GENERAL INSTRUCTION

- **Authors: Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.**
- **Authors: Please check ALL author names for correct spelling, abbreviations, and order of first and last name in the byline and affiliation footnote.**
- **Authors: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.**
- **Authors: Please note that once you click "approve with no changes," the proofing process is now complete and your article will be sent for final publication and printing. Once your article is posted on Xplore, it is considered final and the article of record. No further changes will be allowed at this point so please ensure scrutiny of your final proof.**
- **Authors: Unless invited or otherwise informed, a mandatory Excessive Article Length charges will be incurred if your article is over the page limit set by the society in the Information for Authors.**

QUERIES

Q1. Author: Please provide the postal code for the affiliation of the author Marco Miranda in the first footnote.
Q2. Author: Please update Ref. 29.

Trail-Traced Threshold Test (T4) With a Weighted Binomial Distribution for a Psychophysical Test

Yuxin Gong, Haogang Zhu, Marco Miranda, David P. Crabb, Haolan Yang, Wei Bi, and David F. Garway-Heath

***Abstract*—Clinical visual field testing is performed with commercial perimetric devices and employs psychophysical techniques to obtain thresholds of the differential light sensitivity (DLS) at multiple retinal locations. Current thresholding algorithms are relatively inefficient and tough to get satisfied test accuracy, stability concurrently. Thus, we propose a novel Bayesian perimetric threshold method called the Trail-Traced Threshold Test (T4), which can better address the dependence of the initial threshold estimation and achieve significant improvement in the test accuracy and variability while also decreasing the number of presentations compared with Zippy Estimation by Sequential Testing (ZEST) and FT. This study compares T4 with ZEST and FT regarding presentation number, mean absolute difference (MAD between the real Visual field result and the simulate result), and measurement variability. T4 uses the complete response sequence with the spatially weighted neighbor responses to achieve better accuracy and precision than ZEST, FT, SWeLZ, and with significantly fewer stimulus presentations. T4 is also more robust to inaccurate initial threshold estimation than other methods, which is an advantage in subjective methods, such as in clinical perimetry. This method also has the potential for using in other psychophysical tests.**

***Index Terms*—Bayesian, perimetric threshold test, spatial weight, standard automated perimetry, visual field.**

Manuscript received August 2, 2020; revised December 17, 2020; accepted January 28, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61702027 and in part by the Major Project of Science and Technology of Yunnan Province under Grant 2019ZE005 and in part by research funded by the National Institute for Health Research (NIHR) under its Invention for Innovation (i4i) program under Grant II-LA-0813-20004. (*Corresponding author: Haogang Zhu.*)

Yuxin Gong is with the School of Biological Science and medical Engineering, Beihang University, Beijing 100191, China (e-mail: gongyuxinbuaa@163.com).

Haogang Zhu is with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 10019, China, and also with the Beijing Advanced Innovation Centre for Big Data-Based Precision Medicine, Beihang University, Beijing 10019, China (e-mail: haogangzhu@buaa.edu.cn).

Marco Miranda and David F. Garway-Heath are with the Faculty of Brain Sciences, Visual Neurosciences, Institute of Ophthalmology, University College London, London WC1E 6BT, U.K., and also with the NIHR Biomedical Research Centre, Moorfields Eye Hospital and University College London Institute of Ophthalmology, London, U.K. (e-mail: m.miranda@ucl.ac.uk; david.crabb.1@city.ac.uk).

Haolan Yang is with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 10019, China (e-mail: haolanyang@buaa.edu.cn).

Wei Bi is with the Zsbatech Corporation, Beijing 100011, China.

David P. Crabb is with the School of Health Sciences, City University London, Northampton FTuare EC1V 0HB, U.K. (e-mail: d.garwayheath@nhs.net).

Digital Object Identifier 10.1109/JBHI.2021.3057437

I. Introduction

PSYCHOPHYSICS is the scientific study of the relationship between the physical properties of sensory stimuli and the behavioral sensations and perceptions that are elicited by these stimuli. Psychophysical tests are widely used in many fields, such as audiology [1], vision [2], [3], taste and smell [4], and pain [5], by designing methods to obtain estimates of psychophysical functions describing processes of underlying sensory mechanisms [6]. The psychophysical function depicts the probability of a stimulus being detected. It's S-shape [7], [8] can be described by parameters such as the threshold and slope, which can serve as disease and variability quantifiers.

In vision and hearing studies, it is practical to measure the sensitivity with many trials using computer-generated stimuli. In contrast, for the chemical-based senses, the physical presentation of the stimulus is not easily accomplished without human intervention, and the longer recovery time of the chemical senses prevents the rapid successive presentation of stimuli [4]. These factors limit the number of psychophysical trials in a testing session before fatigue and boredom set in [9].

Many eye diseases, such as glaucoma, show evidence of their initial deficits in the periphery. Moreover, the pattern, shape and location of visual field deficits can indicate the most likely location of damage to the visual pathways, and the effectiveness of a treatment can be monitored by testing the visual field. Standard automated perimetry (SAP) is used in the diagnosis and monitoring of glaucoma and other diseases affecting vision. It can measure the differential light sensitivity (DLS) across a person's retina and the corresponding visual pathway [10]; an illustration is shown in Fig. 1.

Visual field testing is performed with commercial perimetric devices and employs psychophysical techniques to obtain DLS thresholds at multiple retinal locations [11], which is a subjective test that aims to measure a sensitivity threshold in a living

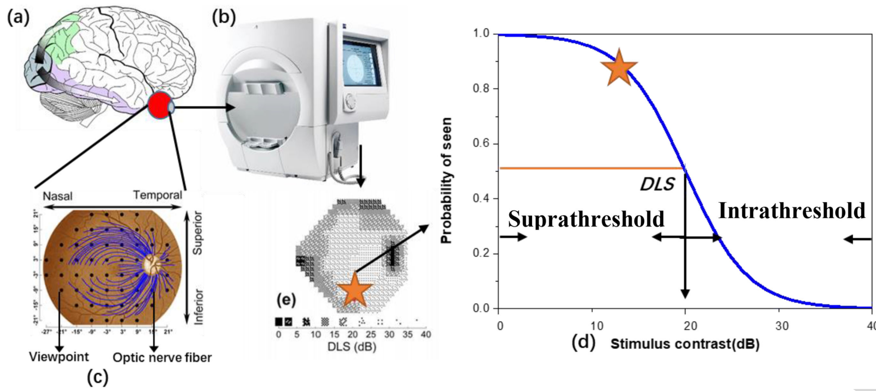

Fig. 1. (a) SAP measuring the differential light sensitivity (DLS) of the retina and corresponding visual pathway. (b) Contrast stimulus from SAP is projected on different locations of the retina. The response from a subject is captured when the stimulus is perceived. (c) The DLSs are measured at various locations (dots) on the retina. The eye ball using24-2 to divide into 54 viewpoints, which interval between horizontal and vertical is 6 degrees and only 52 points get analyzed. The point $(0^\circ, 0^\circ)$ indicates central vision that corresponds to the fovea on the retina. The optic nerve head is the anatomical blind spot. The test locations are correlated with not only their neighbors but also the optic nerve fibers (some of which are represented by blue curves) passing through them. (d) The DLS threshold at a location on the retina is derived at the 50% probability of the visual system responding to a contrast stimulus. (e) The DLS ranges between 0 dB (high contrast stimulus, damage) and approximately 35 dB (low contrast stimulus, healthy) and it can be displayed as a grayscale, where the darker shading represents a lower DLS.

organism and is prone to variability. Besides, it is also easily affected by many factors, such as patient motivation, fatigue and attention and technician performance. Thus, an ideal perimetric threshold algorithm in visual field testing should reduce the testing time without losing the testing accuracy, and it should also be robust to mistakes made while testing. Patient's erroneous answers increase test times and may result in fatigue artifacts that decrease in the quality of the threshold estimates [12]. Unfortunately, the development of computational and statistical methods for analyzing data from SAP has not kept pace with advances in other aspects of eye-related research [10]. Early versions of algorithms for perimetric threshold tests are based on a computationally simple staircase strategy, such as The full threshold (FT) strategy [13] and FASTPAC algorithms [14], and have been studied in detail using both computer simulation and clinical studies [15]–[18]. However, these methods have the drawback that the improvement in the accuracy is at the expense of an increase in the examination duration (test presentation), which can lead to unstable results from incorrect patient responses [19]. Besides, it uses fixed steps to achieve threshold estimation, which is time consuming and inefficient to recover from errors caused by incorrect patient responses. To decrease the test presentation and improving the test accuracy, Watson and Pelli [20] developed a new perimetric algorithm based on Bayesian adaptive threshold procedures. The Bayesian method combines prior knowledge about the expected distribution of the thresholds. The initial or prior probability density function (PDF) and each response made by the patient (in the case of perimetry, these are "seen" or "not seen") are used to alter the expected distribution of the final thresholds (subsequent or posterior PDF) [21]. The family of Swedish interactive threshold algorithms (SITAs)and ZEST are three popular methods from which SITA use both a staircase and maximum likelihood methods [22]–[24], the ZEST algorithm is merely based on maximum likelihood procedures and is computationally simpler than that of SITA [25]–[28]. Although SITA and ZEST could reduce the test time and improve the test accuracy compared with the traditional FT algorithms, the ideal balance between both parameters is still difficult to achieve. Noted that the SITA-faster is much shorter with about the same precision that SITA, it can better get the balance between test accuracy and test time than SITA-fast and SITA-standard, but its variability remains high in the threshold methods.

The Bayesian methods, such as ZEST, have several drawbacks that limit their capability to achieve satisfactory test performance. First, The ZEST doesn't notice the spatial information in the perimetric testing, which describe as an algorithm to threshold a single location in the visual field, not be used at multiple locations. Besides, the fixed shape of the likelihood is another drawback for ZEST, means that the amount of information obtained in each measurement round is completely equivalent, which is not reasonable. In fact, the likelihood function is related to the previous threshold measurement result (patient's threshold estimates and variance), should be nonstationary (heteroscedastic) since we want to modify the optimal threshold estimate with a substantial correction when we have large confidence, and vice versa. Thus, it is necessary to optimize the likelihood function by correcting its distribution using each feedback message from the patient. This can reduce test duration and improve test error performance significantly. To solve these problem, Nikki J. Rubinstein propose SWELZ [29] to reduce test presentation without affecting test accuracy and stability by incorporating spatial information to ZEST. SWeLZ extends the ZEST procedure to update visual sensitivity estimates across multiple locations after each test presentation, and using the spatial weight between current and its neighbor test points to scale the likelihood function of the neighbor test points to update current and its neighbor test points concurrently.

