# Classification of High Frequency Pupillary Responses using Schur Monotone Descriptors in Multiscale Domains

**Bin Shi, Kevin P. Moloney, Ye Pan, V. Kathlene Emery, Brani Vidakovic, Julie A. Jacko, Francois Sainfort**
School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0205, USA

September 21, 2004

**Abstract**

This paper addresses the problem of classifying users with different visual abilities based on their pupillary response data while performing computer-based tasks. Multiscale Schur Monotone (MSM) summaries of high frequency pupil diameter measurements are utilized as feature vectors (or input vectors) in this classification. Various MSM measures, such as Shannon, Picard, and Emlen entropies, the Gini coefficient and the Fishlow measure, are investigated to assess their discriminatory characteristics. A combination of classifiers, motivated by Bayesian paradigm, is proposed to minimize and stabilize the misclassification rate in training and test sets with the goal of improving classification accuracy. In addition, the issue of wavelet basis selection for optimal classification performance is discussed. The members of the Pollen wavelet library are included as competitors. The proposed methodology is validated with extensive simulation and applied to high-frequency pupil diameter measurements collected from 36 individuals with varying ocular abilities and pathologies. The expected misclassification rate of our procedure can be as low as 21% by appropriately choosing the Schur Monotone summary and using a properly selected wavelet basis.

# 1 Introduction

The HCI (human-computer interaction) community is interested in understanding the the unique interaction needs and behaviors of individuals with visual impairments who retain visual capabilities, albeit at a below 'normal' level [*Biglan et al,* 1988]. Therefore, there is a need for methods and procedures that can provide meaningful classification of individuals with varying visual abilities. In the human visual system, the pupil functions as a gain control device, which responds to external stimuli, such as luminance changes, color and pattern changes, onset of motion, attention and social signaling, in a very subtle way. It has been widely accepted [*Backs & Walrath,* 1992, *Hess & Polt*, 1960, *Hess & Polt*, 1964] that pupillary response (in terms of the dynamic pupil size) is becoming an important mmeasure of mental workload, which may be useful for classifying users with different abilities.

However, the pupil has an extremely complex control mechanism, which is moderated by several variables [*Sahraie & Barbur*, 1997], as well as various neural control pathways [*Barbur,* 2003]. As such, it is very difficult to tease out the underlying differences in mental workload or information processing when merely looking at point differences in pupil diameter. The inherent complexity of pupillary behavior requires that robust and valid measures be developed to extract the meaningful components from dynamic pupil behavior. While smoothing large aberrations in data values, and using global or local means, may be suitable in helping to highlight even slight changes in pupil diameter for short, simple tasks this averaging typically does not work for longer, more complex tasks that will inherently include more natural fluctuations in pupillary response and a larger number of confounding, non-cognitive effects. This being said, it is necessary to develop analytical techniques that can isolate these small changes in pupillary behavior. A more sensitive tool for the analysis of pupil measurement data may provide a solution to this problem and provide a unique characterization of interaction for individuals who are aging and/or have visual impairments.

This study examines the dynamic pupillary behavior of four groups of individuals, in which known performance differences were exhibited, during a computer-based task. Additionally, this study aims to examine if these behavioral differences can be sufficiently modeled for purposes of user classification, proposing the application of low dimensional summaries of high frequency data. Specifically, a summary measure called the Multiscale Schur

Monotone (MSM) measure is derived to characterize the disbalance properties of the data distribution at different frequency scales. The MSM measure carries information not only about the disbalance characteristics of the data, but also about its correlation structure. Thus, the MSM summary is more likely to be more sensitive to the differences in visual functioning between users than any other single measure, such as correlation and global Schur Monotone measures. The combination of classifiers is proposed to address the inhomogeneous discriminatory information in the pupil diameter measurements.

The remainder of this paper presents these MSM measures and their application in the classification of individuals with varying visual functioning. Section 2 derives a meaningful summary for high-frequency measurements for the purpose of classification, with wavelet transform and Schur Monotone measures briefly reviewed. Additionally, the concept of Schur Monotone summaries in the multiscale domain (MSM) is introduced and its application is illustrated via examples using MSM summaries. Section 3 describes the classifier combining procedure and provides a Bayesian justification. Section 4 discusses the high frequency pupil diameter measurements used in this study. Section 5 illustrates the use of the MSM summaries of the high frequency pupillary behavior to classify the users. The $K$-nearest-neighbor ($K$-NN) classifier, equipped with combining techniques, is used for the classification. Finally, Section 6 discusses the factors affecting the classification performance and the practical implications of the findings for research in HCI.

# 2 Schur Monotone Summaries in Wavelet Domain

In this section, we first briefly review the wavelet transforms. Next, the concepts of Schur Monotone ordering and Schur Monotone (SM) measures are presented. Then, we introduce the Mutliscale Schur Monotone (MSM) measure as a natural way to combine multiscale representations and Schur ordering, and give two illustrative examples to demonstrate this new measure.

## 2.1 Discrete Wavelet Transform

Discrete (orthogonal) wavelet transformations (DWT) have become indispensable tools in the analysis of data a with complex stochastic structure. It turns out to be an appropriate tool to model non-stationary, non-Gaussian and long memory measurements. DWT is simply a linear transformation. Let $y$ be a data vector of dimension (size) $n$. To avoid algorithmic complications we assume that $n$ is an integer power of 2. The vector $d$ representing a wavelet transform of vector $y$ can be written as

$$d = Wy,$$

where $W$ is an orthogonal matrix of size $n \times n$.

In practice, the choice of the orthogonal matrix $W$ is related to the selection of wavelet basis, and ultimately, to the selection of a wavelet filter needed to implement the transform. The details on theory and statistical applications of wavelets can be found in [*Vidakovic,* 1999]. Due to the properties of the wavelet functions, $W$ usually admits the factorization in terms of a series of sparse matrices. A fast algorithm based on filtering to (equivalently) factorize the matrix $W$ and calculate the wavelet-transformed vector $d$ was proposed by [*Mallat*,1989]. This algorithm is easily implemented and is a part of many standard wavelet packages, such as the WAVELAB module for MATLAB by the team from Stanford University (http://www-stat.stanford.edu/∼wavelab/).

## 2.2 Schur Monotone(SM) Ordering

Schur Monotone ordering is the basis to define the SM measure of a vector. This is used to order the vectors in terms of the their "disbalancing" characteristics. The definition of Schur ordering utilizes the inverted order statistic of two normalized vectors with non-negative components. For a pair of $n$-dimensional vectors $x$ and $y$ with non-negative components, the Schur ordering is defined as

$$x \prec y \quad \text{if} \quad \begin{cases} \sum_{i=1}^{k} x_{[i]} < \sum_{i=1}^{k} y_{[i]}, & k = 1, ..., n-1 \\ \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \end{cases} \tag{1}$$

with $x_{[i]}$ and $y_{[i]}$ being the $i$th largest components of $x$ and $y$ respectively. When $x \prec y$, then it is said that $x$ is Schur majorized by $y$.

