirschberg point is at its smallest, at least one pupil is at its largest diameter, and the Hirschberg point is just nasal to the center of the pupil.

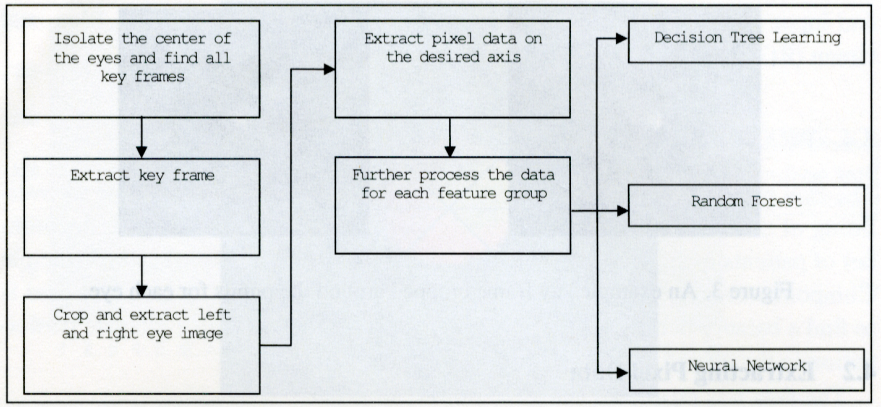

Figure 2. The feature extraction process.

During phase one and phase two, a simple approach was used to identify the one frame that is used for feature extraction. The frame in the exact middle of the stack of frames was selected, adjusting by one in case of even number of frames. The aim is to avoid the fringe frames (the first or last), and analysis of a subset of the data shows that the middle frame has the greatest chance of being one of the best images.

During phase three, the frame selection process was altered. The goal for phase three is to select the frame out of all the key frames where the patient's head is most level. Since the entire classification problem is centered on the reflection of light off the lens and the retina, keeping the angle of reflection the same for both eyes would, theoretically, further enhance any differences between a healthy eye and the eye of a patient who should be referred to a specialist. This idea is further supported by the important features extracted around the crescent reflection and the Hirschberg point that is measured from that crescent. The frame that is most level is determined by the pixel location of the center of the pupil at the y axis. The value is compared between the two eyes for each key frame and the one where the value is the closest is the frame selected. When no key frame could be found where the difference in the height of the eyes was less than five pixels, then the patient data was discarded. This process reduced the number of patients in the entire sample from 723 to 499 for phase 3.

Once the key frame is identified, the image processing has determined the x and y coordinates for the center of the pupil and the radius of the iris. Using this data a left and right eye image is produced. Both images are 62 by 56 pixel images and resemble Figure 3.

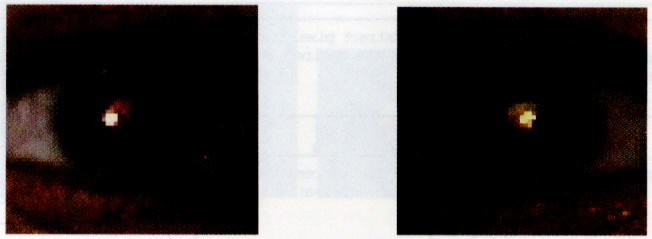

Figure 3. An example key frame cropped around the pupils for each eye.

4.2 Extracting Pixel Data

After the individual eye images are available and cropped, the pixel values are to be extracted from different angles off the center of the eye. During phase one of the project these angles are the 0, 45, and 135 degrees on the center of the eye and across the iris and the pupil. The 90 degree axis will not be used because it typically is partially obstructed by the upper or lower eyelid. Therefore, on the zero axis, 36 pixels are extracted, and on the 45 and 135 axes, 28 pixels are extracted. That makes a total of 92 pixels, with three colors on each pixel, namely, red, blue, and green. Finally, that number is doubled to account for both the left and right eye. Notably, each pixel extracted was averaged against its immediate neighbor in order to smooth out any drastic errors that may have occurred in the image capture process. Figure 4 shows where the data is extracted across the eye image.

Phase two and three used a different process. Since iris color has not shown predictive capability for strabismus and anisometropia [20], the color extraction was isolated to the pixels in the pupil. Also, in order to reduce the dimensionality of the data and focus areas with most predictive information, the axes extracted were focused on the vertical center of each eye (90 degree) and the axis that directed through the nose, which is 45 degree for the left eye and 135 degree for the right, both from the observer's point of view. This process reduced the number of raw pixel features to 180. Figure 5 shows phases two and three feature extraction.