However, this method still dependent on the accurate initial threshold estimate, which is difficult to satisfy in visual field testing; Here, the initial threshold estimation means using previous measurement data to get PDF firstly, and then get an average value for the PDF regarded as the initial threshold. The underestimation or overestimation of the initial threshold may reduce the accuracy and increase the duration of the test [25]. When the initial threshold is inaccurate, the spatial weight will scale the shape of likelihood function for the neighbor test points at the wrong direction, increasing the measurement error of adjacent points. Besides, this method only decrease the test presentation without improving the test accuracy. Kucur proposes a meta-strategy, SORS, capable of using traditional staircase methods or ZEST-like Bayesian strategies at individual locations but in a more efficient and faster manner. In essence, determines which locations should be chosen and in what order they should be evaluated in order to maximally improve the visual field estimate in the least amount of time [30]. Montesano also proposes MacS-ZEST that it uses the detailed two-dimensional structural information provided by macular SD-OCT scans to build a structure-function model for the macula that could be easily employed to inform perimetric testing [31]. In brief, it is a novel approach for structure-function modeling in glaucoma to improve visual field testing in the macula.

Although, such development for ZEST get the improvement in test presentation and accuracy. However, ZEST-related methods still depend on the accurate initial threshold estimate. Theoretically, an ideal visual field testing algorithm does not require an accurate extensive priors derive from big dataset and could be easily adapted to quickly and accurately measure a variety of psychometric functions would provide an enormous benefit to the psychometrics community [32]. Thus, we propose a new perimetric threshold method, called T4, which uses the spatial filter for the spatial connections, combining retinotopic and optic nerve head topic spatial relationships in one metric, and incorporating the spatial weight combine with varying likelihood function based on 6 and binomial probabilities to update multiple location concurrently. Different from scaled-likelihood function of SWeLZ, when a spatial weight decreases, the likelihood function used by SWeLZ become flat (scale compressed in y-axis) but the shape (in x-axis) don't change. In comparison, the proposed likelihood function keeps scale the same (always between 0 and 1) but varies in shape (stretched in x-axis, see Fig. 6). This is useful to improve test accuracy and stability further. Besides, T4 also proposed a new update rule (maximization of 7), which is different with SWELZ. Because SWELZ uses the spatial weight to update neighbor test points not using the spatial weight to help updating current test points. This make T4 can decrease test presentations without decreasing test accuracy and stability compared with ZEST. The most contribution for clinical application is that the initial distribution of T4 is similar with uniform distribution, which make it does not need accurate prior.

This study also compares T4 with ZEST and FT, by evaluating the test presentations, the accuracy, and the test-retest variability between two test results. Meanwhile, we do several verification experiments to explore which part i.e., the proposed varying likelihood function, spatial filter or update rule, is the biggest effect on improving test performance compare with Scale-likelihood function and spatial weight introduced by SWeLZ and the ZEST update rule. The experiments show that T4 significantly outperforms other popular algorithms in terms of test presentation, test accuracy, and test variability. Moreover, T4 showed robust performance when the initial threshold estimate is uniform distribution. Noted that the robust means T4 can get better test error and test stability robustly compared with other two methods not the tolerance when FP increasing.

II. Experiment Setup

A. Overall Description of the Computer Simulation

In the real world, it is difficult to assess the precise error in test results acquired from an algorithm since the exact visual field sensitivity of any patient is unknown. Thus, to verify the three algorithms precisely, computer simulations were used to simulate all the subjects by considering the true distribution of patients' sensitivity and the measurement error caused by individual mistakes, which can be described by the FP and FN, respectively. The patient response to a stimulus at level s was simulated using a frequency-of-seen (FOS) curve defined by:

$$FOS\left(s,v,\delta\right) = 1 - FN - \left(1 - FN - FP\right)\phi\left(s\,|v,\delta\right)\phi\left(s\,|v,\delta\right) \tag{1}$$

Where FN is the false negative response rate while FP is the false positive response rate so as to measure the variability of the patient's response. $\phi(s|v,\delta)$ is the cumulative Gaussian distribution with mean ν and standard deviation (SD) δ, where the mean ν is the level of the true threshold and δ was set to $\min(e^{-0.081v}{+}3.27, 6)$ according to an empirical test [33] because the variance is 6 for locations with a low DLS threshold and gradually decreases with increasing DLS threshold.

$$\phi\left(s\,|v,\delta\right) = \min\left(e^{-0.081v}{+}3.27, 6\right) \tag{2}$$

This simulates the known change in variance at different levels of DLS, hence simulating patient's visual function variance, which is higher for low DLS threshold and lower for high DLS threshold. Moreover, it can also avoid the patient's visual function variance being too high for low DLS. Then, we simulated three types of patient variability by modifying the FP to 5%, 10% and 15%, which represent patients with low, medium and high variability, respectively. The FN was fixed at 5%. By inputting all the initial parameters, we acquired the FOS curve at each DLS level, which represents the patient's response at a certain level according to the FOS rate.

B. Dataset

In this paper, a test-retest dataset, named RAPID dataset, is used which consisting of 218 eyes from 109 glaucoma patients, each of which underwent 10 Humphrey Field Analyzer (HFA) 24-2 visual field tests within 8 weeks. It is assumed that there is no measurable change during the 8 weeks and that the visual ability of any patient is stable, which ensures that the difference among the measurements for the same eye is due to the measurement variability without other effect disturbances. Thus, the average value for the 8 visual fields result can be regarded as the underlying true visual field. To verify that T4 outperforms

TABLE I
THE RAPID DATASET INFORMATION

Characteristics	Median	5th to 95 th percentile
Age(years)	70.3	50.0 to 85.6
IOP(mmHg)	14.0	8.0 to 21.0
SAP MD (dB)	-4.17	-14.22 to 0.88
RNFL thickness(μ)	69.0	45.1 to 95.6
Visual acuity (Snellen)	6/6	6/4 to 6/12
Refractive error(dioptres)	-0.13	-7.48 to 2.95

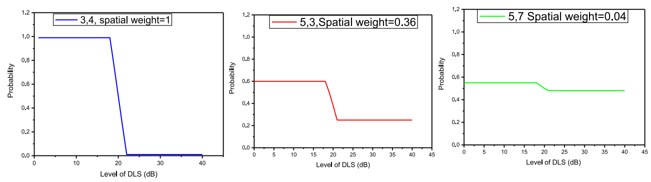

Fig. 2. Illustrative examples of scaled likelihood function negative responses r = 0, where 5 is the current test point, and 3,7 is its neighbor test points, then the varying likelihood function of neighbor test points are scaled according different spatial weight with current test points 5.

ZEST and FT, all algorithms were configured to the 24-2 HFA visual field test grid, and for each patient on each algorithm ten visual field tests were simulated. The dataset was acquired from patients attending the glaucoma clinics at Moorfields Eye Hospital NHS Foundation Trust, which functions as a district general and teaching hospital and a tertiary referral centre; VF testing and imaging was undertaken in the National Institute for Health Research (NIHR) Clinical Research Facility. Collection was undertaken in accordance with Good Clinical Practice guidelines and adhered to the Declaration of Helsinki. The trial was approved by the North of Scotland National Research Ethics Service committee on September 27, 2013 and NHS Permissions for Research was granted by the Joint Research Office at University College Hospitals NHS Foundation Trust on December 3, 2013. All patients provided written informed consent before screening investigations. More detail information about RAPID can be seen in Table I.

III. METHOD

A. Zippy Estimation of Sequential Testing

The ZEST algorithm utilizes the maximum likelihood principle and has been widely used in recent years. At the beginning of each test, an initial PDF is defined to describe the initial distribution of each location [15]. For each location, every possible threshold between 0 dB to 40 dB is quantified by this PDF. Before each stimulus is presented, a mean threshold is estimated for the current PDF and the stimulus intensity equal to the current mean threshold is presented, i.e., initial threshold estimation. Then, the PDF is adjusted according to the subject's response. Here, we use the same initial PDF as Turpin and colleagues did [27]: the initial PDF of each location should be a weighted combination of the normal and abnormal PDF of the patient at a ratio of 1:4. The normal and abnormal PDFs reveal the probability of each possible threshold for a healthy and a glaucomatous visual field, respectively (See Fig. 3). One of the initial PDFs is shown in Fig. 4a. It is evidently that 32 dB

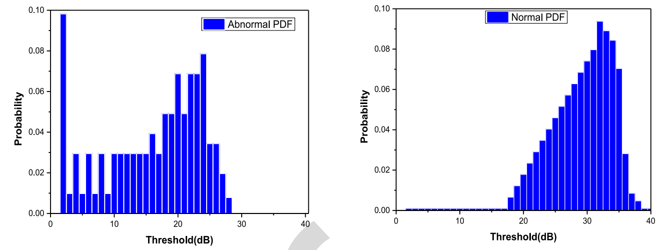

Fig. 3. Example of the initial probability density function (PDF) used in the ZEST algorithm. The left panel is the abnormal PDF, and the right panel is the normal PDF.

has the highest probability of illustrating the initial threshold for this location, then the initial stimulus of 24 dB will be presented according to the mean of the PDF. If the patient responds "yes", then the threshold will have more weight at higher decibel levels, and we multiply the current PDF by the "yes" likelihood function shown in Fig. 4b. If the patient responds "no", then the threshold will have more probability at lower decibel levels, and we multiply the current PDF by the "no" likelihood function shown in Fig. 4c. A normalization step will be carried out after each multiplication to make the sum of the probabilities equal to 1. After the normalization step, a new PDF will be obtained. The new mean is calculated, and a new stimulus contrast equal to that new mean is presented. In ZEST, there are two kinds of likelihood functions that will be used for the different responses. The likelihood used for the "yes" response assumes that the chance of seeing the stimulus at the equal level is 50%, and at much higher levels of DLS, the chance will increase to 99%, while at much lower levels of DLS, the chance will decrease to 1%. A stimulus that is 1 dB higher than the threshold will have a 75% chance of being seen, and a stimulus that is 1 dB lower than the threshold will have a 25% chance of being seen. The "yes" likelihood and "no" likelihood are symmetric. This procedure will be repeated until a certain number of rounds or the variance of the PDF becomes less than a fixed number. The final threshold is the mean of the last PDF. The test termination rule for the number of rounds was set to 10, which is the maximum measurement times for each location, or the terminating variance should be less than 1 dB [15].