## 2.3 Schur Monotone Measure

The Schur Monotone measure is a scalar value assigned to a vector that is sensitive to the Schur Monotone order. There are many available Schur Monotone measures, which have been previously used in economics and biology. In fact, any function $\phi$ such that one of the following two conditions are satisfied

1. $x \prec y \Longleftrightarrow \phi(x) \leq \phi(y)$, and $\phi(ax) = \phi(x)$ for all $a > 0$ or
2. $x \prec y \Longleftrightarrow \phi(x) \geq \phi(y)$, and $\phi(ax) = \phi(x)$ for all $a > 0$

can be used to measure the disbalance of vector $x$. If the first condition is satisfied, function $\phi(\cdot)$ is called a Schur convex measure. If the second condition is true, $\phi(\cdot)$ is called a Schur concave measure. In both cases, $\phi(\cdot)$ is a Schur Monotone measure. In this paper, we are interested in a SM measure defined as

$$\phi_2(x) = -\sum_i \log \frac{x_i}{S},$$

where $S = \sum_i x_i$. This SM measure is usually called Picard entropy (Picard, 1979), which is different from Shannon entropy (Shannon, 1948) by only removing $x_i$'s in the summation terms. Other SM measures utilized in this project include Gini's coefficient (Gini, 1912), Fishlow's measure (Fishlow, 1973) and Emlen's modified entropy measure (Emlen, 1973). . There is a comprehensive theoretical description and comparison for different SM measures in [*Marshall & Olkin,* 1979]. The choice of Picard entropy is substantiated by the relatively good performance in discriminations as shown in Section 5.

## 2.4 Multiscale Schur Monotone Measures

As previously noted, Schur Monotone (SM) measures have been very popular in economics and biology for many years. SM summaries usually serve as a measure of disbalance (or non-uniformity) in an observed vector. Therefore, this measure is expected to provide good discriminative information if the analyzed vectors have different uniformity characteristics. Unfortunately, in some practical examples, the global disbalance (in the time domain) among

the data vectors are too weak to result in powerful discriminatory information. However, if we transform data to the time-scale (wavelet) domain and compare the disbalance at corresponding frequency scales, the discriminatory power of this measure may increase. This increase in sensitivity is apparently due to the unmasking of the balance caused by the interplay of different scale structures and the trends in the data. Through DWT, the data vector is transformed to several wavelet coefficient vectors at different frequency scales (also called resolution levels). Therefore, we are able to define the SM measure at each level, with each measure summarizing the disbalance information of the data vector within distinguishable scales. This natural concept is named the Multiscale Schur Monotone (MSM) measure. The computation of MSM is illustrated in Figure 1.

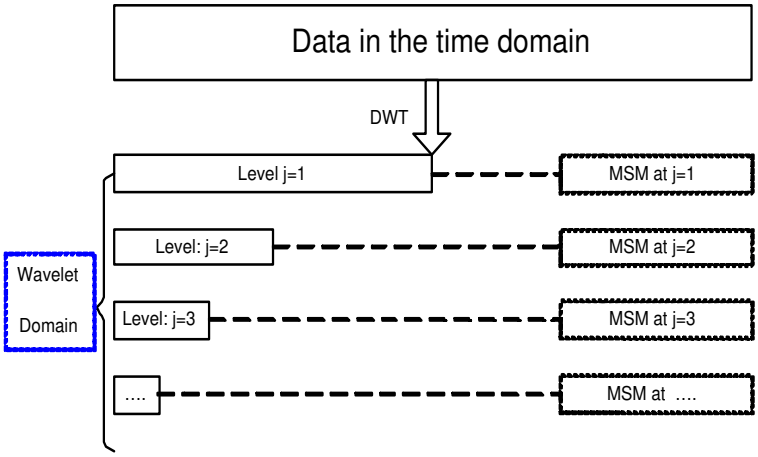


Figure 1: Computation Diagram of Multiscale Schur Monotone Measures.

We provide an example to illustrate the case when the Multiscale Schur Monotone measure is beneficial compared to global, time domain disbalance measures. Assume that the exemplary datasets are simulated by the following two functions:

$$\begin{aligned} f(t) &= \texttt{Doppler} + fGn(H = 0.2); \\ g(t) &= \text{a fixed permutation of } f(t) \end{aligned}$$

where `doppler` is a standard nonstationary testing function commonly used in nonparametric regression [*Donoho & Johnstone*, 1994] and $fGn(H = 0.2)$

is fractional Gaussian noise with Hurst exponent $H = 0.2$ [*Mandelbrot, et al.* 1968]. The time series plots of the typical data simulated from $f(t)$ and $g(t)$ are presented in Figure 2. Clearly, these two functions do not differ from each other as far the Schur Monotone measures in the time domain is concerned, since this measure is invariant with respect to permutation. However, if we map the data into the wavelet domain and compute the MSM measures, the different scale levels show difference in their disbalancing measures. The apparent differences are demonstrated by Figure 3, which is obtained by the analysis of simulated data. We simulated $R = 200$ sample paths from $f(t)$ and $g(t)$ respectively, each of length $N = 2048$. Next, the MSM measures were computed for each sample path. The disbalance measure employed here within each scale in MSM is Picard entropy as defined in 2.3, though other disbalance measures will show similar results. To examine the differences of the MSM measures between $f(t)$ and $g(t)$, we provide the box plots for the MSM at each scale, which are included in Figure 3. The ability to distinguish $f(t)$ and $g(t)$ using MSM measures can be explained by the disbalancing property of DWT. The total inequality exhibited in the time domain is allocated to different frequency scales depending on the correlation structure. Statistically speaking, $f(t)$ tends to have higher values of MSM measures in the first three scales than $g(t)$ with larger probability. This pattern is more pronounced in the finer scales than in the coarser ones because of the smoothing effect of wavelet filtering.

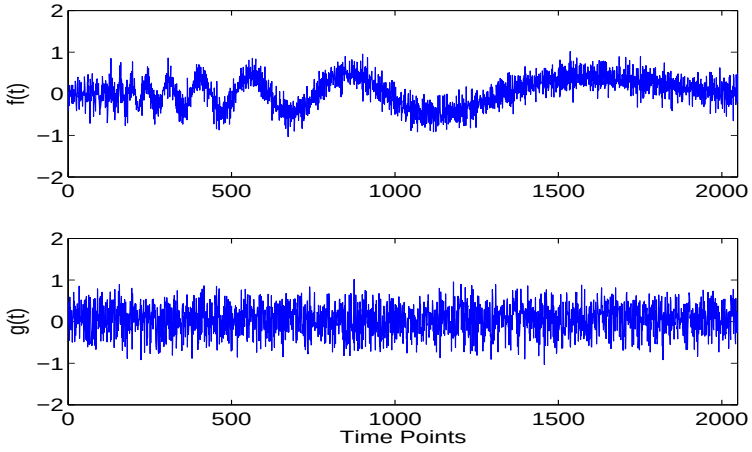


Figure 2: Typical time series plot for the data simulated from functions $f(t)$ and $g(t)$, $t = 1, 2, ..., 2048$..

