A Computer Assisted Method for Nuclear Cataract Grading From Slit-Lamp Images Using Ranking

Wei Huang*, Kap Luk Chan, *Member, IEEE*, Huiqi Li, *Senior Member, IEEE*, Joo Hwee Lim, *Member, IEEE*, Jiang Liu, and Tien Yin Wong

***Abstract*—In clinical diagnosis, a grade indicating the severity of nuclear cataract is often manually assigned by a trained ophthalmologist to a patient after comparing the lens' opacity severity in his/her slit-lamp images with a set of standard photos. This grading scheme is often subjective and time-consuming. In this paper, a novel computer-aided diagnosis method via ranking is proposed to facilitate nuclear cataract grading following conventional clinical decision-making process. The grade of nuclear cataract in a slit-lamp image is predicted using its neighboring labeled images in a ranked image list, which is achieved using a learned ranking function. This ranking function is learned via direct optimization on a newly proposed approximation to a ranking evaluation measure. Our proposed method has been evaluated by a large dataset composed of 1000 different cases, which are collected from an ongoing clinical population-based study. Both experimental results and comparison with several existing methods demonstrate the benefit of grading via ranking by our proposed method.**

***Index Terms*—Computer-aided diagnosis (CAD), grade, nuclear cataract, ranking, slit-lamp images.**

I. Introduction

CATARACT, the "clouding" or opacity developed in the crystalline lens of human eyes, obstructs the passage of light and is the leading cause of vision loss globally [1], [2]. In a world health report published in 1998, about 43% of global blindness was caused by cataracts [1]. This number had increased to 47.8%, which represents about 18 million people, by 2002 [2]. Among various causes of cataracts, aging is the most common one. It is due to the fact that proteins within the lens of aged population are prone to bind (a process known as cross-linking) and become stiffer to form cloudy spots (cataracts) [3]. In the U.S., more than 50% of all Americans at the age of 65 or above suffer from age-related cataracts. This number increases to 70% among those over 75 years old [4]. In Singapore, about 35% of Singapore Chinese population over the age of 40 have age-related cataracts [5].

Based on the locations of developed opacity, age-related cataracts are categorized into three types: posterior subcapsular cataract, cortical cataract, and nuclear cataract [6]. Among them, posterior subcapsular cataract forms at the back of the lens; while cortical cataract forms in the lens cortex and extends its spokes from the outside of the lens to the center [6]. Nuclear cataract, which forms in the nucleus, is the most common type [2], [6]. Accurate diagnosis and timely treatment of nuclear cataract is essential to prevent vision loss. Clinical diagnosis of nuclear cataract is often conducted with the help of slit-lamp photography, from which slit-lamp images are produced depicting eye conditions of patients for ophthalmologists to diagnose their cataract disease [7], [8].

In clinical diagnosis, a grade of nuclear cataract is often manually assigned by trained ophthalmologists to each slit-lamp image by comparing its opacity severity with a set of standard images [7], [8]. To measure the opacity severity quantitatively, several grading systems have been established [9], [10]–[12]. For instance, Fig. 1 shows a set of four standard slit-lamp images used in the Wisconsin Cataract Grading System [9]. These images together represent an increasing severity of cataract indicated by increasing integer-valued grades (from 1 to 4). In clinical grading, an ungraded slit-lamp image is compared with these standard images and a/an decimal/integer-valued grade is assigned to indicate its opacity severity. Hence, if a slit-lamp image is assigned a grade 2.5 by ophthalmologists, it means that the patient has a nuclear cataract disease that is not as severe as the one depicted by the standard image of grade 3, but more severe than the one of grade 2. Although this manual grading scheme is utilized in clinical practice, it is often argued to be subjective. In [9], it is reported that only around 65% inter-observer agreement can be reached when different ophthalmologists are told to assign grades to the same slit-lamp images following the same grading system. Furthermore, ophthalmologists are likely to suffer from fatigue after inspecting numerous images and prone to unconsciously grade them imprecisely.

Therefore, nowadays, automatic, objective and quantitative diagnosis of nuclear cataract in slit-lamp images becomes necessary, and it has been investigated by several research groups. The Wisconsin group [13], [14] extracted anatomical structure on the visual axis of the lens. Sulcus intensity and intensity ratio between anterior lentil and posterior lentil were selected

Manuscript received April 29, 2010; revised July 10, 2010; accepted July 15, 2010. Date of publication July 29, 2010; date of current version December 30, 2010. This work was supported in part by the National Medical Research Council (NMRC), NMRC/STaR/0003/2008 and in part by the Singapore Bio Imaging Consortium (SBIC) under Grant C-011/2006. *Asterisk indicates corresponding author.*

*W. Huang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: n060101@ntu.edu.sg).

K. L. Chan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: eklchan@ntu.edu.sg).

H. Li, J.H. Lim, and J. Liu are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore (e-mail: huiqili@i2r.a-star.edu.sg; joohwee@i2r.a-star.edu.sg; jliu@i2r.a-star.edu.sg).

T.Y. Wong is with the National University of Singapore, Singapore National Eye Center and Singapore Eye Research Institute, Singapore (e-mail: ophwty@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2010.2062197

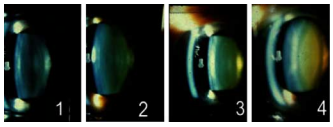

Fig. 1. Standard images with grades indicating the severity of nuclear cataract disease within slit-lamp images according to Wisconsin Cataract Grading System.

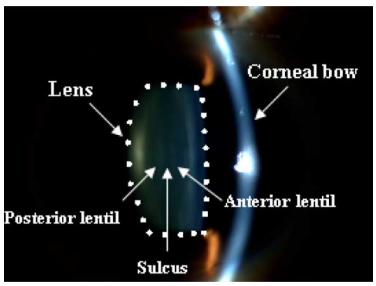

Fig. 2. An illustration of lens structure.

as features using linear regression (Fig. 2). The nuclear cataract grading task was considered as a classification problem [14]. The Johns Hopkins group also regarded the nuclear cataract grading task as a classification problem. They analyzed the intensity profile on the visual axis and three features were extracted: nuclear mean gray level, slope at the posterior point of profile, and the fractional residual of the least-square fit [15]. Hence, both of these studies considered the nuclear cataract grading task as a classification problem, but they only utilized features on the visual axis, whereas the whole area of the lens, which is usually analyzed in the clinical diagnosis, is overlooked. In [16] and [17], the lens contour was automatically detected and features were selected from the segmented lens area according to the Wisconsin cataract grading protocol. This study considered the nuclear cataract grading task as a regression problem and support vector regression was adopted to predict grades. Generally speaking, existing studies on automatic grading of nuclear cataract within slit-lamp images are scarce, and most existing studies consider the grading task as either a classification or regression problem.

In this study, we, computer scientists and clinicians working closely together, propose a novel computer-aided diagnosis (CAD) method, which takes the grading of nuclear cataract disease within slit-lamp images as a ranking task. Generally speaking, ranking aims to sort a list of items according to a system of rating or a record of performance, and the idea of ranking has been adopted in several research works on eye images. In [18], a ranking of components of several color spaces is provided to indicate that, the green component in RGB color space is preferred for blood vessel detection in fundus images. In [19], Spearman's rank correlation coefficient, a nonparametric rank correlation measure, is used to reveal the relationship between the degree of retinal contraction and the degree of metamorphopsia in retinal images. In [20], several red-free fundus images are ranked according to their abilities to discern the margin of internal limiting membrane peeling to evaluate a proposed spectral imaging technique. Although the idea of ranking has been incorporated in several research efforts on eye images, none of them is related to the grading of nuclear cataract disease within slit-lamp images.

The intuition to formulate nuclear cataract grading via ranking is explained as follows. From the classification perspective, grading of nuclear cataract can be conducted by classifying ungraded slit-lamp images to specific classes, and their grades can be suggested by assigned classes therein [13]–[15]. From the regression perspective, grades of nuclear cataract can be directly predicted as outputs of a regression procedure [16], [17]. For these methods, slit-lamp images with clinicians' grades are often used in their training process (i.e., tuning parameters in classifiers or regression procedures), but these images are often not used explicitly in the subsequent and most important phase—the grading process. In the conventional clinical decision-making process of nuclear cataract diagnosis, ophthalmologists usually utilize slit-lamp images manually graded at previous clinical appointments in their current appointment (a popular clinical decision support technique known as "case-based reasoning," which handles new cases based on clinical results of previous cases). Hence, it inspires us to propose a new nuclear cataract grading scheme, which can explicitly incorporate previously graded cases in the interpolation of new ungraded cases following the conventional clinical decision-making process. In this study, we consider the nuclear cataract grading task as a ranking process following the intuition, and ranking can provide a better fit to the task.

The flowchart of our "grading via ranking" scheme is illustrated in Fig. 3. The main idea is to sort slit-lamp images into a ranked image list according to the degree of severity of the nuclear cataract disease within all images. The grade of a new slit-lamp image can be interpolated using its neighboring images with clinicians' grades in the ranked list. In order to achieve such a ranked slit-lamp image list, we incorporated the "learning to rank" technique [21], which aims to learn a ranking function, so that slit-lamp images can be sorted into a ranked list via the learned ranking function. Generally speaking, there are many "learning to rank" methods in literature, and many of them often employ existing ranking evaluation measures (e.g., normalized discounted cumulative gain (NDCG), mean average precision (MAP) [22], [23]) to learn ranking functions with diverse indirect optimization techniques [24]–[26]. The reason to conduct indirect optimization, instead of adopting more intuitive direct optimization techniques, is because existing ranking evaluation measures are often neither continuous nor differentiable so that direct optimization is usually infeasible to apply [22], [23].