4.3 Producing Feature Groups

Since the goal of this project is to investigate feature groups to determine if they provide enough distinction for accurate classification, four feature extraction strategies were chosen: Pixel Values, PCA Pixel Values, Four Color Slope Values, and One Color Slope Values. Not all the strategies are used in all three phases. Phase one uses all four as an initial review of how the data might perform. Phases two and three only use Pixel Values and One Color Slope Values feature groups.

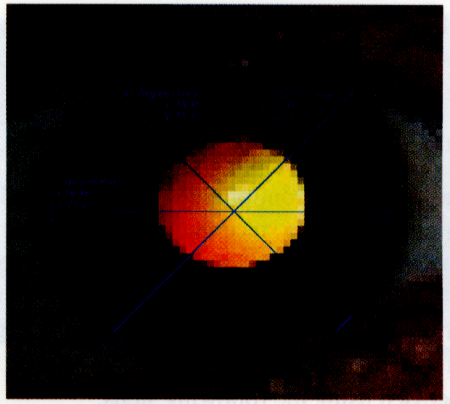

Figure 4. Pixel color extraction.

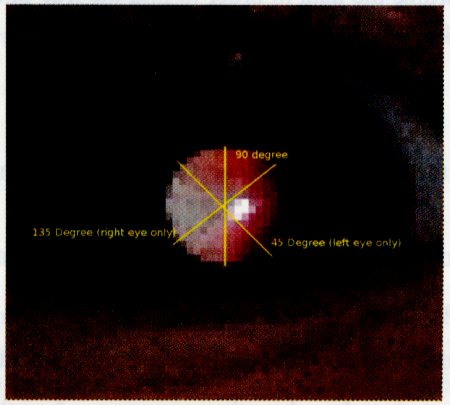

Figure 5. Phases two and three pixel color extraction.

The reason that the PCA Pixel Values and the Four Color Slope Values feature groups are not used in the later phases is because they are both dimension reduction strategies. Since the phase two and three datasets use significantly fewer data points, reducing the dimensions of the raw data was not seen as necessary.

Pixel Values

Here the strategy is to use all the pixel color values that were extracted from the eye images. While this is not expected to yield good results, it is a starting point for establishing the efficacy of the data. Assembly of the data is fairly straightforward,

and input files for the classifiers are produced by appending the Red, Blue, and Green (RGB) values for each axis and each eye.

4.4 PCA Pixel Values

The raw pixel values produced from phase one will probably be prone to overfitting of the data in the classifiers, since the 552 features almost match the number of patients in the dataset. Therefore, the values were processed using Principle Components Analysis (PCA) to reduce the dimensionality of the data. The goal is to find a balance between a small number of features and a high aggregate participation in the overall variance of the data.

The data was processed using the following steps:

(1) Start with the 552 features and standardize the data. When using PCA, a good practice is to standardize the data in order to find the true dimension variance. A common standardization formula was used:

$$\frac{X_i - \mu}{\sigma}.$$

Where X_i is the vector X minus the mean μ and divided by the standard deviation σ of the vector X.

(2) Calculate the covariance matrix using the standardized dataset.

(3) Run PCA against the dataset.

(4) Transform the data and compute the principle components.

(5) Reduce the dimensions based on the percentage of participation in the variance.

After the data was processed, 15 principle components were selected. These accounted for 74.23% of the participation in the total variance of the dataset. No further principle components were included because the remaining 537 components each contributed less than half of a percent of the overall variance. Therefore it was deemed including any additional components would not provide additional classification capabilities.