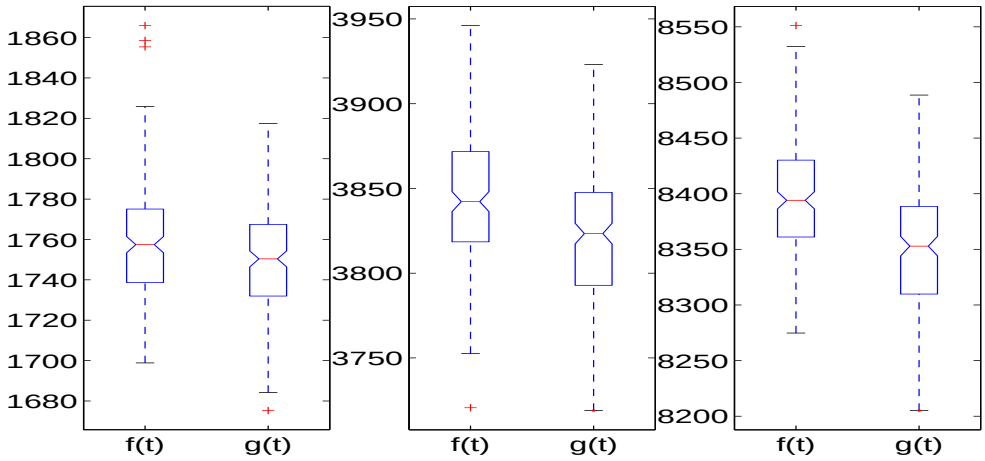


Figure 3: Boxplots of the MSM measures for $R = 200$ replicates of simulated sample paths generated from $f(t)$ and $g(t)$, each with $t = 1, 2, ..., 2048$. The three panels correspond to the first finest three scales. The righthandside panel represents the finest scale, while the lefthandside panel corresponds to the coarsest level.

# 3 $K$-Nearest-Neighbor Classifiers and Their Combinations

The Nearest Neighbor (NN) method is one of the simplest ideas of modeling the regression (or classification) function between the response and predictor variables. It can be expressed as

$$\hat{y}(x_j) = \frac{1}{K} \sum_{x_i \in N_K(x_j)} y_i \tag{2}$$

where $\hat{y}(x_j)$ is the fitted value of the response at $x_j$ and $N_K(x_j)$ is the set containing the first $K$ nearest points to $x_j$ in the predictor variable space. In our classification problem, the response variable $y$ (user group) is categorical and takes only discrete values (e.g., Control, Group 1, etc.). The closeness concept used here is based on Euclidean distance. The Nearest Neighbor method assumes minimal assumptions on the underlying data and is very flexible with respect to finding an arbitrary boundary. The crucial part of $K$-NN modeling is the tuning of parameter $K$. It is well known

[*Hastie, et al,* 2001] that in classification problems with parameter $K = 1$, the misclassification rate is zero for training data set. However, the classification boundary resulting from the Nearest Neighbor method depends very much on the data adequacy of the training set. As a result, the boundary is often very wiggly and unstable in the test set. In other words, the $K$-NN classifiers are often associated with the problem of having a large amount of variance in the prediction for the independent test data set.

An individual classifier usually performs best for certain types of data. However, due to the complexity of certain types of datasets and/or those with a small number of sample paths, the true properties of a population are not able to be fully understood by single classifier. In another words, the inhomogeneous property of the dataset makes it difficult to find a single $K$-NN classifier that optimally fits the data. Although optimal results are not produced from individual classifiers, each classifier describes the dataset by emphasizing certain local aspects of features. It has been observed that the misclassification of data by different single classifiers does not necessarily overlap. Thus, each single classifier has its own values for predicting the classes even if the results are not optimal. The non-overlapped misclassified measurements suggest that those individual classifiers provide complementary information for the prediction. Therefore, a scheme using a combination of the classification results may result in better prediction performance. A diagrammatic representation of the classifier combining procedure is presented in Figure 4. In this paper, the classifiers to be combined are $K$-NN with different tuning parameters $K$. We employed $R = 8$ and $C_1, C_2, ..., C_R$ are $K$-NN classifiers with $K = 3, 4, \ldots, 10$ in our simulation studies afterward.

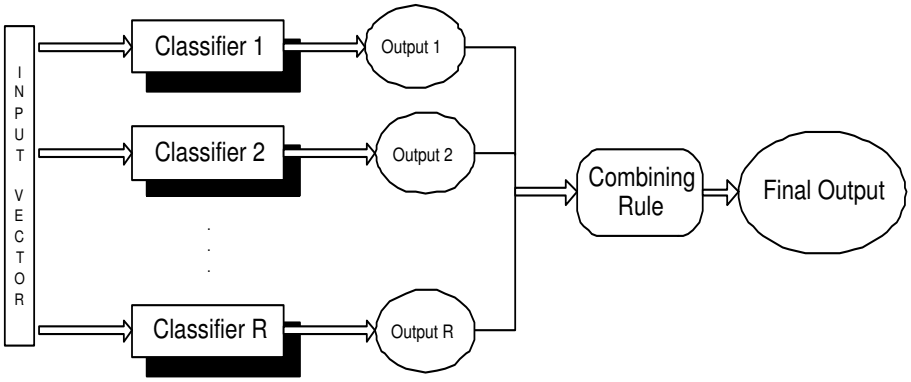


Figure 4: Diagram of Classifier Combining

The combined classifier was originally proposed as an ad hoc procedure, which was then justified by Bayesian decision theory [*Kittler et al,* 1998]. Consider a classification problem where four classes ($y = 0, 1, 2, 3$) are to be distinguished. Suppose that there are $R$ possible classifiers available denoted as $C_1, C_2, ..., C_R$ and each input $x_j$ is assumed to have prior probability $P[y(x_j) = k]$, with $k = 1, 2, ..., 4$ to be concluded correctly from class $k$ regardless of the choice of model. According to Bayesian theory, the predicated class $\widehat{y(x_j)}$ of measurement $j$ with feature vector $x_j$, $j = 1, 2, ..., N$ is

$$\widehat{y(x_j)} = \arg \max_{k \in \{0,1,2,3\}} P[y(x_j) = k | C_1, C_2, ..., C_R] \tag{3}$$

Using Bayes theorem, the posteriori probability in (5) could be expressed as

$$\begin{aligned} P[y(x_j) = k | C_1, C_2, ..., C_R] &= \frac{P[C_1, C_2, ..., C_R | y(x_j) = k] P[y(x_j) = k]}{P[C_1, C_2, ..., C_R]} \\ &= \frac{P[C_1, C_2, ..., C_R | y(x_j) = k] P[y(x_j) = k]}{\sum_{m=0}^{3} P[C_1, C_2, ..., C_R | y(x_j) = m] P[y(x_j) = m]} \end{aligned} \tag{4}$$

Several combination rules can be derived from (4), based on different assumptions on the model probability distribution $P[C_1, C_2, ..., C_R | y(x_j) = k]$ and the prior probability $P[y(x_j) = k]$. These combining rules are summarized in Table 1. The final decision using combined classifiers is

$$\widehat{y(x_j)} = \arg \max_{k \in \{0,1,2,3\}} G(k) \tag{5}$$

where the decision criteria function $G(k)$ can be found in Table 1 corresponding to the different combining rules.