In this study, we propose a new "learning to rank" method, which incorporates a newly proposed approximation to a ranking evaluation measure for learning ranking functions via direct optimization, within our "grading via ranking" scheme. The contribution of this study lies in two aspects. 1) Unlike the conventional way to grade nuclear cataract disease within slit-lamp images by considering it as either a classification or regression task, our study is the first attempt to regard nuclear cataract grading as a ranking task. 2) Technically, a new "learning to rank" method is proposed with a new approximation to a ranking evaluation measure, which can be directly optimized for learning ranking functions. The organization of the paper is as follows. In Section II, a brief review of "learning

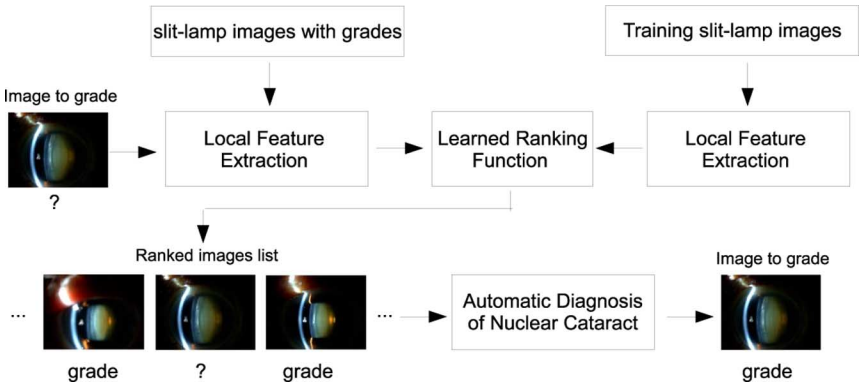

Fig. 3. Flowchart of our nuclear cataract "grading via ranking" scheme.

to rank" methodology is given. Several conventional ranking evaluation measures are presented in Section II-A. Various "learning to rank" methods are introduced and categorized into three groups as well as their pros and cons discussed in Section II-B. Section III presents our proposed "grading via ranking" scheme. Section III-A and Section III-B cover our newly proposed "learning to rank" method. First, a new approximation to a ranking evaluation measure is introduced in Section III-A. A corresponding learning algorithm incorporating direct optimization on the newly proposed approximation for learning ranking functions is elaborated in Section III-B. When ranked slit-lamp image lists are obtained, the nuclear cataract grading strategy is introduced in Section III-C. In Section IV, a large dataset composed of 1000 slit-lamp images from 1000 cases with different nuclear cataract disease obtained from an ongoing population-based study is used to evaluate our "grading via ranking" scheme. Dozens of experiments are conducted to evaluate the ranking performance of our newly proposed "learning to rank" method with comparison to several popular "learning to rank" methods. Then, our "grading via ranking" scheme incorporating these "learning to rank" methods is compared with two existing nuclear cataract grading schemes to evaluate the nuclear cataract grading performance. Experimental results are analyzed from a statistical point of view as well. In Section V, the conclusion of this study is drawn.

II. Review of Learning to Rank Approach

"Learning to rank" is an emerging approach in machine learning and information retrieval in recent years [21], [23]. Generally speaking, "learning to rank" is made up of two steps in a sequence: learning and ranking.

Learning A set of m image lists $d^{(j)} = \left\{d_1^{(j)}, d_2^{(j)}, \ldots, d_{m^{(j)}}^{(j)}\right\}$ are provided with their corresponding relevance $r^{(j)} = \left\{r_1^{(j)}, r_2^{(j)}, \ldots, r_{m^{(j)}}^{(j)}\right\}$, $j = 1, \ldots, m; m^{(j)}$ denotes the number of images within the list $d^{(j)}$. A ranking function f is learned from these training data. Generally speaking, ranking function f is defined in terms of each individual image: $f(d_i^{(j)})$, $i = 1, \ldots, m^{(j)}$ with its output as the score of each image (a real number). The learned ranking function will be used to sort the image collection in the ranking step.

Ranking For a list of n images $d = \{d_1, d_2, \ldots, d_n\}$, the purpose of ranking is to sort images within the list in a/an descending/ascending order of relevance measured by the score of each image calculated from the learned ranking function f.

A. Conventional Ranking Evaluation Measures

There are several measures proposed to evaluate the ranking performance, including winners take all (WTA), mean reciprocal rank (MRR), mean average precision (MAP), and normalized discounted cumulative gain (NDCG) [22], [23]. WTA is a simple measure which only takes the top-most image in the ranked image list into consideration. If the top-most one is relevant, $\mathrm{WTA} = 1$; otherwise, $\mathrm{WTA} = 0$. MRR utilizes the reciprocal of the first relevant image's position as its value. For MAP and NDCG, they are more complex but popular than WTA and MRR.

MAP is a ranking evaluation measure dedicated for the case of binary relevance judgment, in which associated images are assumed to be either relevant or irrelevant [23]. In a ranked image list, precision at position $n (p@n)$ is calculated to measure the ranking performance of top n images

$$p@n = \frac{\text{No. of relevant images (among top n images of the list)}}{n}. \tag{1}$$

Then, average precision (AP) of the list is calculated as the average of all precisions at different positions where the images are relevant. The definition is given in (2). Finally, MAP is calculated as the mean of all AP among all lists [23]

$$AP = \frac{\sum_n p@n \times r_n}{\text{No. of all relevant images}}$$
$$r_n = \begin{cases} 1, & \text{image n is relevant} \\ 0, & \text{image n is irrelevant} \end{cases}. \tag{2}$$

NDCG, on the contrary, is suitable for the case of multiple-level relevance judgement [22]. Its definition is as follows:

$$\mathrm{NDCG} = N_M^{-1} \times \mathrm{DCG} = N_M^{-1} \sum_{n=1}^{M} g(r_n)$$
$$d(n) = N_M^{-1} \sum_{n=1}^{M} \frac{2^{r_n} - 1}{\log_2(1+n)} \tag{3}$$

where N_M is a normalization term denoting the maximum of discounted cumulative gain (DCG), n represents positions in a ranked image list composed of M images, r_n is the degree of relevance of image located at position n, $g(r_n)$ is a gain function and $d(n)$ is a discount function represented by a monotonically increasing exponential function $g(r_n) = 2^{r_n} - 1$ and a monotonically decreasing logarithmic reduction factor $d(n) = 1/\log_2(1+n)$, respectively, in the original definition of NDCG [22]. The range of NDCG is within $[0, 1]$, and higher values indicate better ranking performance.

B. "Learning to Rank" Methods: Categories, Pros and Cons

For most "learning to rank" methods, the basic assumption is that, if a ranking function can be learned by optimizing its ranking performance in terms of a given ranking evaluation measure on the training data, high ranking accuracy is expected from the same measure when the learned ranking function is used to rank other data. In general, most existing "learning to rank" methods can be categorized into three types of approaches: pointwise approach, pairwise approach, and listwise approach.

1) Pointwise Approach: For an image list $d^{(i)} = \{d_1^{(i)}, d_2^{(i)}, \ldots, d_{n^{(i)}}^{(i)}\}$, pointwise approach aims to assign each image a discrete category: $\{(d_1^{(i)}, c_1^{(i)}), (d_2^{(i)}, c_2^{(i)}), \ldots, (d_{n^{(i)}}^{(i)}, c_{n^{(i)}}^{(i)})\}$, in which $\{c_1^{(i)}, c_2^{(i)}, \ldots, c_{n^{(i)}}^{(i)}\} \in C$; $C = \{c_1 \succ c_2 \succ \cdots c_m\}$ is a set of m ordered categories, where $\succ$ denotes the relevance order between various categories. Since elements in C are ordered discrete values, pointwise approach is also known as ordinal regression, which is between regression (outputs: real values that can be ordered) and classification (outputs: nonordered discrete values) [27], [28]. Representative pointwise approaches include constrained ordinal regression [28], Pranking [29], OAP-BPM [30], ranking with large margin principals [31], etc. Although pointwise approach is convenient to implement due to its close resemblance to regression and classification, its drawback is obvious: it can only deal with relevance judgement in the form of absolute relevance. Nonabsolute preference, such as pairwise preference and partial (total) list orders, cannot be handled by pointwise approach.

2) Pairwise Approach: Unlike pointwise approach, which takes separated images as instances, pairwise approach focuses on image "pairs" instead. In learning, image pairs $\left(d_\alpha^{(i)}, d_\beta^{(i)}\right)$ $\left(\alpha, \beta \in \{1, 2, \ldots, n^{(i)}\}, \alpha \neq \beta\right)$ are collected from an image list $d^{(i)} = \{d_1^{(i)}, d_2^{(i)}, \ldots, d_{n^{(i)}}^{(i)}\}$. For each pair, a label $r \in \{+1, -1\}$ is assigned indicating the relative relevance of two images $\left(d_\alpha^{(i)}, d_\beta^{(i)}\right)$: $+1$ shows $d_\alpha^{(i)}$ is more relevant than $d_\beta^{(i)}$, and -1 to the contrary. The idea of pairwise approach is similar to binary-class classification, and pairwise approaches often formulate the ranking task as a classification problem accordingly [24], [26], [32]. Conventional classification methods, such as boosting, support vector machine (SVM), and neural network, have been incorporated leading to corresponding pairwise methods, such as RankBoost [24], RankingSVM [26], and RankNet [32], respectively.