Four Color Slope Values

In this feature group, the rate of change in color is extracted from four segments of the color line for each axis. The formula for calculating the slope of each segment

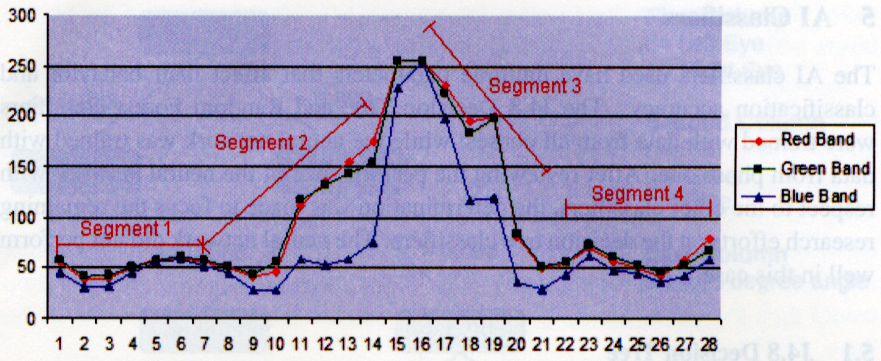

Figure 6. An example of the color segments along the 45 degree axis.

allows the rate of change between every data point to contribute to the final slope value:

$$\frac{[\sum_{i=1}^{j}(p_i + p_{i+1})]/(j-1)}{j}.$$

Where j is the number of pixels in the segment, and p is the pixel color for that particular color band (red, green, or blue). Once processed, 72 features result: 4 slope values times 3 colors times 3 axes times 2 eyes. Figure 6 is an example color line with the segments diagrammed.

One Color Slope Values

This feature group is drawn from taking the slope across the entire color line. This is similar to the four color slope values in that it uses the same formula for calculating the slope. The difference is that j is equal to the length of the entire color line. Once the data are extracted and calculated, 18 features are produced in phase one (3 colors times 3 axes times 2 eyes) and 12 features are produced in phase two and three (3 colors times 2 axes times 2 eyes).

5 AI Classifiers

The AI classifiers used have multiple parameters that affect their behavior and classification accuracy. The J4.8 Decision Tree and Random Forest classifiers were trained with data from all phases; while the neural network was trained with data from phase one. After reviewing the performance of the neural network with respect to the other classifiers, the determination was made to focus the remaining research efforts on the decision tree classifiers. The neural network did not perform well in this case.

5.1 J4.8 Decision Tree

The Weka J4.8 decision tree classifier allows for some customization of its operation based on command line options. These options tune the performance of the classifier and could possibly lead to better classification results [25]. The Weka J4.8 documentation lists ten tunable options; however, the most pertinent are the confidence threshold and the minimum number of instances per leaf. The confidence threshold allows the user to customize the pruning of the tree. Decreasing the value of the parameter has the effect of more aggressively pruning the tree and removing leaves that do not substantially contribute to classification. The minimum number of instances per leaf affects how the tree is constructed and causes further splits in order to meet the minimum number of classes at the leaf layer [25].

For all of the J4.8 classifier training in this project the default parameters were used. This means that a 10-fold cross validation is used for testing. Pruning is used with a 0.25 confidence threshold and a minimum of two instances per leaf [25]. Further experimentation was done with the other options; however, the results were not any better than what was produced by the options discussed. For an example of a pruned decision tree, Figure 7 shows the tree produced from the phase two, one-slope feature set.

5.2 Random Forest

The Weka random forest classifier allows for customization of its operation based on several command line options: the number of trees produced, the number of features used, and the depth of each tree. The random forest produces he best results by building large numbers of trees, typically greater than 1000, using all of the features, and allowing the depth of the trees to grow as large as necessary.

When testing with the data from each of the phases, the typical behavior observed would be an increased accuracy rate as larger number of trees is used. However, at a point in the train and test cycle, the accuracy would plateau. At