Table 1: Summary of combining rules for multiple classifers

| Rule | Decision Criteria $G(k)$ | Assumptions |
|---|---|---|
| Product | $\frac{\prod_{i=1}^{R} P[y(x_j)=k\|C_i]}{(P[y(x_j)=k])^{R-1}}$ | $P[C_1, C_2, ..., C_R\|y(x_j)=k] = \prod_{i=1}^{R} P[C_i\|y(x_j)=k]$ |
| Mean | $\frac{\sum_{i=1}^{R} P[y(x_j)=k\|C_i]}{R}$ | $P[y(x_j)=k\|C_i] = P[y(x_j)=k](1+\delta_{ki}), \quad \delta_{ki} << 1$ |
| Median | median$\{P[y(x_j)=k\|C_i],$ $i=1,2,...,R\}$ | $P[y(x_j)=k\|C_i] = P[y(x_j)=k](1+\delta_{ki}), \quad \delta_{ki} << 1$<br>Outlier may exist. |
| Max | max$\{P[y(x_j)=k\|C_i],$ $i=1,2,...,R\}$ | $P[y(x_j)=k\|C_i] = P[y(x_j)=k](1+\delta_{ki}), \quad \delta_{ki} << 1$<br>$\frac{\sum_{i=1}^{R} P[y(x_j)=k\|C_i]}{R} \approx \max_{i=1}^{R} P[y(x_j)=k\|C_i]$<br>$P[y(x_j)=0] = P[y(x_j)=1] = ... = P[y(x_j)=3]$ |
| Min | min$\{P[y(x_j)=k\|C_i],$ $i=1,2,...,R\}$ | $P[C_1, C_2, ..., C_R\|y(x_j)=k] = \prod_{i=1}^{R} P[C_i\|y(x_j)=k]$<br>$\frac{\prod_{i=1}^{R} P[y(x_j)=k\|C_i]}{R} \approx \min_{i=1}^{R} P[y(x_j)=k\|C_i]$<br>$P[y(x_j)=0] = P[y(x_j)=1] = ... = P[y(x_j)=3]$ |
| Majority Voting | $\sum_{i=1}^{R} \Delta(y(x_j)=k\|C_i)$ | $P[y(x_j)=k\|C_i] = P[y(x_j)=k](1+\delta_{ki}), \quad \delta_{ki} << 1$<br>$\Delta(y(x_j)=k\|C_i) = \mathrm{I}(k = \arg\ \max_{k\in\{0,1,2,3\}} P(y(x_j)=k\|C_i))$<br>$P[y(x_j)=0] = P[y(x_j)=1] = ... = P[y(x_j)=3]$ |

# 4 High Frequency Pupil Dataset

In this section, we briefly describe the datasets used in this study and how the data was preprocessed to fit the further analysis.

## 4.1 Dataset description

The equipment used to collect the pupillary response data during this study was the Applied Science Laboratories (ASL) Model 501 head-mounted optics system. Pupil size was recorded, at a rate of 60 Hz, for each participant over 105 trials of a computer-based task using a graphical user interface (GUI). A camera records the pupil image, which has been brightened by a near-infrared beam that illuminates the interior of the eye. Pupil size is assessed as the number of pixels attributed to the pupil's image, which has been determined by real-time edge detection processing of the image. Actual pupil diameter measurements (in millimeters) are then calculated by multiplying each pixel count by a scaling factor, which is based on the physical distance of the camera from the participant's eye.

The dataset is comprised of pupillary response data streams for 36 individuals, as described in Table 4. In this table, $N$ refers to the number of individuals comprising this user group. Visual acuity refers to the range of Snellen visual acuity scores of the better eye for participants of each group. AMD? refers to the presence (Yes) or absence (No) of this ocular disease in individuals within each group. Number of data sets refers to the number of 2048-length data sets that were obtained from the data streams for each group. For this study, data was collected from four groups of individuals, classified by visual acuity and the presence or absence of age-related macular degeneration (AMD). Visual acuity, an individual's ability to resolve fine visual detail, was assessed via the protocol outlined in the Early Treatment of Diabetic Retinopathy Study (ETDRS) [*University of Maryland School of Medicine,* 2002]. The experimental protocol from this study is fully described in studies by Jacko and colleagues [**?**].

## 4.2 Preprocessing

Studies of pupillary response are faced with the problem of how to remove blink artifacts. A blink generally lasts about 70-100 msec. (producing an artifact spanning 4-6 observations under 60 Hz sampling) during which time

Table 2: Group characterization summary.

| Group | $N$ | Visual Acuity | AMD? | Number of Data Sets |
|---|---|---|---|---|
| Control | 19 | 20/20 - 20/40 | No | 111 |
| #1 | 6 | 20/20 - 20/50 | Yes | 59 |
| #2 | 5 | 20/60 - 20/100 | Yes | 57 |
| #3 | 6 | $>$ 20/100 | Yes | 124 |

the camera registers a loss and a pupil diameter of zero is recorded. Thus, it is generally relatively straightforward to detect and eliminate these contiguous zero observation artifacts from the record. However, on either side of a blink, one may also observe highly unusual recordings because the pupil may be measured inaccurately when the eyelid partially obscures the pupil. The result may be an impossibly small value for the pupil's size.

To ensure that the analysis is conducted on pupil constriction or dilation and not on misleading discontinuities caused by blinks or partial blinks, one must either remove the blink observations from the data entirely or replace them with linearly interpolated values. Blinks (i.e., zero recordings) have been found to account for approximately 3-4% of all observations, with partial blinks accounting for another 1% of the total number of observations. The blink-removal procedure removes all observations having zero values (i.e., the blink) as well as any extreme values that occur within six additional observations on either side of the zero value (i.e., partial blinks). Figure 5 presents a preprocessed result of the typical measurements from a healthy subject (Control Group).