B. C-ZEST Model

C-ZEST Model, a modified version of SWELZ without using growth pattern, which uses the same method with SWeLZ by incorporating spatial weight to update current and its neighbor test points concurrently while other steps are the same with ZEST, because it is easily used to discuss about the impact for different spatial filter methods and varying likelihood functions. Noted that the prior of each locations is assigned a uniform distribution so that it can avoid the influence of prior distribution, and the neighbor test points are selected according to spatial weight range from [0.1,1] that is the same with T4 method. Firstly, C-ZEST Model tests the locations in order while using the spatial weight between current and neighbor test points to scale the likelihood function of neighbor test points, and using them to update neighbor test points concurrently for each

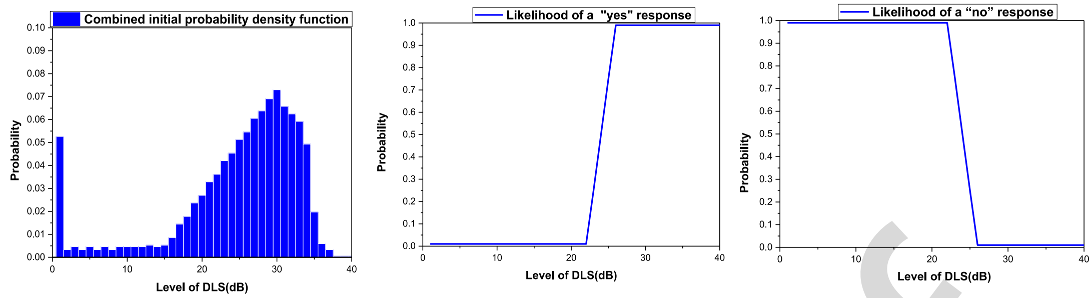

Fig. 4. (a) Combined initial PDF used for the ZEST algorithm; there is one mode in the PDF, 32 dB, which means that this value represents a good chance of being the threshold of this test location. This PDF is derived from a weighted combination of normal and abnormal thresholds. (b) The likelihood of a "yes" response, which suggests that the patient is more likely to have a higher threshold. (c) The likelihood of a "no" response, which suggests that the patient is more likely to have a lower threshold.

presentation. After that, the new PDFs of current test point and its neighbor test points are generated for the test location by multiplying the current PDF with scaled likelihood function. The likelihood function represents the probability that the observer with see the stimulus and the test terminates when the standard deviation of PDF at each location is less than 1 dB or 10 test presentation, the final threshold estimation at each location is the mean of the final PDF for that location. Here, the principle of the scaled likelihood function can be seen in Fig. 2. Suppose that 5 is the current test point of negative response, and 3,7 is its neighbor test points, then the varying likelihood function of neighbor test points are changed with different spatial weight for current test points 5. The lower spatial weight, the likelihood function become more flat (scale compressed in y-axis) but the shape (in x-axis) don't change.

IV. T4 Problem Formalization

ZEST can converge quickly and achieve better measurement accuracy if the patient's true visual function distribution is similar with the assumed initial distribution. However, it is difficult to obtain an initial distribution that approximates the true distribution of a patient, which causes a decrease in measurement accuracy and a significant increase in the number of measurements. Thus, T4 aims to construct an initial distribution of the patient's visual function threshold that can exclude as much artificial decision information as possible, hence weakening the dependence on an accurate initial distribution of the patient's visual function. Here, we assume that the patient's true visual function threshold has the same probability within the 0 to 40 dB interval. To express the belief about the parameters $\mu_m$and σ_m, prior initial distributions are imposed as two Gaussian distributions:

$$p\left(\mu_m\right) = N\left(\mu_\mu, \sigma_\mu\right) \text{and} p\left(\sigma_m\right) = N\left(\mu_\sigma, \sigma_\sigma\right) \tag{3}$$

where μ_m is the initial visual function threshold and σ_m is the variance of the visual function threshold. To make the initial distribution non-informative, similar to a uniform distribution, we usually set μ_μ= 20 dB and σ_μ= 10^3 dB. Moreover, prior parameters for σ_m are set as informative, with μ_σ= 10 dB and σ_σ= 20 dB. Noted that in our experiment μ_μ σ_σ are the same value selected from [0,40] randomly. This is aimed to make T4 have the same prior with C-ZEST and FT in our experiments. Thus, the prior of T4 has high uncertainty about the threshold before observing any response from the subject. The current Bayesian methods, such as ZEST, uses a fixed shape of the likelihood function, which cannot consider heteroscedasticity. This specification can increase the measurement times while decrease accuracy. Thus, SWeLZ uses varying likelihood function to update current and neighbor test points concurrently to decrease test times. However, it can't achieve improvement for test accuracy and stability. One of the reason is that the scaled likelihood function cannot be utilized to measure the relation between current and its neighbor test points accurately. Thus, we consider the patient's current visual function threshold and variance as independent variables in the likelihood function to express the information obtained by each measurement round. When given a stimulus of a certain intensity, the likelihood function used to correct the initial distribution is dependent on the mean of the patient's visual function threshold μ_m and the varianceσ_m. Let the visual field be divided into a set of M locations$\{x_m\}_{m=1}^M$, where x_m is a vector containing the coordinates of each location. The stimuli are presented sequentially at one individual location each time, and the responses from the subject are recorded. The ith stimulus is presented at location $x_{n_i}, n_i \in \{1, 2, \ldots, M\}$ with a sensitivity level s_i, and the response from the subject is $r_i \in \{0, 1\}$, where $r_i = 1$ indicates a positive response and $r_i = 0$ indicates no response. The probability of having a positive response $r_i = 1$ to a stimulus at level s_i at location x_m when $m = n_i$ is governed by a reverse cumulative Gaussian distribution with mean μ_m and SD σ_m:

$$p\left(r_i = 1 \left| s_i, \mu_m, \sigma_m\right.\right) = f_m\left(s\right) = \frac{1}{2}\left[1 - erf\left(\frac{s_i - \mu_m}{\sigma_m\sqrt{2}}\right)\right] \tag{4}$$

where erf(y) is the error functio. The center μ_m represents the current estimate of the threshold, and the SD σ_{m} indicates the uncertainty about this threshold. For convenience, this likelihood function is denoted by $f_m(s_i)$ for a location for which the patient has a positive response. The likelihood function of a negative response can be expressed as $1 - f_m(s_i)$. Given N stimuli $s = \{s_i\}_{i=1}^N$ and responses $r = \{r_i\}_{i=1}^N$ from the patient, the aim is to find the best fit of μ_m and σ_m to estimate the threshold and its uncertainty, respectively. μ_m and σ_m are then used to plan the next stimulus, the details of which will be described in the subsequent sections.

A. Incorporating the Spatial Weight and Prior Information About the Threshold

Conventional algorithms, ZEST, treat each location of the visual field as an independent unit during testing, with each location being measured independently. This strategy fails to take advantage of the spatial relationship between different locations of the visual field and its neighbors. SWELZ uses the spatial weight to update multiple locations concurrently, and the spatial weight derived from spatial filter methods i.e., Correlation model and geometric model [29]. Here, T4 uses a more explainable spatial filter model, combining retinotopic and optic nerve head topic spatial relationships in one metric(RONH model). Firstly, T4 assumed that the retina of each subject comprises M locations that can be denoted by $\{x_m\}, m = 1, 2, \ldots, M$. The spatial weight between two locations $x_m, m \in 1, 2, \ldots, M$ and $x_n, n \in 1, 2, \ldots, M$ can be expressed by w_{mn}. The closer the correlation value is to 1, the larger the relationship between the two points; the closer the value is to 0, the smaller the spatial weight between the two locations. Visual field locations in the different vertical hemifields are not related due to the physiological distribution of optic nerve fibers, thus the correlation is automatically set to zero [31]. On the other hand, $w_{mn} = 1$ if and only if $m = n$, i.e., locations x_m and x_n are the same, otherwise, $w_{mn} < 1$. This relationship can be represented as follows:

$$w_{mn} = \begin{cases} e^{-\frac{1}{2}\left(\frac{dist_{mn}^2}{\sigma_{\mathrm{d}}^2} + \frac{\angle_{mn}^2}{\sigma_{\angle}^2}\right)}, & \text{if } m \text{ and } n \text{ in the same hemifield} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $dist_{mn}$ is the Euclidian distance between the points x_m and x_n in the visual field, and $\angle_{mn}$ is the difference between the angles at which the optic nerve fibers crossing points p and q enter the optic nerve head, which are two factors that can better describe the spatial relationship between two locations of the visual field [34], [35]. σ_d and $\sigma_\angle$ are scale parameters. For the HFA 24-2 test grid, these parameters are chosen to be $\sigma_d = 6^\circ$ and $\sigma_\angle = 14^\circ$. Specifically, $\sigma_{\mathrm{d}} = 6^\circ$ is the angular distance between two neighboring locations, x_m and x_n, in the 24-2 visual field test pattern, and $\sigma_\angle = 14^\circ$ is the reported 95% confidence interval of the population variability in the nerve fiber entrance angle into the optic nerve head [34]. When the two points lie on different hemifields of the visual field [35] $w_{mn} = 0$. Once the formula of spatial weight between different locations is known, one can compute the spatial weight among locations, which can be seen in Fig. 5. Noted that the assumptions on the connectivity of the ONH render T4 a testing algorithm that is specific for glaucoma, because the spatial relationships following optic nerve head bundles are only true in some sense for diseases that affect the retinal nerves.

In Fig. 5 spatial weight is presented in a greyscale where black colors depict no relationship with the location in focus, and white FT represents the location itself ($w_{pq} = 1$). The brighter the color, the stronger the relationship with the location in focus.

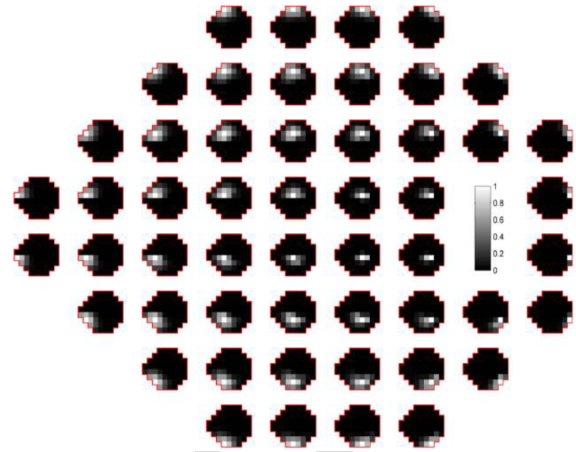

Fig. 5. Spatial weight among different locations shown on a 24-2 visual field. Each location is replaced by a smaller 24-2 visual field, which indicates the spatial weight between this location and any other location. The gray bar indicates the level of correlation.