There are several advantages with the pairwise approach. First, existing classification methodologies can be conveniently adopted [24], [26], [32]. Second, pairwise preference, rather than absolute relevance, is relatively easy to obtain for some scenarios [26]. Nevertheless, there are also drawbacks inherently. First, the ranking procedure is often considered as a classification process, the problem of learning ranking functions is accomplished by minimizing classification errors therein. However, the purpose of ranking is not the same as that of classification [24], [26]. Second, pairs are usually assumed to be generated i.i.d (independent identically distributed). This assumption is sometimes too strong to meet for some applications [32]. Third, the number of generated pairs may vary largely from lists to lists. It is likely to result in learning ranking functions biased towards lists with more data pairs [33].

3) Listwise Approach: Recently, more and more researchers focus on another type of "learning to rank" method—listwise approach [25], [34], [35], [36]. In learning, a set of m image lists $d^{(i)} = \{d_1^{(i)}, d_2^{(i)}, \ldots, d_{n^{(i)}}^{(i)}\}$ and their corresponding relevance $r^{(i)} = \{r_1^{(i)}, r_2^{(i)}, \ldots, r_{n^{(i)}}^{(i)}\}$ are given $(i = 1, \ldots, m)$. Listwise approach aims to learn a ranking function f to output a score for each image. A ranked image list is achieved in a/an decreasing/increasing order of these scores $\{f(d_1^{(i)}), f(d_2^{(i)}), \ldots, f(d_{n^{(i)}}^{(i)})\}$. The learning of a ranking function is often carried out by minimizing a loss function in terms of the difference between the generated ranked image list and its ground truth. Representative listwise methods include AdaRank [25], ListNet [34], RankCosine [35], FRank [36], etc.

For many listwise approaches, their loss functions are defined based on conventional ranking evaluation measures (e.g., NDCG, MAP, etc.). These position-based measures are often neither differentiable nor continuous in terms of discrete positions in ranked lists. Hence, direct optimization is usually infeasible to implement for learning ranking functions, and various indirect optimization techniques are utilized as alternatives [25], [34]–[36]. In Section III, we introduce a new "learning to rank" method based on the listwise approach, incorporating direct optimization for learning ranking functions within our "grading via ranking" scheme.

III. Methodology

In this section, our newly proposed nuclear cataract "grading via ranking" scheme is presented. First, a new "learning to rank" method is introduced. A new approximation to a ranking evaluation measure is proposed in Section III-A. The approximation is utilized to learn ranking functions via a direct optimization algorithm elaborated in Section III-B. Once ranked slit-lamp image lists are obtained by learned ranking functions, the grade of nuclear cataract in each slit-lamp image is interpolated with the help of its neighboring images in the ranked image list. The grading strategy is introduced in Section III-C.

A. "Learning to Rank" Part I-A New Approximation to NDCG: S-NDCG

As mentioned in Section II-A, NDCG is a conventional position-based ranking evaluation measure, which can handle multiple-level relevance judgement. It is chosen here because nuclear cataract grades annotated by ophthalmologists in this study

are also in the form of multiple values. From the original NDCG definition in [22], we have the NDCG for our application as given below

$$\mathrm{NDCG} = N_M^{-1} \times \mathrm{DCG} = N_M^{-1} \sum_{x \in \chi} \frac{2^{r(x)} - 1}{\log_2(1 + \pi(x))} \quad (4)$$

where x is a slit-lamp image and χ is the set of images to be ranked, $r(x)$ and $\pi(x)$ are annotated grade of nuclear cataract disease within image x and position of image x in the ranked image list, respectively, N_M is a normalization term denoting the maximum of DCG as before, which can be obtained when all images are sorted in a perfect order of decreasing severity of nuclear cataract disease.

Unfortunately, optimization cannot be directly applied on NDCG for learning ranking functions, since the measure itself is neither continuous nor differentiable in terms of discrete position $\pi(x)$. We first approximate position $\pi(x)$ as follows:

$$\begin{aligned} \pi(x) &\simeq 1 + \sum_{y \neq x, y \in \chi} \mathrm{sign}(s_y - s_x) \\ &= 1 + \sum_{y \neq x, y \in \chi} \mathrm{sign}(f(\hat{y}) - f(\hat{x})) \end{aligned} \quad (5)$$

where $\hat{x}$ represents a d-dimensional feature vector of slit-lamp image x, s_x is the score of image x computed from ranking function $f(\hat{x})$, which can be in a linear form (i.e., $f(\hat{x}) =< \theta, \hat{x} >$, where $<, >$ denotes an inner product between θ and $\hat{x}$). Hence, θ is also a d-dimensional vector performing scaling on the d-dimensional feature space and there are d parameters in it to learn). The ranking function can also be in a nonlinear form [e.g., an exponential form: $f(\hat{x}) = \exp(< \theta, \hat{x} >)$] in this study. $\mathrm{sign}(s_y - s_x)$ is an signum function, whose value is positive when $s_y \geq s_x$ and negative otherwise. Hence, when the score of image x is smaller than that of image y (i.e., $s_x < s_y$), $\mathrm{sign}(s_y - s_x)$ becomes positive and $\pi(x)$ becomes larger due to (5), which matches the fact that images with lighter symptom (reflected by smaller score s_x) should be ranked in the rear of a ranked image list (i.e., larger value of position $\pi(x)$) in a descending order of severity of nuclear cataract disease.

Furthermore, we overcome the step transition characteristics of signum function— $\mathrm{sign}(\zeta)$ (ζ denotes its variable) by approximating it via a continuous hyperbolic tangent function—$\tanh(\zeta)$ [37]. An illustration of this approximation is shown in Fig. 4. The approximation step is as follows:

$$\begin{aligned} \mathrm{sign}(\zeta) \simeq \tanh(\zeta) &= \frac{\sinh(\zeta)}{\cosh(\zeta)} = \frac{\frac{e^{\zeta} - e^{-\zeta}}{2}}{\frac{e^{\zeta} + e^{-\zeta}}{2}} \\ &= \frac{e^{\zeta} - e^{-\zeta}}{e^{\zeta} + e^{-\zeta}} = \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1}. \end{aligned} \quad (6)$$

In this way, we can obtain a new continuous approximated position $\pi'(x)$

$$\begin{aligned} \pi'(x) &\simeq 1 + \sum_{y \neq x, y \in \chi} \frac{\exp(2\alpha(s_y - s_x)) - 1}{\exp(2\alpha(s_y - s_x)) + 1} \\ \alpha &> 0 \end{aligned} \quad (7)$$

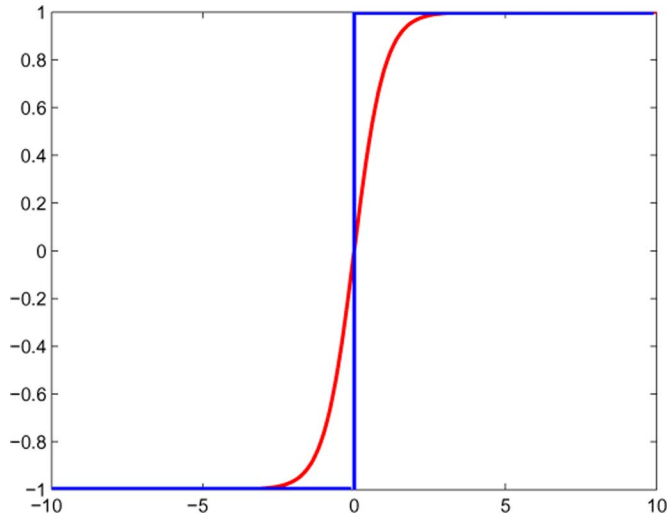

Fig. 4. An illustration of approximating a discrete signum function via a continuous hyperbolic tangent function.

where, $\alpha > 0$ is a positive scaling constant. Hence, a new continuous and differentiable approximation to NDCG, surrogate-normalized discounted cumulative gain (S-NDCG), can be proposed as follows:

$$\begin{aligned} \text{S-NDCG}(x) &= N_M^{-1} \sum_{x \in \chi} \frac{2^{r(x)} - 1}{\log_2(1 + \pi'(x))} \\ &\simeq N_M^{-1} \sum_{x \in \chi} \frac{2^{r(x)} - 1}{\log_2\left(2 + \sum_{y \neq x, y \in \chi} \frac{\exp(2\alpha(s_y - s_x)) - 1}{\exp(2\alpha(s_y - s_x)) + 1}\right)}. \end{aligned} \quad (8)$$

B. "Learning to Rank" Part II-Ranking Functions Learning via Direct Optimization On S-NDCG

A corresponding algorithm to directly optimize S-NDCG for learning ranking functions is listed in Table I. The key step here is to compute the gradient of S-NDCG with respect to the learned parameter $\theta(\partial \text{S-NDCG}(x)/\partial\theta)$ in Steps T4 and T5 of Table I. Detailed derivation is elaborated in Appendix A. The gradient can be computed as shown in (9) at the bottom of the next page.