Because of the difficulty in collecting these measurements, especially from individuals with AMD, the original datasets were cut into equal length pieces to exploit their usage. Another reason for the segmentation is that the original data streams were not equally long among participants. The segmentation is conducted after the 'Six Law' filtering, as mentioned above. The overall dataset contains the sum of 351 segments of measurements, after segmentation and necessary outlier detection, each having a length of 2048 readings. The distribution of the number of the segments among the four groups (Control, #1, #2 and #3) is reported in Table 4.

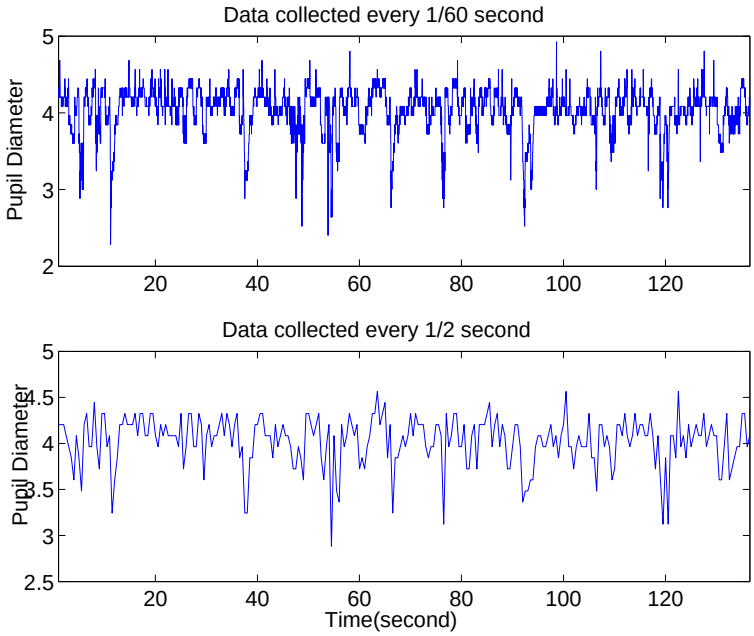


Figure 5: Typical measurements with different resolutions from a heathy subject (Control Group).

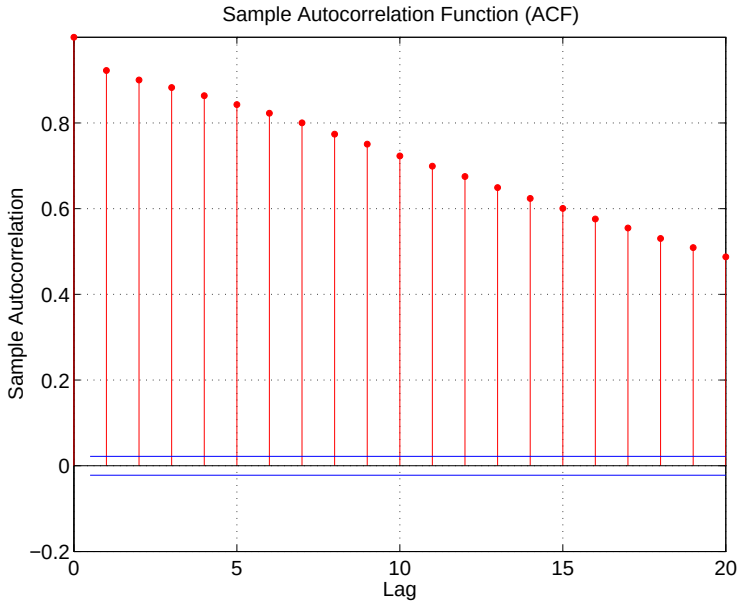


Figure 6: Sample autocorrelation of the measurements presented in Figure 5.

# 5 User Classification using Multiscale Schur Monotone Measures

In this section, we attempt to classify the user groups based on their high frequency pupil diameter measurements. Due to the high dimensionality of these measurements, it is necessary to derive low dimensional summaries from these measurements to characterize these users. The disbalance feature of the pupil-diameter measurements could be a good summary measure to describe the eye behaviors during computer-interaction tasks. However, by simply looking at the statistics of the SM measure in the time domain as shown in Table 3, no significant differences were found to distinguish these four groups. To increase the sensitivity of the disbalance measures, we employed MSM measures. This choice is also motivated by the fact that there are apparent long range correlations within these measurements as displayed in Figure 6. MSM measures characterize the measurement by considering both the disbalance and correlation structure simultaneously, which is not possible in the time domain. Summary statistics of MSM measures are provided in

Table 4. As we can see from this table, the differences among these groups are reflected by the MSM measures, especially at the finer scales. For example, at the finest scale, level 1, the mean MSM measures of group #1 and #2 are much smaller than that of control group. Group #3 tends to have similar mean disbalance at level 1 as the control group while they apparently have different medians. These results are interesting, as the MSM measures may provide evidence of the erratic effect that ocular disease (in the case of Groups 1 and 2), particularly central field deficiencies, have on pupillary response behavior, as previously discussed. While the distinction between Groups 1 and 2 and the Control group is expected, given the presence of ocular disease in the experimental groups, the similarity of the Control group and Group 3 is unexpected. As can be seen in Table 2, the Control group and Group 3 are the most diverse with respect to both the presence of AMD and the level of visual acuity. However, the MSM measures at level 2 provide considerably more distance between these two groups.

The addressed differences in the MSM measures imply the discriminatory information. To fully integrate this information, we propose to use the combined $K$-NN classifiers to develop a statistical classification procedure to automatically distinguish the MSM measures of the different user groups.

Table 3: Summary statistics of Schur Monotone Measures ($10^4$) in the time domain.

| | Control | #1 | #2 | #3 |
|---|---|---|---|---|
| Min | 1.563 | 1.5624 | 1.5625 | 1.5632 |
| Mean | 1.5619 | 1.5617 | 1.5618 | 1.5618 |
| Median | 1.5617 | 1.5617 | 1.5617 | 1.5618 |
| Max | 1.5616 | 1.5616 | 1.5616 | 1.5616 |
| Std. Dev. | 0.0003 | 0.0002 | 0.0002 | 0.0002 |

A 5-fold cross-validation scheme is used to guarantee the robustness of our procedure. The datasets described in Section 4 are randomly divided into two parts: 80% of the measurements are used as training set, which is used to estimate the classification model; and the remaining 20% of the measurements are regarded as test set, used to validate the classification model. The default scheme of combining $K$-NN classifiers (see Section 3) is employed to classify the MSM measures computed from pupil diameter measurements. The cross-validation is repeated twenty times in order to

Table 4: Group characterization summary in terms of the Schur Monotone measures in the wavelet domain. Level 1-4 represent the first, second, third and fourth finest scales respectively.