Based on the spatial weight map, one can not only update the current posterior distribution using the proposed likelihood, but also update its neighboring locations according to computed correlation. 4 defines the probability of a positive response when $m = n_i$. However, with the definition of the spatial weight, it is desirable to borrow the stimuli and their responses from the neighboring locations when$m \neq n_i$.

For locationx_m, the likelihood of the ith responses at location x_{n_i} is defined as a binomial distribution weighted by the spatial weight w_{mn_i}:

$$p\left(r_i \left| s_i, w_{mn_i}, \mu_m, \sigma_m\right.\right) = \frac{f_m(s_i)^{w_{mn_i} r_i}(1 - f_m(s_i))^{w_{mn_i}(1-r_i)}}{f_m(s_i)^{w_{mn_i}} + (1 - f_m(s_i)^{w_{mn_i}})} \tag{6}$$

If $w_{mn_i} = 1$, i.e., when $m = n_i$, the ith stimulus is presented at x_m, the denominator becomes 1 and 6 becomes a binomial distribution defined exactly by 4. When $w_{mn_i} < 1$, i.e., the ith stimulus is not presented at x_m but is a neighboring location x_{n_i}, the distribution is "stretched" by the spatial weight w_{mn_i} and the denominator guarantees that the probability in 6 sums to 1. The impact of the spatial weight w_{mn_i} on the binomial distribution is illustrated in Fig. 6. A smaller w_{mn_i} indicates weaker spatial weight and therefore stretches the distribution to a flatter shape with larger uncertainty around the center. Therefore, when using the response from x_{n_i} at x_m, the uncertainty of the distribution increases when x_{n_i} is far away from x_m. Particularly, when $w_{mn_i} \to 0$, i.e., $x_{n_i}$is far from x_m such that their correlation approaches 0, 6 becomes a flat line at 0.5, indicating that the largest uncertainty about the response $\to \infty$. This result is intuitive because when a stimulus, x_{n_i}, is far away from x_m, it does not provide any information about the distribution of x_m. By using the spatial weight w_{mn_i}, the likelihood function of x_m is able to "borrow" information from its neighboring locations thus improving the measurement efficiency of T4 when compared with conventional threshold algorithms.

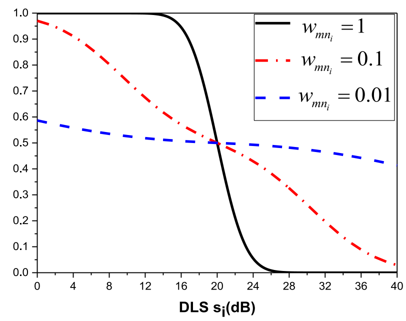

Fig. 6. Illustrative examples of weighted binomial distributions (6) with negative responses r = 0. The mean and SD of 6 were set to 20 and 2.5, respectively. $f_m(s_i)$ at $w_{mn_i} = 1$and the weighted distributions at $w_{mn_i} = 0.1$ and $w_{mn_i} = 0.01$ are plotted.

B. Inference About the Threshold and Its Uncertainty

For $\mu_m$and σ_m, the iterative formula of the posterior distribution of a patient at a certain location can be derived by multiplying 5 and 6 for all N stimuli $s = \{s_i\}_{i=1}^N$, responses $r = \{r_i\}_{i=1}^N$ and their spatial weights $w = \{w_i\}_{i=1}^N$

$$p\left(\mu_m, \sigma_m \left| r, s, w\right.\right) \propto \Pi_{i=1}^N p\left(r_i \left| s_i, w_{mn_i}, \mu_m, \sigma_m\right.\right) p\left(\mu_m\right) p\left(\sigma_m\right) \qquad (7)$$

As shown in 7, the inference about the threshold μ_m and its uncertainty σ_m is carried out by maximizing the log of 7 with the constraint that 0 dB $\leq \mu_m \leq$ 40 dB for conventional perimetry tests. The maximization was carried out using the trust-region algorithm, which is a class of iterative schemes for solving unconstrained optimization problem and have strong global convergence properties [36]. Then, the values of the estimated mean μ_m and variance σ_m are updated. Note that 7 contains the likelihood function of all the historical measurements and is a cumulative multiplication process. A likelihood function will be added to the right side of 7 after each stimulus, mainly to fully consider all the previous measurement information, including the likelihood function of the current test location and its related locations. Thus, T4 is very different from SWeLZ where only uses the spatial weight to update neighbor test points without full utilizing neighbor test points to help updating current points, that is one reason why the SWeLZ can't improve test accuracy. Here, the update rule of T4 improves more than SWeLZ only be effectiveness when using proposed likelihood function. The reason is that the Scale-likelihood function cannot be sensitive to measure the relation between current and its neighbor test points, i.e., the threshold of neighbor and current test points cannot be updated accurately by using scaled likelihood function.

C. Proposing the Next Stimulus

The T4 algorithm aims to propose the location and level of the next stimulus. It maintains a pool of candidate locations that requires further testing to confirm the threshold. This pool consists of locations where the number of stimuli presented falls below a set amount, i.e., the maximum terminate times; and those with SD σ_m larger than a set value. The next location is then selected to be the one randomly from the candidate pool.

For the simulations in this study, the candidate pool consisted of locations where the minimum amount of presentations per location was below 10 and $>= 2$ or σ_m was higher than 1 dB.

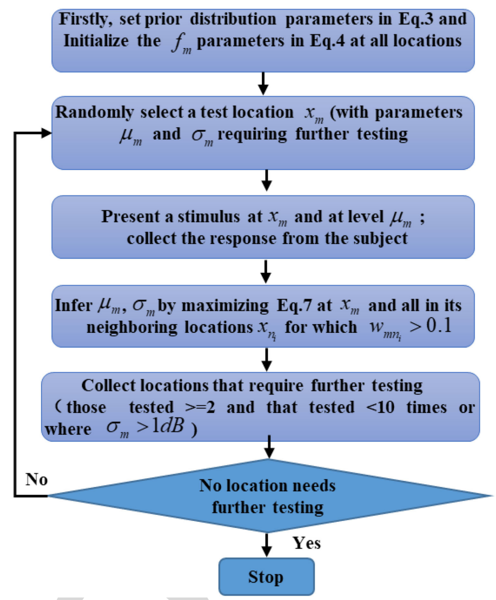

Fig. 7. Summary of the T4 procedure.

D. Putting Things Together: The Testing Procedure

The test procedure of T4 can be summarized in Fig. 7. The number of iterations of the procedure is equal to the number of stimuli presented to the subject during the test and is used as a surrogate for test duration.

Suppose that the candidate location set is C_l, we first initialize the f_m in Eq. 4 and set the prior distribution parameter in Eq. 3 for all of the location, i.e., 52 points, and adding all of the viewpoints to the candidate location set C_l. Next, randomly selecting a test location as the current test point, x_m, extracted from candidate location set, and getting the μ_m and σ_m for the current points for requiring further testing. Then, we present a stimulus at level μ_m.for the $x_m$and collect the response from the subject. After that, we get the likelihood function at x_m by using Eq, 4 after receiving the patient's response (yes or no). Meanwhile, the likelihood functions of neighbor test points corresponding to x_m are calculated by using Eq. 6 and w_{mn_i} range from [0.1,1] concurrently. Then, the μ_m and σ_m of current test point is inferred by using Eq. 7, that is, using the likelihood function both current and its neighbor test points to update current μ_m and σ_m. After that, we collect the points from C_l that locations tested $>= 2$ and $<= 10$ times or $\sigma_m < 1$ dB. When the C_l is empty the T4 is terminated and output the threshold estimation for all of 52 points. Or else, we should repeat the second step, that is, random selecting test location, x_m from C_l, and continue the next step until the $C_l$is empty. For each location, the level correspondent to the mode at the last update is taken as the threshold estimation.

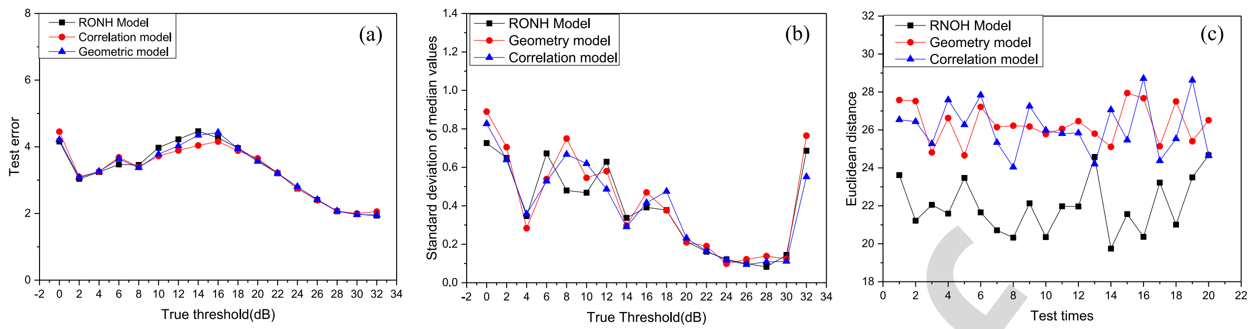

Fig. 8. Experiment of C-ZEST using different spatial filters. (a) The mean values of median test errors stratified by true sensitivities for C-ZEST with three spatial filters, RONH, Correlation and Geometric models from 20 repeated tests. (b) The SD of median test error from 20 repeated tests. (c) The Test-retest result measured by the Euclidean distance between the true and tested VF from 20 repeated tests. The C-ZEST uses the same scale likelihood and update rules with those of SWeLZ but the spatial filters are different. All the experiments are carried out with FP = 5%, FN = 5%.

V. Experiments and Results

A. The Verification of T4 Spatial Filter

In order to investigate the impact of using different spatial weight derived from different spatial filter methods. Correlation Model, Geometric model are used to make comparison with the RONH model used in T4 (Eq. 5). Here, Correlation Model was derived from a previously published spatial filter [37], and the average of two filter values was used to determine the edge weight of the edge shared between each pair of locations. Edge weights were rescaled linearly to have maximum weight of 0.55 and a minimum weight of 0. Geometric model was derived from a computational model relating retinal ganglion cells to the angle of their insertion at the optic disc [38]. C-ZEST method is used as traditional method to investigate whether the RONH model has advantage compared with other methods on improving test performance and stability. Noted that the test presentation set to 150 in verification experiments of spatial filter, varying likelihood function as well as update rules, so that making the comparison results of test accuracy, stability, as well as test-retest are reasonable. Fig. 8(a) is the mean value of median test error performance corresponding to each input threshold for the three spatial filter methods repeating 20 times. We can see that RONH model shows the similar performance with other two models in terms of mean value of median test error, and the SD of median test error for repeating 20 times (see Fig. 8(b)). However, RONH model still have improvement compared with other two model in the Test-Rest experiment (see Fig. 8(c)) range from 0-40. Thus, using a principle approach to incorporate spatial information (RONH model) can improve the test-retest performance without enlarging the test error performance evidently compared with other spatial filter methods.