Since the local optimizer of gradient ascent cannot guarantee a global optimal solution, we run T iterations of ranking functions learning θ_t initialized by previously learned $\theta_{(t-1)}$, where t is the tth iteration. Hence, after conducting the training phase in Table I, there are T ranking functions learned with their corresponding learned θ. Then, a validation phase is incorporated afterwards to select an optimal ranking function $f_{\mathrm{opt}}(\hat{x})$ as the one with the highest NDCG value (4), after applying all T learned ranking functions from the training phase to rank the validation set of images. Once $f_{\mathrm{opt}}(\hat{x})$ is determined, it can be used to rank new incoming images.

C. Nuclear Cataract Grading via Ranking of Slit-Lamp Images

When the learned optimal ranking function $f_{\mathrm{opt}}(\hat{x})$ is applied on the new incoming images, they can be sorted to form a ranked

TABLE I
LEARNING RANKING FUNCTIONS VIA DIRECT OPTIMIZATION ON S-NDCG IN OUR "LEARNING TO RANK" METHOD

Inputs	1. Slit-lamp images for training: $\{x \in \chi\}$
	2. Slit-lamp images for validation: $\{x_v \in \chi_v\}$
	3. Number of Iterations: T
	4. Learning rate: η
Training	
T1.	Initialize parameter θ of the ranking function $f(\hat{x})$ as θ_0
T2.	For t = 1 to T
T3.	Set $\theta = \theta_{t-1}$
T4.	Feed $\{x \in \chi\}$ to Equation 9 to calculate the gradient
T5.	Update θ via *gradient ascent*: $\theta = \theta + \eta \cdot \frac{\partial \text{S-NDCG}(x)}{\partial \theta}$
T6.	Set $\theta_t = \theta$
T7.	End for T2
Training	T learned ranking functions $f(\hat{x})$ with T corresponding learned parameters θ
Validation	
V1.	For j = 1 to T
V2.	Feed j^{th} learned ranking function $f_j(\hat{x})$ to $\{x_v \in \chi_v\}$ to rank validation images
V3.	Calculate its corresponding NDCG value using Equation 4
V4.	End for V1
V5.	Determine $f_{opt}(\hat{x})$ as the one with the highest NDCG value
Outputs	Optimal learned ranking function: $f_{opt}(\hat{x})$

image list in an order of nuclear cataract severity (according to scores calculated using $f_{\text{opt}}(\hat{x})$). An ungraded slit-lamp image x is sorted together with other slit-lamp images with clinicians' annotated grades in the ranked image list. Grade g_{x_i} of the ungraded slit-lamp image x located at position i of the ranked image list is interpolated using both scores from itself (s_{x_i}) and its neighboring images $(s_{x_{i-1}}, s_{x_{i+1}})$ as well as their annotated grades $(g_{x_{i-1}}, g_{x_{i+1}})$. The grading strategy is as below

$$g_{x_i} = \begin{cases} g_{x_{i+1}} & \text{if } g_{x_{i+1}} = g_{x_{i-1}} \\ g_{x_{i+1}} + \frac{s_{x_i} - s_{x_{i+1}}}{s_{x_{i-1}} - s_{x_{i+1}}} \times (g_{x_{i-1}} - g_{x_{i+1}}) & \text{if } g_{x_{i-1}} > g_{x_{i+1}} \\ g_{x_{i-1}} + \frac{s_{x_{i-1}} - s_{x_i}}{s_{x_{i-1}} - s_{x_{i+1}}} \times (g_{x_{i+1}} - g_{x_{i-1}}) & \text{if } g_{x_{i-1}} < g_{x_{i+1}} \end{cases} \tag{10}$$

where $s_{x_i} = f_{\text{opt}}(\hat{x_i})$; $s_{x_{i-1}} = f_{\text{opt}}(\hat{x_{i-1}})$; $s_{x_{i+1}} = f_{\text{opt}}(\hat{x_{i+1}})$. In this study, interpolated grades are decimal numbers, not integers.

IV. EXPERIMENTS AND DISCUSSION

A. Data Description and Implementation of Our "Learning to Rank" Method

The performance of our newly proposed "grading via ranking" scheme has been evaluated with a large dataset comprising of 1000 slit-lamp images from 1000 cases with different nuclear cataract disease severity obtained from an ongoing population-based study, the Singapore Malay Eye Study (SiMES) [38]. All images were captured by a Topcon DC-1 digital slit-lamp camera with FD-21 flash attachment. The slit beam was adjusted to completely fill the pupil, bisecting the lens from 12:00 to 6:00 at a 45° of angle. Focus was placed on the sulcus of the lens. Each slit-lamp image was saved as a 24-bit color image of the size 2048×1536 pixels. A clinical grade was provided to each slit-lamp image by senior ophthalmologists indicating the severity of nuclear cataract disease following the Wisconsin Cataract Grading System [9]. In this study, lens region, which is believed to be discriminative in identifying and diagnosing nuclear cataract disease in its conventional clinical diagnosis, was detected by an active shape model (ASM) method [16], [17], [39]. A 6-D local feature vector was extracted from the detected region within each slit-lamp image following previously published clinical work [9]. Detailed description of each feature vector dimension is explained as below.

Mean intensity (first dimension): The average intensity inside the lens region (to assess the nuclear opacity).

Color on posterior reflex (second to fourth dimensions): The posterior subcapsular reflex is suitable to judge the quality of the opacity color [9]. The position of central posterior subcapsular reflex was obtained via ASM. Mean values in each channel (hue, saturation and value) of the HSV color space in the region of central posterior subcapsular reflex were utilized as the next three dimensional features.

Visual axis profile analysis (fifth to sixth dimensions): Intensity change along the visual axis is important for nuclear cataract grading [9]. The visual axis profile was obtained from the intensity distribution on a horizontal line through central posterior reflex. A low-pass Chebyshev filter [40] was applied to smooth the profile. The first derivative of the profile was analyzed and the edge of lens nucleus was obtained. The mean intensity within the sulcus and intensity ratio between anterior lentil to posterior lentil were used as the last 2-D features.

All 1000 slit-lamp images were equally divided into 10 subsets for a ten-fold cross validation [41]: there are 800 training images (8 subsets), 100 validation images (1 subset) and 100 testing images (1 subset), respectively, in each fold. For our "learning to rank" method, we empirically set $T = 30$, $\eta = 0.05$, and $\alpha = 0.001$ as inputs in Table I.

$$\frac{\partial \text{S-NDCG}(x)}{\partial \theta} = N_M^{-1} \sum_{x \in \chi} \left(-\frac{2^{r(x)} - 1}{(\log_2(1 + \pi'(x)))^2} \cdot \frac{1}{(1 + \pi'(x)) \ln 2} \right) \cdot \left(\sum_{y \neq x, y \in \chi} \frac{-4\alpha \cdot \exp(-2\alpha(f(\hat{x}) - f(\hat{y})))}{(\exp(-2\alpha(f(\hat{x}) - f(\hat{y}))) + 1)^2} \cdot \left(\frac{\partial f(\hat{x})}{\partial \theta} - \frac{\partial f(\hat{y})}{\partial \theta} \right) \right) \tag{9}$$

B. "Learning to Rank" Algorithms for Ranking Performance Comparison

Besides our "learning to rank" method, we also incorporated three other popular "learning to rank" methods to learn ranking functions within our "grading via ranking" scheme. These methods include *RankBoost* [24], *AdaRank* [25] and *RankingSVM* [26]. Their basic ideas as well as implementation strategies in our experiments are explained below.

1) Rankboost: Freund *et al.* adopted the well-known boosting approach [41] into "learning to rank" methods, and proposed RankBoost as a pairwise approach to learn ranking functions [24]. The learning of ranking function is conducted by optimizing a total loss function (L) defined on the sum of losses from all lists (l) in an exponential form

$$\begin{aligned} L_{\text{total}} &= \sum_{l} \sum_{d_i \succ d_j} L(d_i \succ d_j) \\ &= \sum_{l} \sum_{d_i \succ d_j} \exp(-(f(d_i) - f(d_j))) \end{aligned} \tag{11}$$

where, $d_i \succ d_j$ denotes that image d_i is more relevant than image d_j; $f(d_i)$ is a ranking function f operating on image d_i. In our experiments, we used the toolbox from [42] for its implementation. The parameter to specify in this method is the number of iterations, which is of the same value as that of our "learning to rank" method for consistent experimental settings. The weak ranker of RankBoost was implemented as in its original work: each weak ranker was derived from a ranking feature by comparing its score on a given instance. Each weak ranker has a binary output of $\{0, 1\}$ [24].

2) AdaRank: Xu *et al.* proposed AdaRank as a listwise approach following adaptive boosting (AdaBoost [41]) to learn ranking functions [25]. The basic idea of AdaRank is to repeatedly construct weak rankers and linearly combine them together to form ranking functions. After T iterations, the learned ranking function f_T can be represented as follows:

$$\begin{aligned} f_T &= \sum_{k=1}^{T} \alpha_k h_k \\ \alpha_k &= \frac{1}{2} \ln \frac{\sum_{i=1}^{m} P_k(i)(1 + E(\pi(d_i, f_k), y_i))}{\sum_{i=1}^{m} P_k(i)(1 - E(\pi(d_i, f_k), y_i))} \\ P_k(i) &= \frac{\exp(-E(\pi(d_i, f_{k-1}), y_i))}{\sum_{j=1}^{m} \exp(-E(\pi(d_j, f_{k-1}), y_j))} \end{aligned} \tag{12}$$

where h_k is the kth weak ranker and α_k is its weight to update, $\pi(d, f)$ is a permutation of images set d resulted from ranking function f. $E(\pi(d, f), y) \in [-1, +1]$ is a ranking performance measure assessing the agreement between ranked permutation π and its ground truth y. In this study, we used Kendall's Tau coefficient [43] as the measure E with a definition: $E = P - Q/(1/2)(n(n-1))$, where n denotes the number of images within set d; $(1/2)(n(n-1))$ represents the number of image pairs from set d with n images; P and Q are numbers of concordant and discordant pairs, respectively [43]. The range of Kendall's Tau coefficient is within $[-1, +1]$, and an increasing value implies a better agreement between $\pi(d, f)$ and y. In this study, we constructed weak rankers as in its original work as well: features having the optimal weighted performance among all features are chosen as weak rankers [25]. For consistent experimental settings, the number of iterations in AdaRank is set to the same value as that of our "learning to rank" method.