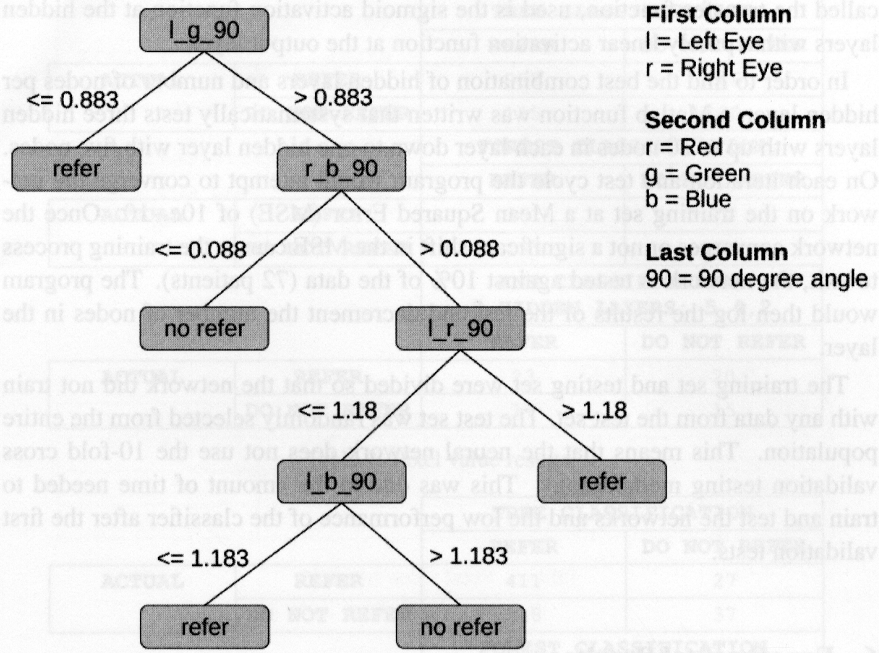

Figure 7. Decision tree generated from the phase two one slope feature set.

this point the train and test cycle were halted, and the best accuracy was deemed to have been found. This behavior is supported through mathematical examination and explaining that random forests do not over-fit as more trees are added, but produce a limit on the generalized error [2].

5.3 Neural Network

When using the artificial neural network implementation provided by Matlab, the default settings are taken [12]. In general, the inputs are not normalized between 0 and 1, but were kept at their original values, 0 to 255, because there was already a reasonable upper and lower bound on the data and there would not be any fluctuations or spikes that would cause the classifier to miss any subtleties in the data due to an overpowering spike. The exception to the normalization process is in the case of the principle components analysis in phase one. The reason for normalizing the dataset at that phase is due to the process of gathering the principle components. Part of the procedure is to start with normalized data. The activation function, also

called the transfer function, used is the sigmoid activation function at the hidden layers with a purely linear activation function at the output layer.

In order to find the best combination of hidden layers and number of nodes per hidden layer, a Matlab function was written that systematically tests three hidden layers with up to 20 nodes in each layer down to one hidden layer with five nodes. On each iteration and test cycle the program would attempt to converge the network on the training set at a Mean Squared Error (MSE) of 10e−10. Once the network converges or not a significant shift in the MSE causes the training process to exit, the network is tested against 10% of the data (72 patients). The program would then log the results of the test and decrement the number of nodes in the layer.

The training set and testing set were divided so that the network did not train with any data from the test set. The test set was randomly selected from the entire population. This means that the neural network does not use the 10-fold cross validation testing methodology. This was due to the amount of time needed to train and test the networks and the low performance of the classifier after the first validation tests.

6 Experimental Results

6.1 Phase One

Tables 2 through 5 show the confusion matrix results of the phase one classification results. Tables 6 through 9 summarize the performance of each classifier using specificity, sensitivity, and accuracy metrics.

6.2 Phase Two

Tables 10 and 11 show the confusion matrix results of the phase two classification results. Tables 12 and 13 summarize the performance of each classifier using specificity, sensitivity, and accuracy metrics.

6.3 Phase Three

Tables 14 and 15 show the confusion matrix results of the phase three classification results. Tables 16 and 17 summarize the performance of each classifier using specificity, sensitivity, and accuracy metrics.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	292	146
	DO NOT REFER	155	130
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	361	77
	DO NOT REFER	155	130
		ANN CLASSIFICATION 3 HIDDEN LAYERS: 5,4,2	
		REFER	DO NOT REFER
ACTUAL	REFER	23	20
	DO NOT REFER	14	15

Table 2. Pixel value results.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	411	27
	DO NOT REFER	248	37
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	374	64
	DO NOT REFER	194	91
		ANN CLASSIFICATION 1 HIDDEN LAYER	
		REFER	DO NOT REFER
ACTUAL	REFER	48	4
	DO NOT REFER	20	0

Table 3. PCA Pixel value results.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	258	180
	DO NOT REFER	149	136
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	353	85
	DO NOT REFER	162	123

Table 4. Four color slope value results.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	438	0
	DO NOT REFER	285	0
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	340	98
	DO NOT REFER	233	52

Table 5. One color slope value results.

522 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	66.67%	45.61%	58.37%
FOREST	82.42%	45.61%	67.91%
ANN	53.49%	51.72%	52.78%

Table 6. Pixel value metrics.