B. The Verification of T4 Varying Likelihood Function

SWeLZ uses the spatial weight between current and its neighbor test points to update their threshold estimation using Scale-likelihood function. Here, we regard likelihood function of SWeLZ as Scale-likelihood function. The spatial weight can make current and its neighbor test point update concurrently by using varying likelihood function, we regard this as Borrow point. SWeLZ can decrease the test presentation compared with ZEST without decreasing the test accuracy and stability. However, it can't decrease time presentation while improving test accuracy and stability concurrently, because the scale-likelihood function is not sensitive to measure the difference between current and its neighbor test points by the likelihood function. The T4 proposes new likelihood function (See Eq. 6) that can change both the shape **(in x-axis)** and scale **compressed in y-axis** of likelihood function to update neighbor test points not like SWeLZ that just **scale compressed in y-axis but the shape (in x-axis) don't change**. Thus, it can better measure the correlation relation between the current and its neighbor test point in term of likelihood function. When updating current point, its neighbor test points can be more accurate updated concurrently.

Fig. 9(a) illustrates the mean value of median test error for 20 repeated experiments corresponding to each threshold. It is evidently that the test error improve significantly, especially for 18 to 34 dB, which prove the proposed likelihood function can be more effectiveness to borrow point's message to improve test error.

Fig. 9(b) illustrates the SD of the median test error for the experiments of repeated 20 times. We can see that the SD of using proposed likelihood function still have evident improvement compared with that of scale-likelihood function. This mainly because the likelihood function of T4 is more sensitive to measure the relation between current and its neighbor test point that can make the test points fit the optimal threshold estimation at the more correct direction compared with SWeLZ.

Fig. 9(c) illustrates the test-retest experiment for 20 times. Here, the Euclidean distance of median values are used to measure the degree of deviation between the predicted median values and diagonal line values. The improvement of test stability proves the shape and scale of likelihood function are all effective to improve the performance of borrow point performance, and can improve test error and stability concurrently.

C. The Update Rule Verification for T4

As discussed above, the varying likelihood function has big effect on improve the test error and stability compared with

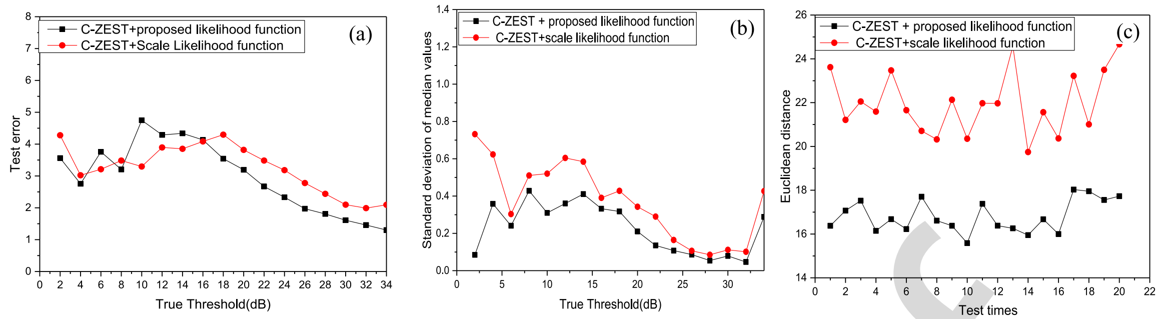

Fig. 9 Experiment of C-ZEST using different likelihood function. (a) The mean values of median test error for C-ZEST with different Likelihood functions, proposed likelihood function and Scale-likelihood function repeating 20 times. (b) The SD of median test error values repeating 20 times for C-ZEST with the two likelihood functions. (c) The Test-retest result measured by the Euclidean distance between diagonal line values and the predicted test results for repeating 20 times. Here C-ZEST uses the same spatial filter i.e., RONH mode with T4l, and the update rule is the same with SWeLZ, but the likelihood functions are different. All the experiments are at FP = 5%, FN = 5%.

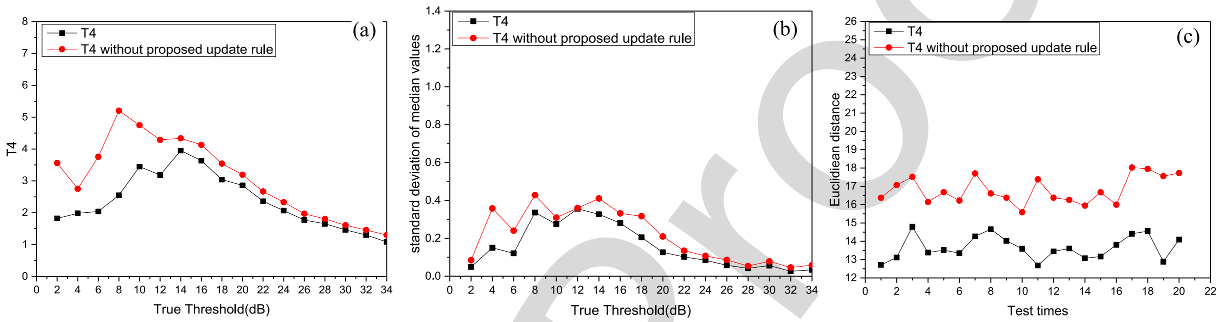

Fig. 10 (a) The Test error for T4 and T4 without update rule measured by the average median values repeating 20 times. Fig. 10(b) The SD of median test errovalues for repeating 20 times Fig. 10(c) is the Test-retest result measured by the Euclidean distance between diagonal values and the predict test result for repeating 20 times. Here C-ZEST use the same Scale likelihood and update rule with SWeLZ but the Spatial filter are different, All the experiments are at FP = 5%, FN = 5%.

Spatial filter factor. However, SWeLZ only focus on using the spatial weight of current point to update its neighbor test point without giving consideration for using the neighbor test point's message to update the current points. Thus, this update rule of SWeLZ can't fully utilize neighbor points that it has potential to improve test accuracy and stability further. As for T4, when it tests the current point, the likelihood function of neighbor test points are used to update the threshold estimate of the current point. Thus, if the current point is updated at the wrong direction resulted by inaccurate spatial weight or patient's mistake response, the other likelihood functions of its neighbor test points help it to fix the threshold estimation of current points. This can improve test error and stability performance further, prove by Fig. 10(a)–(c).

In Fig. 10(a), it shows that T4, comprises proposed update rule and likelihood function, improve the mean value of median test error compared with C-ZEST, using the same proposed likelihood function and spatial filter without T4 update rule, especially for the range from [0,26]. Thus, the proposed update rule can fully utilize neighbor test point message and can improve test error effectiveness are proved.

Fig. 10(b) illustrates the SD of median test error values repeated for 20 times corresponding to each thresholds. It is evidently that the SD of T4 improve more evidently than ZEST without proposed update rules. The main reason is that the proposed update rule can fix the test error using the likelihood function of neighbor test points, and the Posterior probability of μ_m and σ_m See Eq,7) by maximum of Eq. 7 can more better fit the optimal threshold estimate and making SD decreased.

Fig. 10(c) is the mean value of the Euclidean distance for median values to measure the Test-retest performance. We can see that the proposed update rule improves the test-retest further compared with T4 without update rules, decreasing from 17.5 to 13.5 in term of Euclidean distance. Thus, the proposed update rule can further improve the test error and test stability concurrently.

D. The Comparison Experiments

The impact of varying likelihood function, and update rule of T4 are proved to have effect on improving the test error and stability. In this section, we aim to use the T4 to compare with other general algorithms i.e., ZEST and FT. Here, ZEST uses the accurate prior that is the same initial PDF as Turpin and colleagues did [27] (see Fig. 3), aiming to get the optimal performance of ZEST. Besides, we do not use the ZEST with uniform distribution prior to make comparison, because ZWeLZ with uniform distribution have already discuss above, and ZEST show the similar performance in test accuracy and stability with SWeLZ except test presentation. Meanwhile the initial threshold of FT, similar with T4 and C-ZEST, random selecting from [0, 40] so that making comparison with T4 at the same condition,