| | Statistics | Level4 | Level3 | Level2 | Level1 |
|---|---|---|---|---|---|
| Control | Min | 583.02 | 1358.64 | 5467.64 | 19404.85 |
| Group | Mean | 470.32 | 1100.71 | 3739.06 | 15526.69 |
| | Median | 459.57 | 1093.07 | 3629.92 | 15703.12 |
| | Max | 390.71 | 944.39 | 2861.20 | 12335.15 |
| | Std Dev. | 54.45 | 106.53 | 578.94 | 2154.64 |
| Group | Min | 564.35 | 1559.91 | 5169.51 | 19261.05 |
| #1 | Mean | 447.78 | 1073.71 | 3458.55 | 13626.51 |
| | Median | 444.43 | 1006.24 | 3282.31 | 14814.35 |
| | Max | 389.91 | 875.12 | 1954.61 | 6770.97 |
| | Std Dev. | 37.34 | 166.40 | 1001.68 | 3610.46 |
| Group | Min | 516.56 | 1297.45 | 4777.68 | 18714.94 |
| #2 | Mean | 429.85 | 1030.24 | 3423.02 | 14160.38 |
| | Median | 431.18 | 1024.13 | 3538.62 | 15387.57 |
| | Max | 380.67 | 885.53 | 2187.12 | 7871.65 |
| | Std Dev. | 30.24 | 79.17 | 728.43 | 3287.20 |
| Group | Min | 733.43 | 2366.90 | 9005.10 | 25460.81 |
| #3 | Mean | 452.76 | 1304.61 | 4738.77 | 15989.98 |
| | Median | 432.37 | 1198.93 | 4800.60 | 18413.07 |
| | Max | 375.40 | 872.74 | 1933.66 | 4266.89 |
| | Std. Dev. | 58.81 | 385.70 | 2215.99 | 7435.45 |

estimate the mean and standard deviation of the misclassification rate. The classification result is summarized in Table 5. The MSM measures at the three finest scales are considered here as the input vector in the classification model and the wavelet basis used to conduct DWT is Daubechies wavelet with two vanishing moment. The product combining rule, which assumes independence among the classifiers to be combined, does not work very well in our study because we combine a family of $K$-NN classifiers with different tuning parameters, which are quite likely to depend on each other. Therefore, the results for the product combining rule are reported here and thereafter. The rest of the combining rules mentioned in Section 3 work comparably in terms of the misclassification rate for the test set, although the MAX rule seems to slightly outperform the others.

Table 5: Error rates after combining the Nearest Neighbor classifiers using MSM measures at the three finest levels.

| | rule | mean | median | **max** | min | majority voting |
|---|---|---|---|---|---|---|
| Test | avg. | 0.2566 | 0.2592 | **0.2408** | 0.2592 | 0.25 |
| Error | std. | 0.0345 | 0.0344 | 0.036 | 0.0382 | 0.0462 |
| Training | avg. | 0.2474 | 0.2551 | **0.245** | 0.2257 | 0.2489 |
| error | std. | 0.0098 | 0.0089 | 0.0104 | 0.0166 | 0.0093 |

# 6 Discussions

In this section, we discuss the possibility of improving the classification performance of the default model used in Section 5 by choosing the coarsest level and wavelet basis in DWT.

## 6.1 Choice of the Coarsest Level

The number of scales included in MSM measures is a parameter to be decided in our classification model. This is equivalent to choose the coarsest level in DWT, which affects the size of the input vector and is further related to the fitting quality of the classifier. More than enough scales of the coarsest level may result in overfitting, while not enough DWT levels results in oversmoothing. Table 7 illustrates the fact that if we use MSM measures from the two finest scales, the performance will be decremented at least 3%,

which implies an underfitting case. On the other hand, if we include MSM measures at the first four scales, the performance will still be decremented at least 2% - a case of overfitting. Therefore, MSM measures at the first three finest scales are the optimal choice for our pupil diameter classification.

Table 6: Error rates after combining the Nearest Neighbor classifiers using MSM measures at the finest two levels (underfitting).

| | rule | **mean** | median | max | min | majority voting |
|---|---|---|---|---|---|---|
| test | avg. | **0.2763** | 0.2796 | 0.2875 | 0.2855 | 0.2987 |
| error | std. | 0.0347 | 0.0316 | 0.035 | 0.0334 | 0.0328 |
| training | avg. | **0.232** | 0.236 | 0.2465 | 0.232 | 0.2523 |
| error | std. | 0.0105 | 0.0088 | 0.0112 | 0.0097 | 0.0081 |

Table 7: Error rates after combining the Nearest Neighbor classifiers using MSM measures at the finest four levels (overfitting).

| | rule | **mean** | median | max | min | majority voting |
|---|---|---|---|---|---|---|
| test | avg. | **0.2605** | 0.2664 | 0.2678 | 0.2711 | 0.2704 |
| error | std. | 0.0463 | 0.0471 | 0.043 | 0.0442 | 0.0504 |
| training | avg. | **0.2342** | 0.2384 | 0.2391 | 0.223 | 0.237 |
| error | std. | 0.0116 | 0.0123 | 0.0134 | 0.0127 | 0.0116 |

## 6.2 Wavelet Basis Selection

The wavelet basis has substantial influence on the transformed coefficients of pupil diameter measurements and is, therefore, an important factor in determining the classifier quality. We formulate an optimization study to search for the best wavelet basis, which results in the most accurate classification. The search will be limited to the Pollen wavelets library. Pollen wavelets are a family of wavelet basis with a continuous mapping from $[0, 2\pi]^{N-1}$ to a set of "wavelet solutions" in terms of the quadratic mirror filters of $h = \{h_0, h_1, \cdots, h_{2N-1}\}$, where $N$ is the number of vanishing moments. Pollen representation of all wavelet solutions of length 4 ($N = 2$) and length 6 ($N = 3$) is given in Tables 8 and 9. The Daubechies wavelet family is included in the Pollen library as a special case.

There are many measures of classifier performance. Some popular measures include scatter-matrix and Bayesian risk, among others. Though these

Table 8: Pollen paramterization for $N = 2$ (four-tap filters).$[s = 2\sqrt{2}]$.

| $n$ | $h_n$ for $N = 2$ |
|---|---|
| 0 | $(1 + \cos(\theta) - \sin(\theta))/s$ |
| 1 | $(1 + \cos(\theta) + \sin(\theta))/s$ |
| 2 | $(1 - \cos(\theta) + \sin(\theta))/s$ |
| 3 | $(1 - \cos(\theta) - \sin(\theta))/s$ |

Table 9: Pollen paramterization for $N = 3$ (six-tap filters).$[s = 2\sqrt{2}]$.