3) RankingSVM: Joachims *et al.* extended the popular SVM technique into a "learning to rank" method, and proposed RankingSVM as a pairwise approach [26]. The main idea of RankingSVM is similar to conventional SVM, which aims to tune parameters by minimizing the sum of empirical loss and regularizer[44]. The constrained optimization function in terms of partial-order relationships within data pairs can be represented as follows:

$$\begin{aligned} &\min V(\omega, \epsilon) = \frac{1}{2}\omega^T\omega + C\sum_{i,j,q} \epsilon_{i,j,k}; \quad k = 1, \ldots, n \\ &s.t. \quad \forall (d_i, d_j) \in r_k^{\star} : \omega\varphi(d_i) \geq \omega\varphi(d_j) + 1 - \epsilon_{i,j,k} \end{aligned} \tag{13}$$

where ω is a normal vector perpendicular to the separating hyperplane, C is a trade-off between empirical loss and regularizer, ϵ is a slack variable measuring the degree of misclassification. The constraint $\omega\varphi(d_i) > \omega\varphi(d_j)$ reveals that image d_i is more relevant than d_j. RankingSVM is well formulated in the framework of structural risk minimization, and the ranking performance of RankingSVM has been appraised in several studies [26], [33], [45]. In our experiments, we used the binary codes in svm-light toolbox [46] for implementing RankingSVM. The trade-off C (13) is empirically set as 0.01 as suggested.

C. Ranking Experiments and Statistical Analysis

1) Ranking Experiments: For the above three compared "learning to rank" methods (i.e., RankBoost, AdaRank and RankingSVM), 900 images except for the 100 test images in each fold were used for training, as there is no validation needed for these methods. A simple example to rank the same 20 slit-lamp images with ranking functions learned from the same training images is shown in Fig. 5. The number below each image is its clinical ground truth (in this case, they are of integer-valued grades): 4 denotes the most severe symptom of nuclear cataract disease, while 1 represents the lightest symptom. It can be observed that our method achieved the least ranking errors among all four "learning to rank" methods (Our method—2 errors; RankBoost—16 errors; AdaRank—9 errors; RankingSVM—4 errors).

In our experiments, for 100 test images in each fold, they are divided into lists composed of small/medium/large numbers of slit-lamp images. To be specific, 20/50/100 slit-lamp images per list were specified for small/medium/large sets, respectively (i.e., set 20/50/100). In our experiments, we also divided training/validation images into different sub-lists of sizes 20/50/100 accordingly. Hence, the ranking function learning is conducted by maximizing the average S-NDCG over all sub-lists of the image data sets of the same size. The purpose is to test the ranking capability and stability of different "learning to rank" methods when handling various numbers of slit-lamp images. NDCG values (4) were calculated from these ranked

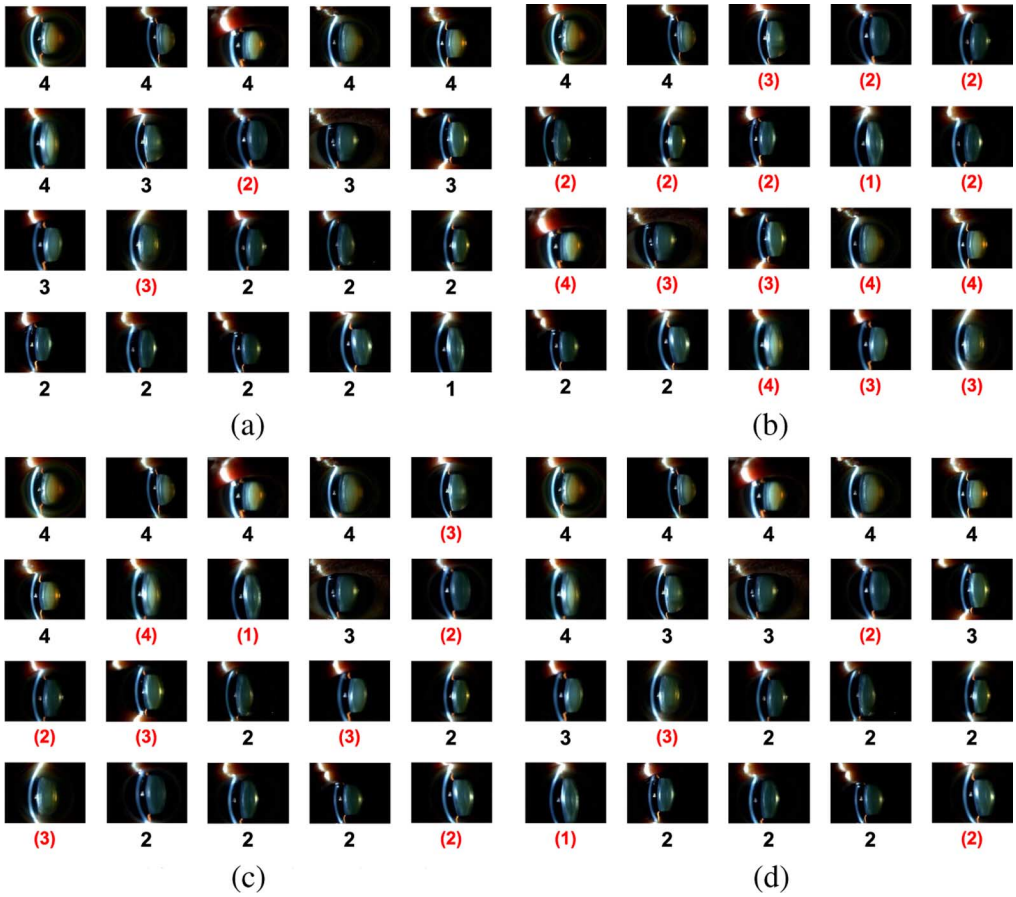

Fig. 5. An example of ranking the same 20 slit-lamp images (ranking errors are highlighted in red within brackets; the clinical ground truth is 4: first to sixth images; 3: seventh to eleventh images; 2: twelfth to nineteenth images; 1: twentieth image). (a) Ranking results by our method. (b) Ranking results by RankBoost. (c) Ranking results by AdaRank. (d) Ranking results by RankingSVM.

TABLE II
NDCG RESULTS OF ALL METHODS ON TEN-FOLD CROSS VALIDATION TEST FOR SETS OF 20 IMAGES (MEAN ± STANDARD DEVIATION)

Fold	Our Method (LRF)	Our Method (non-LRF)	RankBoost	AdaRank	RankingSVM
1	0.9423 ± 0.0315	0.9492 ± 0.0305	0.8407 ± 0.0693	0.9241 ± 0.0364	**0.9615 ± 0.0343**
2	0.8985 ± 0.0578	0.8985 ± 0.0578	0.7729 ± 0.0517	0.8904 ± 0.0642	**0.9695 ± 0.0207**
3	0.9304 ± 0.0468	0.9281 ± 0.0490	0.7917 ± 0.0489	0.9286 ± 0.0498	**0.9720 ± 0.0213**
4	**0.9443 ± 0.0361**	0.9382 ± 0.0375	0.8237 ± 0.0670	0.8874 ± 0.0488	0.9026 ± 0.0850
5	0.9023 ± 0.0828	**0.9166 ± 0.0568**	0.7183 ± 0.1167	0.8428 ± 0.0925	0.8810 ± 0.0802
6	0.8942 ± 0.0932	0.8946 ± 0.0915	0.7659 ± 0.0991	0.8594 ± 0.1102	**0.9089 ± 0.0770**
7	0.9685 ± 0.0169	0.9686 ± 0.0170	0.7925 ± 0.0467	0.9419 ± 0.0394	**0.9724 ± 0.0158**
8	**0.9718 ± 0.0133**	0.9630 ± 0.0149	0.8225 ± 0.0932	0.9627 ± 0.0178	0.9704 ± 0.0226
9	0.9327 ± 0.0507	**0.9354 ± 0.0501**	0.7600 ± 0.0763	0.8916 ± 0.0652	0.9209 ± 0.0768
10	0.9291 ± 0.0555	**0.9305 ± 0.0547**	0.7471 ± 0.0916	0.8660 ± 0.1004	0.9158 ± 0.0588

image lists to measure the ranking performance of different "learning to rank" methods quantitatively.