15 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	93.84%	12.98%	61.96%
FOREST	85.34%	31.93%	64.32%
ANN	92.31%	0%	66.67%

Table 7. PCA Pixel value metrics.

72 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	58.90%	47.72%	54.50%
FOREST	80.59%	43.16%	65.81%

Table 8. Four color slope value metrics.

18 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	100.00%	0.00%	60.58%
FOREST	77.63%	18.25%	54.22%

Table 9. One color slope value metrics.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	305	132
	DO NOT REFER	185	101
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	369	68
	DO NOT REFER	194	92

Table 10. Pixel value results.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	379	58
	DO NOT REFER	224	62
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	353	84
	DO NOT REFER	195	91

Table 11. One color slope value results.

180 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	69.79%	35.31%	56.15%
FOREST	84.44%	32.17%	63.76%

Table 12. Pixel value metrics.

12 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	86.73%	21.68%	61.00%
FOREST	80.78%	31.82%	61.41%

Table 13. One color slope value metrics.

6.4 Analysis

The neural network was tested using fewer samples because the 10-fold cross validation testing was not used. Instead, a holdout test was used with 10% of the population randomly selected as a member of the test set. However, the results showed as just slightly better than flipping a coin. Since the goal was to identify

a good feature set, this sort of result does not warrant stratified cross validation or continuing to use the classification on the remaining phases of the project.

The best overall accuracy of the new feature set was 68%. At times, the system produced sensitivity in the 80 and 90 percent range. Although the accuracy results are not as high as the 77.5% accuracy reported [20], they show that the color slope analysis has merits. The combination of the two works may produce even better

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	86	107
	DO NOT REFER	100	206
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	72	121
	DO NOT REFER	39	267

Table 14. Pixel value results.

		TREE CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	69	124
	DO NOT REFER	78	228
		FOREST CLASSIFICATION	
		REFER	DO NOT REFER
ACTUAL	REFER	65	128
	DO NOT REFER	59	247

Table 15. One color slope value results.

180 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	44.56%	67.32%	58.52%
FOREST	37.31%	87.25%	67.94%

Table 16. Pixel value metrics.

12 FEATURES	SENSITIVITY	SPECIFICITY	ACCURACY
TREE	35.75%	74.51%	59.52%
FOREST	33.68%	80.72%	62.53%

Table 17. One color slope value metrics.

results. This work also illustrates the difference between the three AI techniques as they apply to this domain. The best accuracy was from the random forest. The random forest likely produced the best results because it is able to overcome the over-fitting issue. The number of features and the number of samples lent itself to that sort of a problem.

7 Conclusion

The research presented in this paper contributes to the advancement of the automatic identification of human vision disorder of amblyopia in terms of color slope feature extraction and use of AI techniques. The goal is to build a system for identifying patients that have a high risk of developing the amblyopic condition and should be referred to a specialist. This paper provided a thorough examination of color slope features paired with three distinct classifiers. The primary focus was the pixel color across several axes of the eyes and the rate of change of that color across the iris. The research used a rigorous 10- fold stratified cross validation testing procedure for the two most promising classifiers, Random Forest and J4.8. This means that all patient data was taken into account and no attempts were made to filter out the difficult patient videos in order to gain better classification accuracy. The best result achieved an overall accuracy of 68%.

A number of approaches have been identified as possible future work. Finding the frame that captures the patient exhibiting signs of the amblyopic condition is one of the most important parts of the feature extraction process. It is also one of the most difficult. One research area would be to investigate using more than one frame or an average of all the frames in hopes of capturing the necessary information. Another approach can include the investigation of foveating frames, referring to the true fixation to distinguish it from slightly off-axis fixation [4, 20]. An examination of recorded video reveals frequent cases where the patient's eyes will switch between fixation and slightly off-axis fixation, or foveate. These patients will then switch again only a few frames later. The goal would be to focus the feature extraction only on these foveating frames to determine if the color slope gathered from that frame will yield more accurate results than other frame selection strategies. Another potential approach is to combine color slope analysis with other features used which have produced good results [20], developing a hybrid system with good potential.

Bibliography

[1] S. J. Anderson, I. E. Holliday and G. F. Harding "Assessment of cortical dysfunction in human strabismic amblyopia using magnetoencephalography," (MEG). *Vision Research*, vol. 39 (1999), pp. 1723–1738.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45 (2001), pp. 5–32.