| $n$ | $h_n$ for $N = 3$ |
|---|---|
| 0 | $[1 + \cos(\theta_1) + \cos(\theta_2) + \sin(\theta_1) - \sin(\theta_2) - \cos(\theta_1 - \theta_2) - \sin(\theta_1 - \theta_2)]/2s$ |
| 1 | $[1 - \cos(\theta_1) + \cos(\theta_2) + \sin(\theta_1) - \sin(\theta_2) - \cos(\theta_1 - \theta_2) + \sin(\theta_1 - \theta_2)]/2s$ |
| 2 | $[1 + \cos(\theta_1 - \theta_2) + \sin(\theta_1 - \theta_2)]/s$ |
| 3 | $[1 + \cos(\theta_1 - \theta_2) - \sin(\theta_1 - \theta_2)]/s$ |
| 4 | $[1 - \cos(\theta_1) + \cos(\theta_2) - \sin(\theta_1) + \sin(\theta_2) - \cos(\theta_1 - \theta_2) - \sin(\theta_1 - \theta_2)]/2s$ |
| 5 | $[1 + \cos(\theta_1) - \cos(\theta_2) - \sin(\theta_1) + \sin(\theta_2) - \cos(\theta_1 - \theta_2) + \sin(\theta_1 - \theta_2)]/2s$ |

separability measures are optimal (or almost optimal) under certain assumptions, computational issues like matrix inversion and prior statistical knowledge about the data often make this impractical. For a detailed discussion of these measures, the readers are directed to [*Fukunaga,* 1990]. A more practical and easily implemented measure of the separability is the misclassification rate based on the input vector associated with the wavelet filter $H = (h_0, h_1, ..., h_n)$.

As a result, our search procedure focuses on minimizing the misclassification rate in the test set with respect to the wavelet filters. The first search is done in the Pollen library with $N = 2$. The results are presented in Figure 7. As shown in this figure, the performance varies up to about 9% with different values of $\theta$ and the best performance is achieved around $\theta = 100°$. The scaling and wavelet functions corresponding to this optimal Pollen wavelet basis are plotted in Figure 8. To compare the performance of the different pollen wavelet basis with a different number of vanishing moments, the search is conducted for Pollen wavelets with $N = 3$. The results are shown in Figure 9. For $N = 3$, there is more variability in performance with these different parameters than those in the case of $N = 2$, resulting in worse overall performance compared with $N = 2$. This can be attributed to the locality and smoothness of the wavelet bases. The Pollen wavelet with $N = 3$ is

smoother and, hence, tends to smooth the data more than $N = 2$. It may be the case that some of the discriminatory information has been smoothed, which causes the classification performance to become worse.

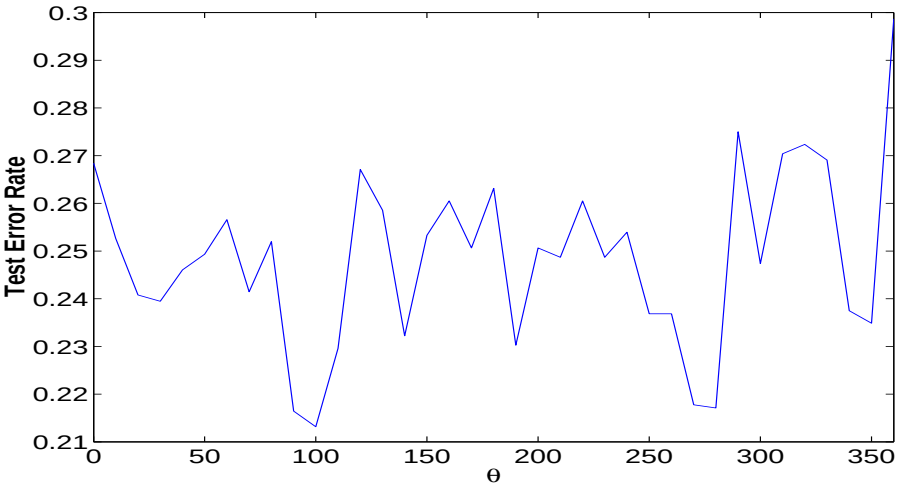


Figure 7: Misclassification rates using a Pollen wavelet basis with different parameter $\theta \in [0, 2\pi]$. The classifiers here are the combined $k$-NN with MsSC measure input vectors. The error rates shown in the figure are the average values for 20 randomly selections of test set from the whole dataset. The minimum error rate here is 21.32%, which is achieved at $\theta = 100°$.

# 7 Conclusions and Future Work

The classification procedure for the user group is developed utilizing the Multiscale Schur Monotone measures in the wavelet domain and ad hoc classifier combination schemes. We investigated the performance of different Multiscale Schur Monotone measures in this particular user classification problem. It turns out that the Picard's entropy measure works best among the considered candidates. We also considered the stabilization of misclassification rates by using combinations of single basic classifiers. This heuristic procedure implies some approximation of Bayesian model averaging. Our user classification example validates this procedure through the relatively low misclassification rate, which resulted in the randomly selected test set. Additionally, we studied the problem of searching for the optimal

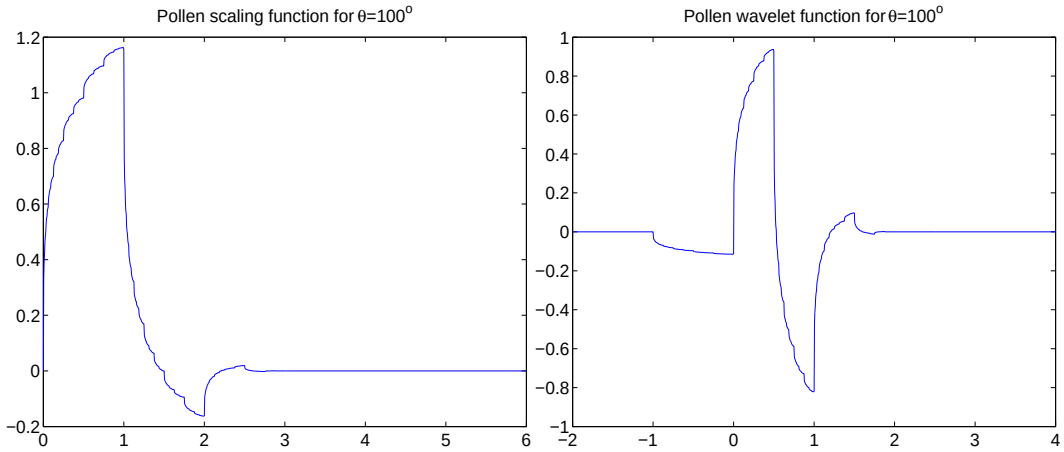


Figure 8: GT wavelet basis (Four-tap Pollen wavelet basis with $\theta = 100^o$).