Detailed NDCG results of all folds are listed in Tables II–IV for set sizes of 20/50/100, respectively. For our newly proposed "learning to rank" method, we incorporated both linear ranking function (LRF) (i.e., $f(\hat{x}) =< \theta, \hat{x} >$) and nonlinear ranking function (non-LRF), which is of an exponential form in this study: $f(\hat{x}) = \exp(< \theta, \hat{x} >)$. From entries in Tables II–IV, it can be observed that, RankingSVM and our method achieve better ranking performance compared with RankBoost and AdaRank (Highest NDCG mean value in each fold is highlighted). To be specific, for set size of 20 (Table II), RankingSVM is superior among five out of the ten folds; while our method dominates in the other five folds (i.e., two folds by our method with LRF; the rest three folds by our method with non-LRF). For set size of 50 (Table III), RankingSVM is superior among four out of the ten folds; while our method is better in five other folds (i.e., four folds by our method with LRF and one fold by our method with non-LRF). For set size of 100 (Table IV), our method performs the best among five out of the ten folds, while RankingSVM dominates the other four folds. For our method (with either LRF or non-LRF), entries from the same fold of sets 20/50/100 (e.g., fold 1 in Tables II–IV) share similar NDCG values, which is indicative of the stability of our "learning to rank" method when ranking image lists of various sizes.

Based on all NDCG values, three box-and-whisker plots of NDCG were generated in Fig. 6 for the sets 20/50/100. In each

TABLE III
NDCG RESULTS OF ALL METHODS ON TEN-FOLD CROSS VALIDATION TEST FOR SETS OF 50 IMAGES (MEAN ± STANDARD DEVIATION)

Fold	Our Method (LRF)	Our Method (non-LRF)	RankBoost	AdaRank	RankingSVM
1	0.9459 ± 0.0468	0.9439 ± 0.0447	0.8091 ± 0.0076	0.9218 ± 0.0477	**0.9696 ± 0.0325**
2	0.9643 ± 0.0001	**0.9794 ± 0.0028**	0.8029 ± 0.0620	0.8703 ± 0.0249	0.9661 ± 0.0017
3	0.9433 ± 0.0133	0.9254 ± 0.0174	0.8061 ± 0.0404	0.9082 ± 0.0159	**0.9651 ± 0.0213**
4	**0.9535 ± 0.0303**	0.9433 ± 0.0436	0.7482 ± 0.1019	0.8138 ± 0.0770	0.8759 ± 0.1189
5	**0.9030 ± 0.0534**	0.8919 ± 0.0717	0.7020 ± 0.0809	0.8395 ± 0.1155	0.8723 ± 0.0996
6	0.8761 ± 0.0513	0.8954 ± 0.0632	0.7678 ± 0.0021	0.8345 ± 0.0322	**0.8982 ± 0.0548**
7	0.9367 ± 0.0220	0.9367 ± 0.0220	0.7774 ± 0.0095	0.9208 ± 0.0281	**0.9774 ± 0.0017**
8	0.9344 ± 0.0128	0.9321 ± 0.0133	0.7746 ± 0.0198	**0.9503 ± 0.0282**	0.9330 ± 0.0096
9	**0.9098 ± 0.0697**	0.9076 ± 0.0667	0.7525 ± 0.0722	0.8563 ± 0.1112	0.8996 ± 0.1053
10	**0.9111 ± 0.0965**	0.8975 ± 0.0911	0.7754 ± 0.0660	0.8325 ± 0.1222	0.8970 ± 0.0473

TABLE IV
NDCG RESULTS OF ALL METHODS ON TEN-FOLD CROSS VALIDATION TEST FOR SETS OF 100 IMAGES (MEAN)

Fold	Our Method (LRF)	Our Method (non-LRF)	RankBoost	AdaRank	RankingSVM
1	0.9414	0.9605	0.8409	0.9373	**0.9834**
2	**0.9734**	0.9374	0.7845	0.8710	0.9699
3	0.9558	0.9262	0.7895	0.9099	**0.9647**
4	0.9054	**0.9521**	0.7706	0.7889	0.8404
5	**0.8789**	0.8754	0.6912	0.7798	0.8241
6	0.8790	0.8725	0.7651	0.8258	**0.8908**
7	0.9409	0.9397	0.7957	0.9279	**0.9758**
8	0.9311	0.9211	0.7930	**0.9590**	0.9343
9	**0.8899**	0.8774	0.7519	0.8051	0.8622
10	0.8648	**0.8986**	0.7276	0.8120	0.8720

box, a red horizontal line is drawn across each box representing the median of NDCG, while the upper and lower quartiles of NDCG are depicted by blue lines above and below the median. A vertical dashed line is drawn up from the upper and lower quartiles to their most extreme data points, which are within a 1.5 IQR (Inter-Quartile Range) [47]. Each data point beyond the ends of 1.5 IQR is marked via a symbol of plus. It can be observed that, boxes of RankingSVM and our method (with both LRF and non-LRF) are located higher than those of RankBoost and AdaRank for all three cases (Fig. 6). It also substantiates our early observation in Tables II–IV that, RankingSVM and our method outperform RankBoost and AdaRank in ranking various numbers of slit-lamp images.

2) Statistical Analysis: It can be observed from entries in Tables II–IV that, our method is comparable with RankingSVM in ranking slit-lamp images. In order to evaluate it from a statistical point of view, we further conducted a statistical analysis composed of one-way analysis of variance (ANOVA) followed by a *post-hoc* multiple comparison test [47].

In one-way ANOVA, means of NDCG values from all methods are compared to test a hypothesis (H_0) that all NDCG means of various methods could be equivalent, against the general alternative that at least one method is different. P-value is used as an indicator to reveal whether H_0 exists or not. In our study, p-values for set sizes of 20/50/100 are all 0, which suggests that H_0 is an invalid hypothesis for all cases. Hence, the next step is to do more detailed paired comparisons. The reason to conduct paired comparison is because the generative alternative against H_0 is too general to reveal which method is superior from statistical point of view. Therefore, multiple comparison test is adopted to investigate it Appendix B.

Entries in Tables V–VII are results of multiple comparison test on NDCG by all methods for set sizes of 20/50/100, respectively. Each row indicates a paired comparison between two "learning to rank" methods, and there are two types of estimations for each paired comparison: one is single-value estimation, which estimates NDCG mean difference by a single value; the other is an interval estimation conducted via a 95% confidence interval (CI) Appendix C, which estimates a range that the NDCG mean difference is likely to be included. For instance, the second row of Table V is about the paired comparison between our method (LRF) and RankBoost for set size of 20. The NDCG mean difference from single-value estimation is 0.1490 (our method (LRF) - RankBoost), which suggests that our method (LRF) is better than RankBoost from single-value estimation perspective. The NDCG mean difference is likely to fall within a 95% CI $[0.1107, 0.1872]$. Since its upper and lower bounds are both positive, it gives a strong indication $(> 95\%)$ that, the NDCG mean difference (our method - RankBoost) is positive. Hence, out method is superior to RankBoost for set size of 20 from both single-value and interval estimation perspectives.

For paired comparisons between our method and RankingSVM (i.e., fourth rows of Tables V–VII for paired comparisons between our method (LRF) and RankingSVM; seventh rows of Tables V–VII for paired comparisons between our method (non-LRF) and RankingSVM), the analysis is similar. For set size of 20 (Table V), our method (with LRF and non-LRF) is 0.0028 and 0.0016 lower than RankingSVM, respectively, from single-value estimation perspective. The 95% CIs for the two paired comparisons are $[-0.0410, 0.0355]$ and $[-0.0399, 0.0366]$, which suggests RankingSVM is marginally better than our method (RankingSVM is superior in 53.59% and 52.16% cases, respectively following a general assumption that each CI is uniformly distributed). For set size of 50 (Table VI), our method (LRF and non-LRF) is 0.0051 and 0.0043 higher than RankingSVM, respectively from single-value estimation perspective. The 95% CIs for the two paired comparisons are $[-0.0484, 0.0586]$ and $[-0.0491, 0.0578]$, respectively, which suggests our method is marginally better than RankingSVM