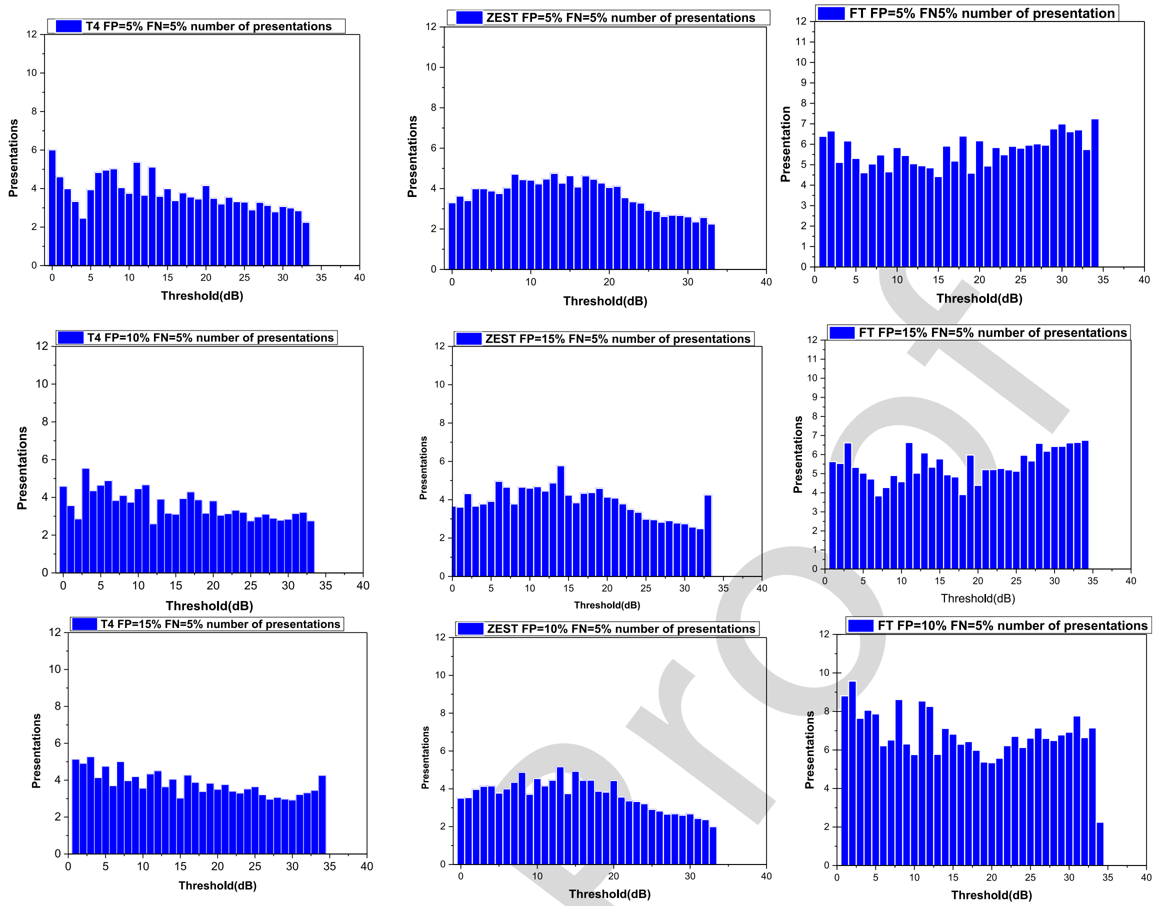

Fig. 11. Test efficiency of T4, ZEST and FT. The left panels show the test efficiency of T4, the middle panels show the test efficiency of ZEST and the right panels show the test efficiency of FT. The top three figures are the performance of the low-variability group, the middle figures are the performance of the medium-variability group, and the bottom three figures are the performance of the high-variability group. Note that the test efficiency is evaluated by the average number of presentations at each input threshold.

i.e., all the stimulus range from [0, 40] are equal probability. The performance of T4, ZEST, and FT for the low-, medium- and high-variability patient groups are illustrated in Figs. 11– 13 so that we can make comparison for the three methods at different variability measured by FP and FN.

Fig. 11 shows the number of presentations required in the testing process for all three algorithms. Fig. 12 illustrates the mean absolute difference (MAD) between the estimated threshold and the true visual fields for the three algorithms. Fig. 13 shows the Test-retest performance of T4, ZEST and FT, which indicates the variability of the difference between two repeated measurement results when testing the same subject with the same algorithm. Noted that the test error is calculated by pointwise firstly and then get the test error corresponding to all of True Threshold. Then we get SD for the Test error corresponding to each True Threshold. All the experiments were repeated 10 times, and then get the average values representing each patient's result used for comparison

1) Test Efficiency: For each algorithm, T4, ZEST and FT, we repeat the experiment for 10 times, and getting the average test presentation to evaluate test efficiency shown in Fig. 11 for each input threshold (dB) on the three variability groups. For the low-variability group, T4 has a mean number of presentations of 3.64, while ZEST and FT have mean number of presentations of 3.68 and 5.71, respectively. The medium- and high-variability groups show the same trend: T4 required 3.59, and 3.82, and ZEST requires 3.67 and 3.89 presentations for the two variability groups, while FT requires 5.49 and 6.77 respectively. Thus, T4 requires a smaller number of presentations compared with the other two algorithms at three variability level. With an increasing FP rate, T4 needs more presentations before the final threshold emerges to correct the mistake made by the patient during the testing process. While the number of presentations required for ZEST and FT does not increase presentation with FP increased. The reason is that FT uses the staircase method that the level of the next stimulus changes with a fixed and it should takes longer to recover from a patient mistake than it does on the other algorithms, i.e., more presentations. Actually it may never recover, as the 2 reversal criteria may be reached beforehand hence increasing variability. Thus, the wrong response may make the FT terminate early. ZEST only use the maximum likelihood strategy, and the variance of the PDF shrinks even if the patient response is wrong, which makes the test duration stay the same in the different patient groups. Noted that the PDF may converge into the sub-optimal that may result in decrease test accuracy, but the presentation is seldom affected.

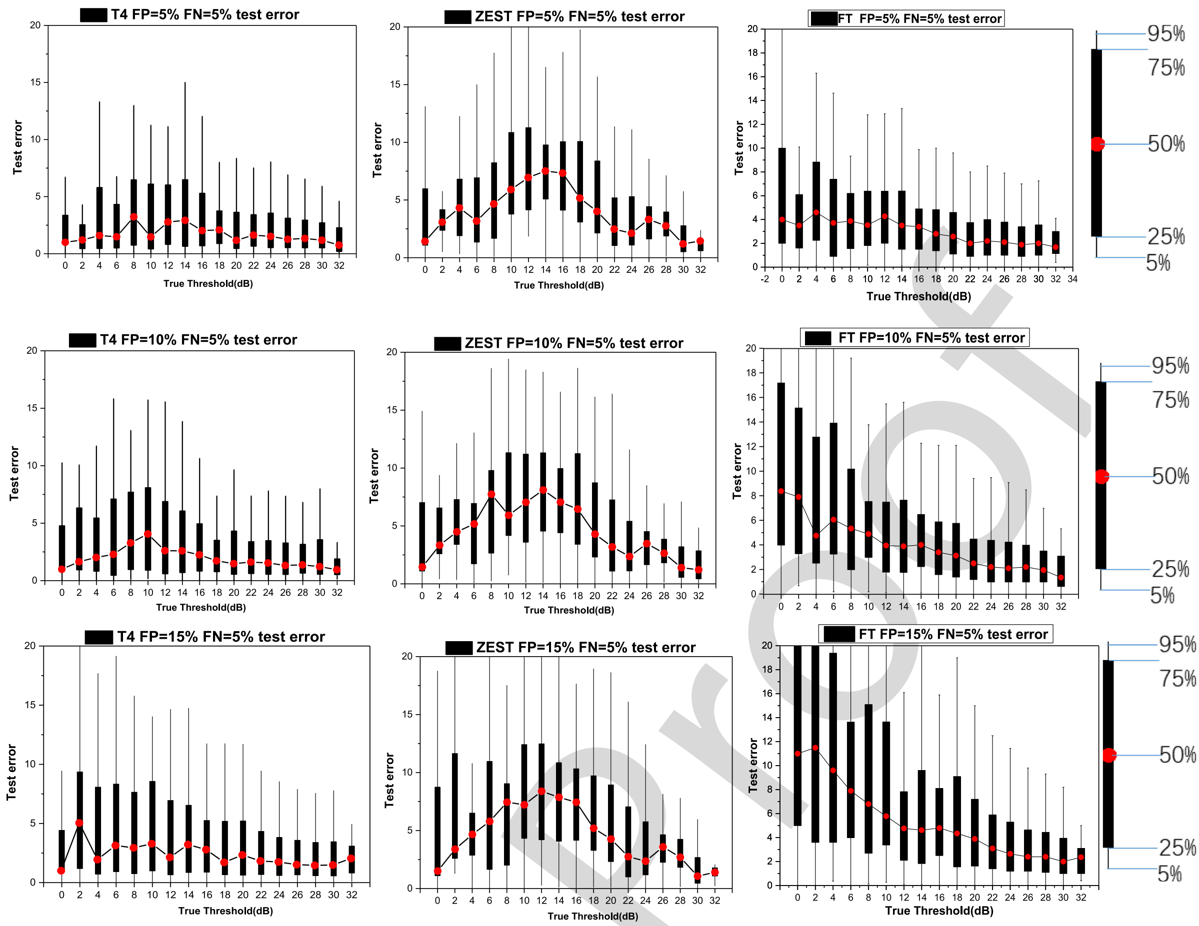

Fig. 12. MAD between estimated threshold and the true visual field for T4, ZEST and FT. The left panels show the test error of T4, the middle panels show the test error of ZEST and the right panels show the test error of FT. The top three figures are the performance in the low-variability group, the middle figures are the performance in the medium-variability group, and the bottom three figures are the performance in the high-variability group.

However, T4 updates the current test point by borrowing the message from neighboring points to help updating the current test points. Thus, with the FP increasing, the correction requires an extra number of stimuli to recover from the wrong threshold estimate and the spatial weight derived from normal dataset cannot have enough ability to update neighbor test points accurately for all of the glaucoma patients. Sometimes the spatial weight are near to the accurate spatial weight for one patient, the neighbor test points can converge to the accurate final threshold estimate quickly. When the spatial weight at disease area is not enough accurate for one patients, the neighbor test points need more presentation to fix the error. So, the SD of presentation is larger than ZEST and FT caused by the spatial weight and more sensitive to patient variability; that is, the number of presentations increases by 6–11% each time the patient variability rises. However, T4 still shows an advantage as it requires less presentations than those of the other two algorithms, i.e., T4 is faster than ZEST and FT in all the patient variability groups because the T4 can update the current and its neighboring points concurrently, which makes it has more chance to correct the wrong response compared with other methods that is the reason why the T4 have lower presentations compared with other two methods.