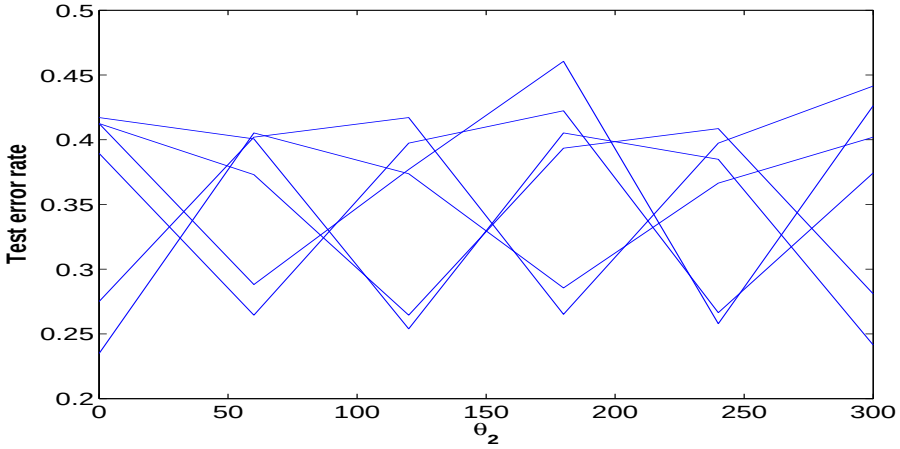


Figure 9: Misclassification rates using Pollen wavelet basis with different parameter $\theta \in [0, 2\pi]^2$. The classifiers here are the combined $k$-NN with MsSC measure input vectors. The error rates shown in the figure are the average values for 20 randomly selections of test set from the whole dataset. The error rates here are obviously quite larger than 4-tap Pollen wavelet filter. The 6 curves correspond to different values of $\theta_1$.

wavelet basis among certain candidate wavelet families. Those families includes Daubechies and Pollen. In the Pollen wavelet family (limited to four tap filters), we found that the basis with parameter $\phi = 100^\circ$ achieves the best classification performance for the pupil-diameter measurements. Overall, the expected misclassification could be at least around 21% by choosing the appropriate wavelet basis and Multiscale Schur Monotone measure. This exciting result is of much importance in the HCI community.

The utility of these analytical tools for applied research in HCI has tremendous potential, as user classification is of primary importance in this field of research. The use of novel statistical methods, as shown in this paper, shows promise for the ability to use complex physiological data from users to better understand their unique needs and behaviors. While further data collection is needed to help increase the amount of data being analyzed, the initial results suggest that the presence of ocular disease and/or acuity loss does result in dynamic, complex differences in pupil behavior. In essence, MSM measures can be used to 'tease-out' differences in the pupillary behavior of individuals with and without ocular disease, possessing a range of visual acuity. The results not only show the fairly reliable classification or distinction of individuals with and without ocular disease (AMD), as the separation of the Control Group and Group 1 illustrates, but the results also illustrate finer distinctions amongst groups with a similar ocular disease, but with varying visual functioning (e.g., visual acuity), as the separation of Groups 1, 2, and 3 illustrates. This ability to separate these groups, based on dynamic pupillary behavior, illustrates the usefulness of these analytical procedures for user classification.

One of the overreaching goals of this study was to examine the use of high-frequency pupillary behavior as a means of quantifiably assessing differences between users during performance of a computer-based task. The results of this study show great potential toward this goal, as MSM measures were used to distinguish the user groups. This distinction between user groups was used to generate a promising predictive model of user classification. The future implications of this study include the application of these, and similar, analytical tools for other high frequency physiological data, such as eye movement, heart rate and brain signals.

# References


[*Backs & Walrath,* 1992] Backs, R.W. & Walrath, L.C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied Ergonomics, 23, 243-254.

[*Barbur,* 2003] Barbur, J. L. (2003). Learning from the pupil - studies of basic measurement mechanisms and clinical applications. L. M. Chalupa & J.S. Werner (Eds.), Cambridge, MA: MIT Press, in press.

[*Biglan et al,* 1988] Biglan, A. W., Van Hasselt, V. B., & Simon, J. (1988). Visual impairment. In V. B. Van Hasselt (Ed.), *Handbook of Developmental and Physical Disabilities* pp471-502 New York, NY: Pergamon Press.

[*Donoho & Johnstone*, 1994] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425-455.

[*Fukunaga,* 1990] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* NY:Academics.

[*Hastie, et al,* 2001] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, New York.

[*Hess & Polt*, 1960] Hess, E. H. & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, **132**, 349-350.

[*Hess & Polt*, 1964] Hess, E. H. & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem solving. *Science*, **143**, 1190-1192.

[*Jacko et al,* 2000] Jacko, J. A., Barreto, A. B., Marmet, G. J., et al. (2000). Low vision: The role of visual acuity in the efficiency of cursor movement. *Paper presented at the Fourth International ACM Conference on Assistive Technologies (ASSETS 2000).*

[*Jacko et al,* 2002] Jacko, J. A., Barreto, A. B., Scott, et al. (2002). Macular degeneration and visual icon use: deriving guidelines for improved access. *Universal Access in the Information Society*, 1(3), 197-206.

[*Jacko et al,* 1999] Jacko, J. A., Dixon, M. A. et al.(1999). Visual profiles: A critical component of universal access. *Paper presented at the Human Factors in Computing Systems*, Pittsburg, PA.

[*Jacko et al,* 2000b] Jacko, J. A., Rosa, R. H. et al(2000). Visual impairment: The use of visual profiles in evaluations of icon use in computer-based tasks. *International Journal of Human-Computer Interaction*, 12(1), 151-164.

[*Kittler et al,* 1998] Kittler, J. et al (1998), On Combining Classifiers, *IEEE Transactions on pattern analysis and machine learning intelligence* , 20(3):228-239.

[*Loewenfeld,* 1999] Loewenfeld, I. E. (1999). The pupil: Anatomy, physiology, and clinical applications (2nd ed.). Oxford, UK: Butterworth-Heinemann.

[*Mallat*,1989] Mallat, S. (1989). A Theory for Multiresolution Signal Decomposition: the Wavelet Representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**:674-693.

[*Mandelbrot, et al.* 1968] Madndelbrot, B. et al. (1968), Fractional Brownian Motion, Fractional Noise and Applications, *SIAM review*, 10, 422-437, 1968.

[*Marshall & Olkin,* 1979] Marshall, A. & Olkin, I. (1979). *Inequalities: theory of Majorization and its applications,* Academic Press.

[*Sahraie & Barbur*, 1997] Sahraie, A., & Barbur, J. L. (1997). Pupil response triggered by the onset of coherent motion. *Graefes Archives of Clincal Experimental Ophthalmology*, **235**, 494-500.

[*University of Maryland School of Medicine,* 2002] University of Maryland School of Medicine, Department of Epidemiology and Preventative Medicine. (1980). Early treatment diabetic retinopathy study, Manual of Operations, Chapter 12 (pp. 1-15). Baltimore, MD: ETDRS Coordinating Center.

[*Van Gerven et al.*, 2004] Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, **41**, 167-174.

[*Vidakovic,* 1999] Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, John Wiley & Sons, Inc., New York, 384 pp.