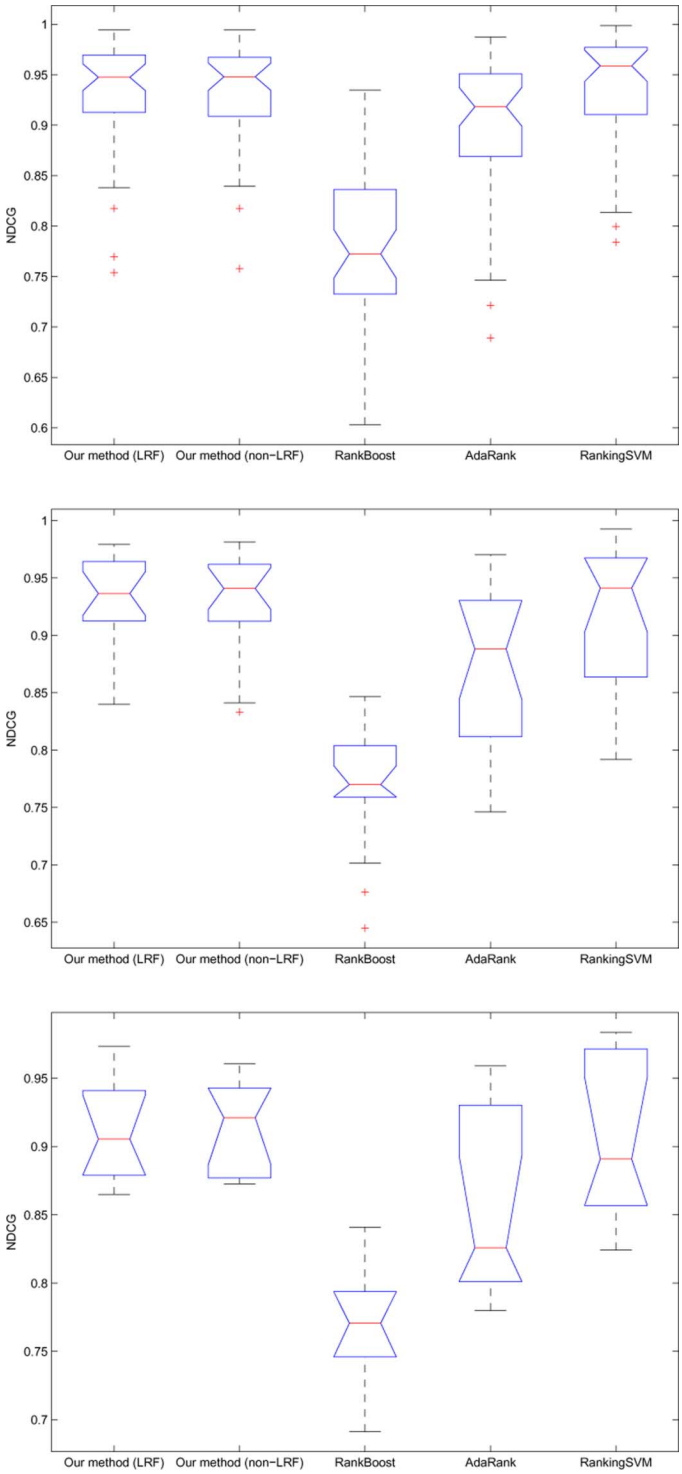

Fig. 6. Box-and-whisker plots of NDCG of achieved ranked image lists for sets of 20/50/100 images (up to down).

(our method with LRF and non-LRF are superior in 54.77% and 54.07% cases, respectively) for set size of 50. It is similar for set size of 100 in Table VII, that our method is marginally better than RankingSVM from both single-value and interval estimation perspectives. To sum up, after conducting one-way ANOVA followed by multiple comparison tests, our method and RankingSVM are comparable in ranking slit-lamp images from the statistical point of view.

D. Discussion

Given our method and RankingSVM are comparably effective in ranking performance, we further conduct a theoretical analysis on the computational complexity of the two "learning to rank" methods to compare their efficiency in ranking.

As mentioned in Section II-B2, RankingSVM is a pairwise approach, which utilizes instance pairs (d_i, d_j) as training data in learning ranking functions [(13)]. The training phase of RankingSVM needs to explicitly form all possible difference vectors $(\varphi(d_i) - \varphi(d_j))$ (13) from all instance pairs (d_i, d_j), and it sets up a standard classification procedure (i.e., SVM) for learning ranking functions. This method is costly in the training phase. Its computational complexity is of a quadratic order of training data size: $O(N^2)$ (indicated by the number of image pairs from N images: $C(N,2) = N!/((N-2)! \times 2!) = N(N-1)/2 \sim O(N^2)$).

For our method (both LRF and non-LRF), it is a listwise approach (Section II-B3). All N training images are fed into the training phase simultaneously for learning ranking functions. The computational complexity of our method is of an order $O(T \times N)$, where T denotes the number of iterations in our method. In this study, $N = 900$ (number of training images for ten-fold cross validation) and $T = 30$, an illustration of computational complexity of the two "learning to rank" methods with respect to the number of training images (N) is shown in Fig. 7. It can be observed that RankingSVM is more costly than our "learning to rank" method in computational complexity, especially when more and more images are incorporated in learning ranking functions. When learned ranking functions are used in performing the grading task, their computational complexity are equivalent since all applied "learning to rank" methods in our "grading via ranking" scheme use (10) for grades interpolation. Thus, although RankingSVM and our method are comparable in ranking performance, our method is more efficient than RankingSVM in learning ranking functions.

E. Nuclear Cataract Grading Test

To evaluate the nuclear cataract grading performance of our "grading via ranking" scheme, we use the same dataset composed of 1000 slit-lamp images. We compared our "grading via ranking" scheme with two existing nuclear cataract grading schemes: "grading via classification" [14] and "grading via regression" [16] applied on the same data. Statistical results of grading nuclear cataract within 1000 slit-lamp images are shown in Table VIII. For our "grading via ranking" scheme, we incorporated our "learning to rank" method as well as three other popular "learning to rank" methods in ranking slit-lamp images. Grades of slit-lamp images were then interpolated using (10) after obtaining ranked images lists for them. For entries in Table VIII, two measures are utilized to quantitatively measure the grading performance. One is grading accuracy, in which accurate gradings are assumed to be achieved when their grading errors (between predicted grades and clinical ground truth annotated by ophthalmologists) are within one integer grade (clinically important); the other is mean error (average grading errors of different schemes). It can be observed that, our newly proposed "grading via ranking" scheme with the new "learning to rank" method (LRF) achieves a 95.4% grading accuracy, which

TABLE V
MULTIPLE COMPARISON TEST RESULTS OF NDCG AMONG ALL METHODS FOR SETS OF 20 IMAGES

Method I	Method II	Estimated Mean Difference (I - II)	95% Confidence Interval
Our Method (LRF)	Our Method (non-LRF)	-0.0011	[-0.0394, 0.0371]
Our Method (LRF)	RankBoost	0.1490	[0.1107, 0.1872]
Our Method (LRF)	AdaRank	0.0354	[-0.0029, 0.0736]
Our Method (LRF)	RankingSVM	**-0.0028**	**[-0.0410, 0.0355]**
Our Method (non-LRF)	RankBoost	0.1501	[0.1119, 0.1884]
Our Method (non-LRF)	AdaRank	0.0365	[-0.0018, 0.0747]
Our Method (non-LRF)	RankingSVM	**-0.0016**	**[-0.0399, 0.0366]**
RankBoost	AdaRank	-0.1136	[-0.1519, -0.0754]
RankBoost	RankingSVM	-0.1518	[-0.1900, -0.1135]
AdaRank	RankingSVM	-0.0381	[-0.0764, 0.0001]

TABLE VI
MULTIPLE COMPARISON TEST RESULTS OF NDCG AMONG ALL METHODS FOR SETS OF 50 IMAGES

Method I	Method II	Estimated Mean Difference (I - II)	95% Confidence Interval
Our Method (LRF)	Our Method (non-LRF)	0.0008	[-0.0527, 0.0542]
Our Method (LRF)	RankBoost	0.1583	[0.1049, 0.2118]
Our Method (LRF)	AdaRank	0.0550	[0.0015, 0.1085]
Our Method (LRF)	RankingSVM	**0.0051**	**[-0.0484, 0.0586]**
Our Method (non-LRF)	RankBoost	0.1576	[0.1041, 0.2110]
Our Method (non-LRF)	AdaRank	0.0542	[0.0008, 0.1077]
Our Method (non-LRF)	RankingSVM	**0.0043**	**[-0.0491, 0.0578]**
RankBoost	AdaRank	-0.1033	[-0.1568, -0.0498]
RankBoost	RankingSVM	-0.1532	[-0.2067, -0.0998]
AdaRank	RankingSVM	-0.0499	[-0.1034, 0.0036]

TABLE VII
MULTIPLE COMPARISON TEST RESULTS OF NDCG AMONG ALL METHODS FOR SETS OF 100 IMAGES

Method I	Method II	Estimated Mean Difference (I - II)	95% Confidence Interval
Our Method (LRF)	Our Method (non-LRF)	-0.0033	[-0.0719, 0.0652]
Our Method (LRF)	RankBoost	0.1427	[0.0741, 0.2112]
Our Method (LRF)	AdaRank	0.0553	[-0.0132, 0.1239]
Our Method (LRF)	RankingSVM	**0.0058**	**[-0.0628, 0.0743]**
Our Method (non-LRF)	RankBoost	0.1460	[0.0775, 0.2145]
Our Method (non-LRF)	AdaRank	0.0587	[-0.0099, 0.1272]
Our Method (non-LRF)	RankingSVM	**0.0091**	**[-0.0594, 0.0776]**
RankBoost	AdaRank	-0.0874	[-0.1559, -0.0188]
RankBoost	RankingSVM	-0.1369	[-0.2054, -0.0684]
AdaRank	RankingSVM	-0.0496	[-0.1181, 0.0190]

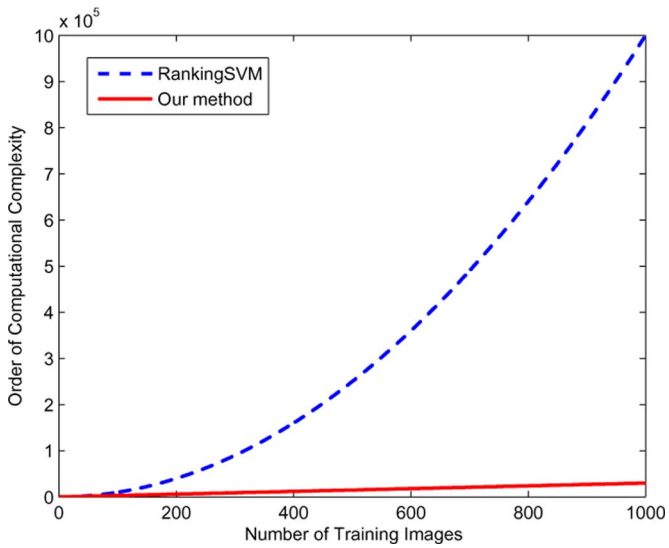

Fig. 7. An illustration of computational complexity comparison between our "learning to rank" method and RankingSVM.

is the highest among all grading schemes. Its mean error is 0.3432, which is the lowest. Also, the newly proposed "grading via ranking" scheme with our "learning to rank" method as well as RankingSVM perform apparently better than that with AdaRank and RankBoost, which suggests that better ranking performance can lead to better grading results in our "grading via ranking" scheme. Grading results achieved from ranked images lists of various sizes by our "grading via ranking" scheme with one particular "learning to rank" method are similar (e.g., grading accuracy of 95.7%, 95.2%, 95.4%, respectively, for set sizes of 20/50/100 for our method with LRF). For the two existing nuclear cataract grading schemes, numbers of slit-lamp images per list do not affect their grading performance since all training slit-lamp images (in our ten-fold cross validation, 900 training images in one fold) are fed into the training phase to tune their parameters simultaneously. Their tuned parameters are used in grading 100 images in each fold and statistical grading results of all ten folds are listed in Table VIII. From the comparison of the nuclear cataract grading performance of our "grading via ranking" scheme against them, it can be observed that, ours performs significantly better.