TABLE II
AVERAGE AND SD OF THE NUMBER OF PRESENTATIONS FOR T4, ZEST, FT FOR EACH PATIENT GROUP

Number of presentations	FP=5%, FN=5%	FP=10%, FN=5%	FP=15%, FN=5%
Average for T4	**151.06**	**155.33**	**178.22**
SD for T4	**44.43**	**42.88**	**49.38**
Average for ZEST	173.53	172.4	172.3
SD for ZEST	21.38	22.85	23.01
Average for FT	351.12	326.34	300.41
SD for FT	43.22	44.53	46.02

To more intuitively compare the number of presentation performances, we get the total presentation number of 109 subjects (52 points) firstly and then get the average value for the 109-presentation result. Then, repeat it for 10 times and get the average value for the result of 10 times. Meanwhile, the calculation steps of SD are that we first get SD for the total presentation number of 109 subjects (52 points) firstly, and then repeat it for 10 times and get the average SD for the result of 10 times. Table II show that the FT requires an average of approximately 320 presentations for the three patient groups, which is approximately twice the number required by T4 (approximately 160 presentations), and ZEST requires

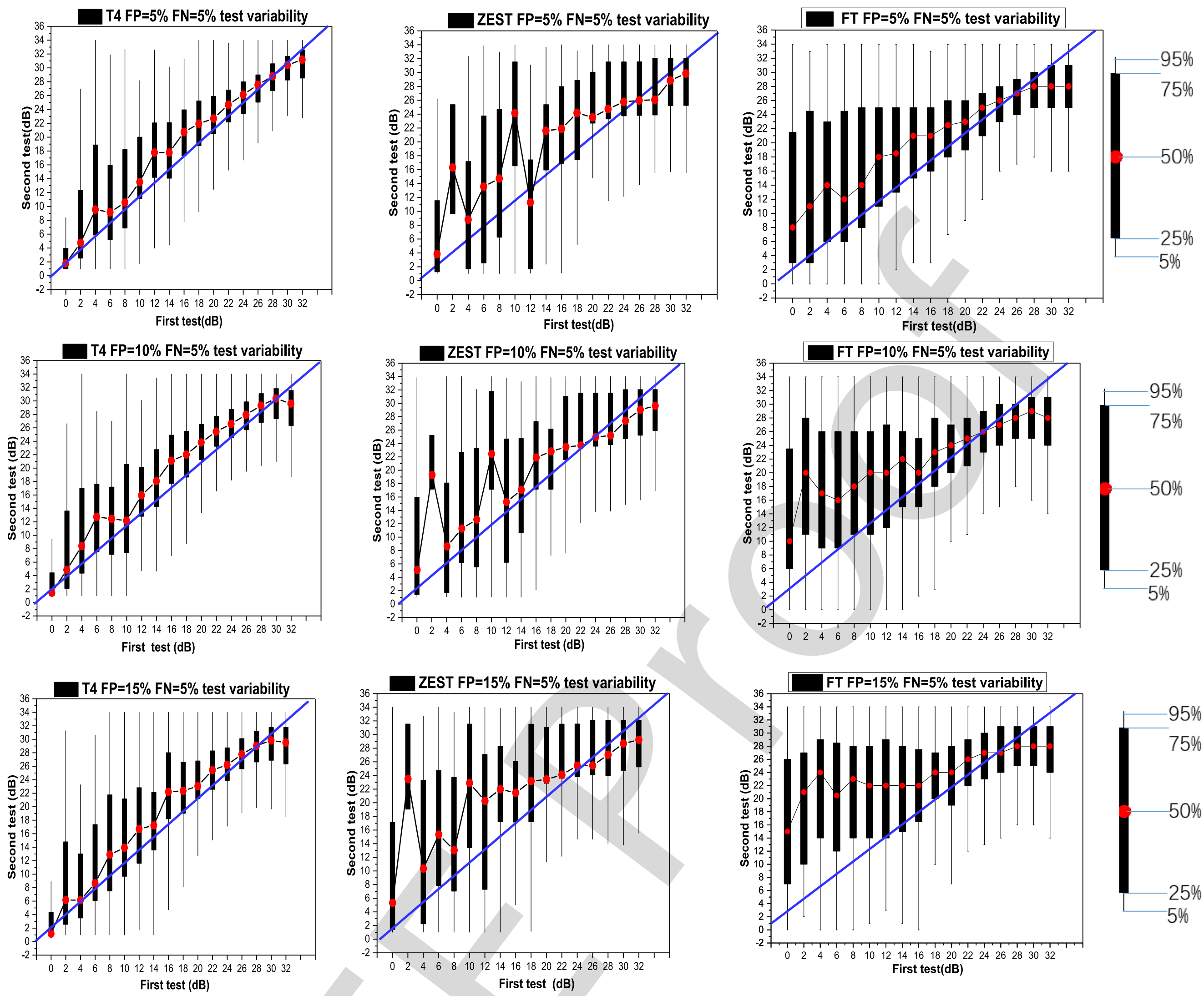

Fig. 13. Test variability of T4, ZEST and FT. The left panels show the test variability of T4, the middle panels show the test variability of ZEST and the right panels show the test variability of FT. The top three figures are the performance in the low-variability group, the middle figures are the performance in the medium-variability group, and the bottom three figures are the performance in the high-variability group. Here, baseline sensitivity represents the results from the first experiment while Retest sensitivity represent the test results for the second experiment.

approximately 173 presentations for one VF test. Thus, it is evidently that T4 can decrease the number of presentations significantly, by nearly 13 presentations, compared with ZEST. In addition, the number of presentations in T4 are sensitive to the changes in the FP, i.e., the FP increases and its SD is larger than that of other algorithms. Thus, the T4 algorithm is more sensitive for the patient's false feedback (FP variability). This makes T4 have a higher SD of presentation than the other two algorithms, but this sensitivity of T4 for incorrect patient response is essential for improving the test accuracy. The total number of test presentations of FT far exceed those of ZEST, which results from the initial threshold estimation being selected from 0 dB to 40 dB, and it is more affected by an incorrect response, making the test duration fluctuate more evidently than in ZEST in the three variability level [13].

2) Test Accuracy: Fig. 12 shows the test error performance for the three algorithms evaluated by the MAD between the estimated results and the true visual fields. The boxplots show the test error distribution for the three algorithms. Here, the test error is calculated by pointwise for 109 patients, and then it is sorted according to the true visual threshold, i.e., the real clinical visual field testing threshold result. Thus, Fig. 12 shows the test error of every true threshold for 109 patients. Noted that each patient is simulated for 10 times and then, the average threshold result is computed regarded as an average performance of one subject, which can make the result more credible (109×520 to 109×52). For the low-variability group, the mean error of T4 is 3.18 dB, while the mean error of ZEST and FT are 5.07 dB and 3.03 dB, respectively. Here, the mean error is the average value for the median sensitivity of all the true threshold (0–34 dB). With increasing FP, the mean test error for all three algorithms moderately increases; that is, the mean error of T4 in the medium-variability group is 4.02 dB while those of ZEST and FT are 5.58 dB and 4.1 dB respectively. In the high-variability group, the mean error of T4 is 4.1 dB while for ZEST and FT it is 5.93 dB and 5.29 dB. Thus, we can see that T4 shows a

TABLE III
AVERAGE DISTANCE VALUE FOR T4, ZEST, FT FOR EACH PATIENT GROUP

Average distance value	FP=5%, FN=5%	FP=10%, FN=5%	FP=15%, FN=5%
Average for T4	**13.24**	**14.58**	**16.68**
Average for ZEST	14.56	25.29	27.44
Average for FT	15.23	29.02	36.27

significant improvement in the test error compared with ZEST. FT outperforms ZEST, but FT require two time as much as ZEST in term of test presentation. Besides, T4 show the similar test error compared with FT at low and medium variability in term of median values but T4 show evident improvement in test stability compared with FT, Meanwhile T4 shows significant improvement at high variability both median values and stability, besides T4 only use half test presentation compared with FT, and SD of T4 show stable performance when FP increasing while FT increase dramatically when the FP increasing. Thus, the T4 is proved to have advantage in test error and stability compared with FT and ZEST.

3) Test Variability: Fig. 13 shows the test-retest variability performance for T4, ZEST and FT. Here, we simulated two visual fields results for 109 subjects corresponding to three variability groups in the dataset. Only data within the 95% confidence interval is shown. Meanwhile, the degree of deviation measured by summation of the Euclidean distance between the median points of the box plot and the diagonal points corresponding to (the first experiment, which can be used to measure the stability of the algorithm. The closer the median distribution of the box plot is to the diagonal points (lower Euclidean distance), the more consistent the algorithm. Noted that Fig. 13 is the example of the experiment result of three methods selected from repeated 10 times experiments. Besides, choosing different experiment as X axis or Y axis may make the median values most above or below the diagonal lines. Thus, we select the images that mostly above the diagonal lines so that make the comparison more evidently. In fact, in our experiment the median values have random above or below the diagonal line. The repeated experiment evaluation can be seen in Table III. For T4, the interval for the difference between the two tests is narrower than ZEST and FT. The variability interval (distance between the upper quartile, 75%, and the lower quartile, 25%) of ZEST and FT becomes wider than T4 for nearly all the sensitivities (dB), which suggests that the difference in the same patient between the two tests is relatively larger than that of T4. In addition, we can see that T4 has the lowest deviation between the median points and the diagonal points: its median distribution almost coincides with the diagonal line. The median distribution of FT become more offset from the diagonal, especially for lower dB.

ZEST, as a whole, have better stability compared with FT that it has better extent of coincides with the diagonal compared with FT, although there is more serious deviation at 2 dB and 10 dB, and FT show better extent of coincides with the diagonal at low variability performance. Meanwhile, ZEST show more stable with FP increasing while FT have drastic increasing. Besides, ZEST needs lower presentation than FT that is another advantage. In theory, the variability of ZEST will improve further if the number of presentations increase, but that only in simulation this will be the case. In real life fatigue will kick in which will increase test variability. Thus, the comparison of variability for T4, ZEST and FT in clinic evaluation need to be discussed in the future. As mention above, to prove the test stability for the three methods, we further repeat the experiment for 10 times and getting the average distance median values between measurement values and diagonal values to represent each test performance for three variability, which can be shown in Table II. We can see that T4 is closer to the diagonal line that it gets 13.24, 14.58, and 16.68 average distance values for three variability. Surprisingly that the Euclidean distance values of T4 do not increase significantly like ZEST and FT, which proves that the T4 has more stability. As for ZEST and FT. the test variability increase with FP increasing. But the FT illustrates more drastic increasing when FP increasing compared with ZEST. Thus, ZEST have better stability. Noted that Table III only proves ZEST with accurate prior is more stable than FT with uniform distribution prior. However, T4 still show more stable performance than that of other two methods although it uses uniform distribution prior and lower presentation.

VI. DISCUSSION

In this paper, it is shown that T4 estimates the visual field threshold more rapidly than ZEST and FT algorithms and with lower test error on the three patient groups on the computer simulation. Moreover, T4 shows a reduced heteroscedasticity compared with ZEST and FT and C-ZEST. Compared with the conventional approach ZEST, C-ZEST, and FT, the reason why T4 achieves a better performance can be concluded as follows.

Firstly, T4 uses new Likelihood function that is more sensitive with changing the spatial weight and can better measure the different between current and its neighbor test points compared with Scale-likelihood function. Here, we prove that the shape and scale are two factor to improve test accuracy and stability. Only changing the scale compressed in y-axis but the shape (in x-axis) don't change is not enough to measure the relation between current and its neighbor test points accurately that is the reason why SWeLZ can't improve test error and stability performance concurrently.

Secondly, T4 uses a novel update rule that it uses neighbor test points to help updating current test points and proposed a Bayesian method to get the threshold estimation. This can correct the patients' mistake by using the test results of its neighboring locations; nearly 20 likelihoods surround one single location ($w_{mn_i} > 0.1$). Thus, T4 is more sensitive for correcting mistake response and easier to approach accurate threshold under the helpful of neighboring points compared with the update rule of SWeLZ. Our experiments prove the effective of our proposed update rules can decrease Test error while improving test stability.