[3] G. W. Cibis, "Video Vision Development Assessment (VVDA): Combining the Bruckner Test with Eccentric Photoreflection for Dynamic Identification of Amblyogenic Factors", *Transactions of the American Ophthalmological Society*, vol. 84 (1994), pp. 643–685.

[4] G. W. Cibis, "Video vision development assessment in diagnosis and documentation of microtropia," *Binocular Vision & Strabismus Quarterly*, vol. 20 (2005), pp. 151–158.

[5] Cibis Eye Care, `http://cibiseyecare.com/`.

[6] P. G. Clark, *Identifying Vision Disorders using Iris Color Analysis*. MS Thesis, Department of Electrical Engineering and Computer Science, University of Kansas.

[7] H. L. Freedman and K. L. Preston, "Polaroid Photoscreening for Amblyogenic Factors. An Improved Methodology," *Ophthalmology*, vol. 99 (1992), pp. 1785–1795.

[8] The Vision in Preschoolers Study Group, "Preschool Vision Screening Tests Administered by Nurse Screeners Compared with Lay Screeners in the Vision in Preschoolers Study," *Investigative Ophthalmology & Visual Science*, vol. 46 (2005), pp. 2639-2648.

[9] T. K. Ho, "Random Decision Forest," *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, vol. 1 (1995), pp. 278–282.

[10] A. R. Kemper, P. A. Margolis, S. Downs and W. C. Bordley, "A Systematic Review of Vision Screening Tests for the Detection of Amblyopia," *Pediatrics*, vol. 104 (2007), pp. 1220–1222.

[11] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *International Joint Conference and Artificial Intelligence*, pp. 1134–1143, 1995.

[12] The MathWorks Inc., "MATLAB," Available at `http://www.mathworks.com/products/matlab/`.

[13] American Academy of Pediatrics, "Use of photoscreening for children's vision screening," *Pediatrics*, vol. 109 (2002), pp. 524–525.

[14] Y. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, 1989.

[15] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[16] D. Robaei, K. Rose, E. Ojaimi, A. Kifley, F. Martin and P. Mitchell, "Causes and Associations of Amblyopia in a Population-Based Sample of 6-Year-Old Australian Children," *Archives of Ophthalmology*, vol. 126 (2006), pp. 878–884.

[17] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall Inc., 1995.

[18] M. D. Sheridan, "The STYCAR graded-balls vision test," *Developmental Medicine and Child Neurology*, vol. 15 (1973), pp. 423–432.

[19] S. B. Steinman, B. A. Steinman and R. P. Garzia, *Foundations of Binocular Vision: A Clinical Perspective*, McGraw-Hill, 2000.

[20] J. Van Eenwyk, A. Agah and G. W. Cibis, "Automated Human Vision Assessment Using Computer Vision and Artificial Intelligence," *Proceedings of the IEEE international Conference on System of Systems Engineering (SoSE 2008)*, Monterey, California, June 2008.

[21] T. Wang, "Eye Location and Fixation Estimation Techniques for Automated Video Vision Development Assessment," in Computer Engineering and Computer Science: University of Missouri – Columbia, 2002.

[22] T. Wang, "Investigation of Image Processing and Computer-Assisted Diagnosis System for Automated Video Vision Developement Assessment," in *Computer Engineering and Computer Science: University of Missouri – Columbia*, 2005.

[23] J. L. Weber and J. Wood, "Amblyopia: Prevalence, Natural History, Functional Effects and Treatment," *Clinical and Experimental Optometry*, vol. 88 (2005), pp. 365–375.

[24] M. Wheeler, "Objective Strabismometry in Young Children," *Transactions of the American Ophthalmological Society*, vol. 40 (1942), pp. 547–564.

[25] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition: Morgan Kaufmann, San Francisco, 2005.

Received April 10, 2011.

Author information

Patrick G. Clark, Department of Electrical Engineering and Computer Science,
The University of Kansas, Lawrence, KS 66045, USA.

Arvin Agah, Department of Electrical Engineering and Computer Science,
The University of Kansas, Lawrence, KS 66045, USA.
E-mail: `agah@ku.edu` (corresponding author)

Gerhard W. Cibis, Cibis Eye Care, Kansas City, MO 64112, USA.