According to our experiment, varying likelihood function and update rule are the main reasons why T4 can improve test accuracy and stability. Spatial filter of T4 (RONH model) can't show evident improvement compared with Correlation model and Geometric model in terms of test accuracy and stability, but RONH shows improvement in Test-retest experiment. This is mainly because spatial filter got from normal dataset is fixed that it cannot change with different glaucoma patients. Thus, in the C-ZEST, the inaccurate spatial weight derived from spatial filter may make neighbor test points are updated at wrong direction that probably enlarging the test error and cannot improve test stability. So test accuracy and stability are tough to be improved when changing the spatial filter methods. However, combining retinotopic and optic-nerve-head-topic spatial relationships in one metric still have effect on the test-retest performance. Besides, T4 has advantage that it does not depend on the accurate prior. In real, the initial accurate threshold estimation is tough to achieve, thus, it is very meaningful to decrease the dependence on accurate threshold.

In conclusion, T4 estimates the true visual fields faster and more accurately and stability than ZEST, C-ZEST and FT robustly. Meanwhile it has significant clinical values because it is less affected by the initial estimate threshold and patient's wrong mistake response than the other current general algorithms.

References

[1] D. McFadden and F. L. Wightman, "Audition: Some relations between normal and pathological hearing," *Annu. Rev. Psychol.*, vol. 34, no. 1, pp. 95–128, Jan. 1983.

[2] B. C. Chauhan, J. D. Tompkins, R. P. LeBlanc, and T. A. McCormick, "Characteristics of frequency-of-seeing curves in normal subjects, patients with suspected glaucoma, and patients with glaucoma," *Invest. Ophthalmol. Vis. Sci.*, vol. 34, no. 13, pp. 3534–3540, Dec. 1993.

[3] S. A. Wallis, D. H. Baker, T. S. Meese, and M. A. Georgeson, "The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision," *Vis. Res.*, vol. 76, pp. 1–10, Jan. 2013.

[4] M. R. Linschoten, L. O. Harvey, P. M. Eller, and B. W. Jafek, "Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure," *Percept. Psychophys.*, vol. 63, no. 8, pp. 1330–1347, Nov. 2001.

[5] J. Sandkühler, "Models and mechanisms of hyperalgesia and allodynia," *Physiol. Rev.*, vol. 89, no. 2, pp. 707–758, Apr. 2009.

[6] C. A. Johnson, "Psychophysical factors that have been applied to clinical perimetry," *Vis. Res.*, vol. 90, pp. 25–31, Sep. 2013.

[7] J. I. Gold and L. Ding, "How mechanisms of perceptual decision-making affect the psychometric function," *Prog. Neurobiol.*, vol. 103, pp. 98–114, Apr. 2013.

[8] M. R. Leek, T. E. Hanna, and L. Marshall, "An interleaved tracking procedure to monitor unstable psychometric functions," *J. Acoust. Soc. Amer.*, vol. 90, no. 3, pp. 1385–1397, Sep. 1991.

[9] C. Hudson, J. M. Wild, and E. C. O'Neill, "Fatigue effects during a single session of automated static threshold perimetry," *Invest. Ophthalmol. Vis. Sci.*, vol. 35, no. 1, pp. 268–280, Jan. 1994.

[10] H. Zhu, R. A. Russell, L. J. Saunders, S. Ceccon, D. F. Garway-Heath, and D. P. Crabb, "Detecting changes in retinal function: Analysis with non-stationary weibull error regression and spatial enhancement (ANSWERS)," *PLoS One*, vol. 9, no. 1, Jan. 2014, Art. no. e85654.

[11] U. Schiefer *et al.*, "Comparison of the new perimetric GATE strategy with conventional full-threshold and SITA standard strategies," *Invest. Opthalmol. Vis. Sci.*, vol. 50, no. 1, p. 488, Jan. 2009.

[12] M. Wall, K. R. Woodward, and C. F. Brito, "The effect of attention on conventional automated perimetry and luminance size threshold perimetry," *Invest. Opthalmol. Vis. Sci.*, vol. 45, no. 1, pp. 342–350, Jan. 2004.

[13] J. L. Keltner and C. A. Johnson, "Effectiveness of automated perimetry in following glaucomatous visual field progression," *Ophthalmology*, vol. 89, no. 3, pp. 247–254, Mar. 1982.

[14] J. G. Flanagan *et al.*, "Evaluation of FASTPAC: A new strategy for threshold estimation with the humphrey field analyser," *Graefe's Arch. Clin. Exp. Ophthalmol.*, vol. 231, no. 8, pp. 465–469, Aug. 1993.

[15] A. Turpin, A. M. McKendrick, C. A. Johnson, and A. J. Vingrys, "Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer simulation," *Invest. Opthalmol. Vis. Sci.*, vol. 44, no. 11, pp. 4787–4795, Nov. 2003.

[16] A. Heijl, A. Lindgren, and G. Lindgren, "Test-retest variability in glaucomatous visual fields," *Amer. J. Ophthalmol.*, vol. 108, no. 2, pp. 130–135, Aug. 1989.

[17] C. A. Johnson, B. C. Chauhan, and L. R. Shapiro, "Properties of staircase procedures for estimating thresholds in automated perimetry," *Invest. Ophthalmol. Vis. Sci.*, vol. 33, no. 10, pp. 2966–2974, Sep. 1992.

[18] S. E. Spenceley and D. B. Henson, "Visual field test simulation and error in threshold estimation," *Brit. J. Ophthalmol.*, vol. 80, no. 4, pp. 304–308, Apr. 1996.

[19] J. M. Wild, I. E. Pacey, E. C. O'Neill, and I. A. Cunliffe, "The SITA perimetric threshold algorithms in glaucoma," *Invest. Ophthalmol. Vis. Sci.*, vol. 40, no. 9, pp. 1998–2009, Aug. 1999.

[20] A. B. Watson and D. G. Pelli, "Quest: A bayesian adaptive psychometric method," *Percept. Psychophys.*, vol. 33, no. 2, pp. 113–120, Mar. 1983.

[21] A. M. McKendrick and A. Turpin, "Advantages of terminating zippy estimation by sequential testing (ZEST) with dynamic criteria for white-on-white perimetry," *Optometry Vis. Sci.*, vol. 82, no. 11, pp. 981–987, Nov. 2005.

[22] B. Bengtsson, J. Olsson, A. Heijl, and H. Rootzen, "A new generation of algorithms for computerized threshold perimetry, SITA," *Acta Ophthalmol. Scand.*, vol. 75, no. 4, pp. 368–375, Aug 1997.

[23] B. Bengtsson, A. Heijl, and J. Olsson, "Evaluation of a new threshold visual field strategy, SITA, in normal subjects," *Acta Ophthalmol. Scand.*, vol. 76, no. 2, pp. 165–169, Apr. 1998.

[24] B. Bengtsson and A. Heijl, "Evaluation of a new perimetric threshold strategy, SITA, in patients with manifest and suspect glaucoma," *Acta Ophthalmol. Scand.*, vol. 76, no. 3, pp. 268–272, Jun. 1998.

[25] P. E. King-Smith, S. S. Grigsby, A. J. Vingrys, S. C. Benes, and A. Supowit, "Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation," *Vis. Res.*, vol. 34, no. 7, pp. 885–912, Apr. 1994.

[26] A. J. Vingrys and M. J. Pianta, "A new look at threshold estimation algorithms for automated static perimetry," *Optometry Vis. Sci.*, vol. 76, no. 8, pp. 588–595, Aug. 1999.

[27] A. Turpin, A. M. McKendrick, C. A. Johnson, and A. J. Vingrys, "Performance of efficient test procedures for frequency-doubling technology perimetry in normal and glaucomatous eyes," *Invest. Ophthalmol. Vis. Sci.*, vol. 43, no. 3, pp. 709–715, Mar. 2002.

[28] A. Turpin, A. M. McKendrick, C. A. Johnson, and A. J. Vingrys, "Development of efficient threshold strategies for frequency doubling technology perimetry using computer simulation," *Invest. Ophthalmol. Vis. Sci.*, vol. 43, no. 2, pp. 322–331, Feb. 2002.

[29] "Incorporating spatial models in visual field test Procedures[J]," *Transl. Vis. Sci. Technol.*, 2016. Q2

[30] S. Kucur Erife, S. Raphael, and A. Andrew. "Sequentially optimized reconstruction strategy: A meta-strategy for perimetry testing," *PLoS One*, vol. 12, no. 10, 2017, Art. no. e0185049.

[31] G. Montesano *et al.* "Improving visual field examination of the macula using structural Information[J]," *Transl. Vis. Sci. Technol.*, vol. 7, no. 6, 2018.

[32] B. Chesley and D. L. Barbour, "Visual field estimation by probabilistic classification," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3499–3506, Dec. 2020.

[33] D. B. Henson, S. Chaudry, P. H. Artes, E. B. Faragher, and A. Ansons, "Response variability in the visual field: Comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes," *Invest. Ophthalmol. Vis. Sci.*, vol. 41, no. 2, pp. 417–421, Feb. 2000.

[34] D. F. Garway-Heath, D. Poinoosawmy, F. W. Fitzke, and R. A. Hitchings, "Mapping the visual field to the optic disc in normal tension glaucoma eyes," *Ophthalmology*, vol. 107, no. 10, pp. 1809–1815, Oct. 2000.

[35] N. G. Strouthidis, V. Vinciotti, A. J. Tucker, S. K. Gardiner, D. P. Crabb, and D. F. Garway-Heath, "Structure and function in glaucoma: The relationship between a functional visual field map and an anatomic retinal map," *Invest. Opthalmol. Vis. Sci.*, vol. 47, no. 12, pp. 5356–5362, Dec. 2006.

[36] A. Kamandi, K. Amini, and M. Ahookhosh, "An improved adaptive trust-region algorithm[J]," *Optim. Lett.*, vol. 11, no. 3, pp. 555–569, 2017.

[37] S. K. Gardiner *et al.*, "Reducing noise in suspected glaucomatous visual fields by using a new spatial filter.[J]," *Vis. Res.*, vol. 44, no. 8, pp. 839–848, 2004.

[38] D. Jonathan, A. M. Mckendrick, and T. Andrew, "An anatomically customizable computational model relating the visual field to the optic nerve head in individual Eyes[J]," *Invest. Ophthalmol. Vis.*, vol. 53, no. 11, pp. 6981–6990, 2012.