A histogram of the difference between ground truth of grades and predicted grades of the newly proposed "grading via ranking" scheme with our new "learning to rank" method (LRF) for set size of 20 is shown in Fig. 8. It can be observed

TABLE VIII
STATISTICAL RESULTS OF NUCLEAR CATARACT GRADING BY OUR PROPOSED GRADING SCHEMES WITH COMPARISON OF OTHERS

Grading Scheme		Set 20 Grading Accuracy	Set 20 Mean Error	Set 50 Grading Accuracy	Set 50 Mean Error	Set 100 Grading Accuracy	Set 100 Mean Error	Average Grading Accuracy	Average Mean Error
Grading via Ranking	Our Method (LRF)	**95.7%**	**0.3378**	**95.2%**	0.3481	**95.4%**	**0.3438**	**95.4%**	**0.3432**
	Our Method (non-LRF)	95.1%	0.3384	93%	0.3990	91.1%	0.5184	93.1%	0.4186
	RankingSVM	93%	0.3626	94%	**0.3479**	93.4%	0.3529	93.4%	0.3545
	AdaRank	85%	0.5169	83.6%	0.5474	83.4%	0.5515	84.0%	0.5386
	RankBoost	78.4%	0.6265	76.4%	0.6330	76.6%	0.6370	77.1%	0.6322
Grading	via Regression	87.3%	0.5652	87.3%	0.5652	87.3%	0.5652	87.3%	0.5652
Grading	via Classification	76.8%	0.8135	76.8%	0.8135	76.8%	0.8135	76.8%	0.8135

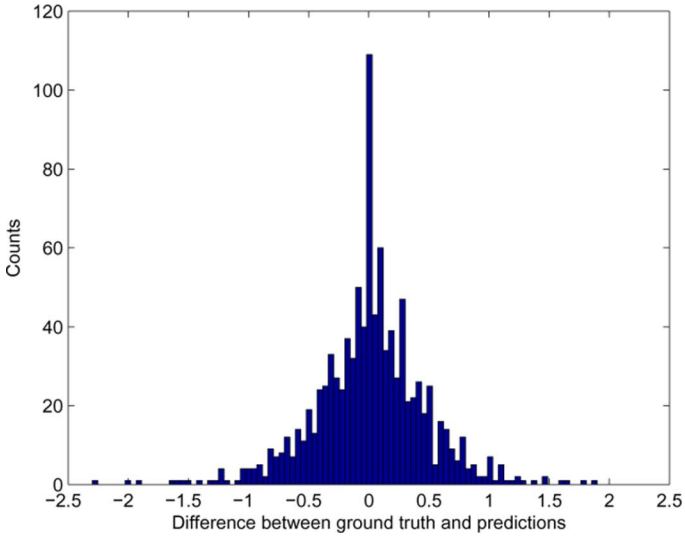

Fig. 8. Histogram of the difference between ground truth of grades and predicted grades of the newly proposed "grading via ranking" scheme with our "learning to rank" method (LRF) for sets of 20 images.

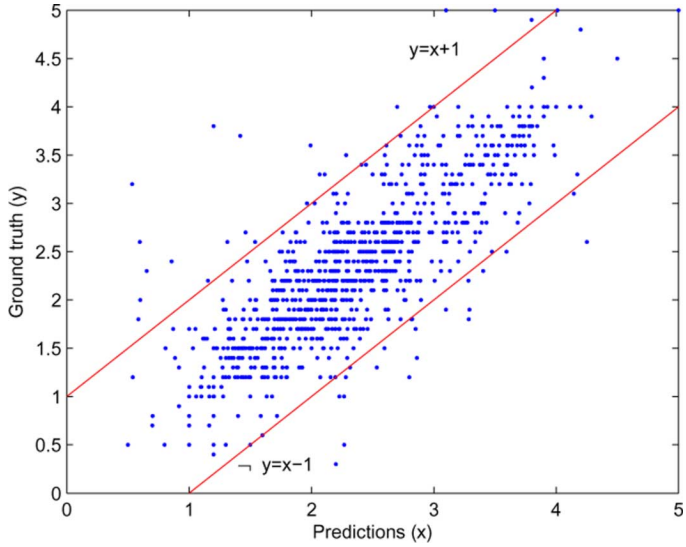

Fig. 9. Comparison between ground truth of grades and predicted grades of the newly proposed "grading via ranking" scheme with our "learning to rank" method (LRF) for sets of 20 images.

that, most results are within one integer grade error. An illustration of comparison between ground truth of grades and predicted grades of the newly proposed "grading via ranking" scheme with our "learning to rank" method (LRF) for set size of 20 is shown in Fig. 9. Points scattered between the two red lines denote incidences of grading errors less than one integer grade. To sum up, it can be concluded that the newly proposed "grading via ranking" scheme with our new "learning to rank" method performs better than existing schemes compared in this study in grading nuclear cataract in 1000 slit-lamp images.

V. CONCLUSION

A novel nuclear cataract grading scheme is proposed following conventional clinical decision-making process in the paper. Grade of an ungraded slit-lamp image is predicted with the help of its neighboring images in a ranked image list, which is achieved using an ranking function learned via a newly proposed "learning to rank" method. A new approximation to a ranking evaluation measure is proposed to incorporate direct optimization on learning ranking functions. Our grading via ranking scheme has been evaluated by a large dataset composed of 1000 slit-lamp images. Experimental results demonstrate that our grading scheme performs better than the other existing grading schemes in grading nuclear cataract disease in the slit-lamp images dataset. Our grading via ranking scheme can easily accommodate different ranking methods to learn ranking functions. It has been demonstrated that both our "learning to rank" method based on direct optimization of S-NDCG and RankingSVM are comparable in ranking performance, while our proposed "learning to rank" method has less computational complexity. Our grading via ranking scheme can be utilized as a training tool for junior clinicians to learn diagnostic decision from images with similar disease severity (neighboring images in a ranked image list) and their diagnosis results given by senior clinicians. It can be used in research to analyze different diagnoses with similar symptoms as well. In the future work, we plan to explore the nuclear cataract grading task via other techniques, such as ordinal regression [48]. We also plan to incorporate other recently proposed methods [49], [50] to efficiently implement RankingSVM in our "grading via ranking" scheme.

APPENDIX A
DERIVATION OF THE GRADIENT OF S-NDCG

After applying the *chain rule*, the gradient of S-NDCG(x) with respect to θ becomes

$$\frac{\partial \text{S-NDCG}(x)}{\partial \theta} = \frac{\partial \text{S-NDCG}(x)}{\partial \pi'(x)} \cdot \frac{\partial \pi'(x)}{\partial \theta} = N_M^{-1} \sum_{x \in \chi} \frac{\partial \frac{2^{r(x)}-1}{\log_2(1+\pi'(x))}}{\partial \pi'(x)} \cdot \frac{\partial \pi'(x)}{\partial \theta} \quad (14)$$

where, the first term of (14) is derived as follows:

$$\frac{\partial \frac{2^{r(x)}-1}{\log_2(1+\pi'(x))}}{\partial \pi'(x)} = -\frac{2^{r(x)}-1}{(\log_2(1+\pi'(x)))^2}\cdot\frac{1}{(1+\pi'(x))\ln 2}. \tag{15}$$

Furthermore, we rewrite $\pi'(x)$ in (15) as follows:

$$\begin{aligned}\pi'(x) &\simeq 1+\sum_{y\neq x, y\in\chi}\frac{\exp(2\alpha(s_y-s_x))-1}{\exp(2\alpha(s_y-s_x))+1}\\ &= 1+\sum_{y\neq x, y\in\chi}\frac{\exp(-2\alpha(s_{xy}))-1}{\exp(-2\alpha(s_{xy}))+1}\\ s_{xy} &= s_x - s_y.\end{aligned} \tag{16}$$

Apply the chain rule to the second term of (14) after incorporating results in (16)

$$\begin{aligned}&\frac{\partial \pi'(x)}{\partial\theta}\\ &= \frac{\partial \pi'(x)}{\partial s_{xy}}\cdot\frac{\partial s_{xy}}{\partial\theta}\\ &= \sum_{y\neq x, y\in\chi}\left(\frac{-2\alpha\cdot\exp(-2\alpha s_{xy})(\exp(-2\alpha s_{xy})+1)}{(\exp(-2\alpha s_{xy})+1)^2}\right.\\ &\qquad\left.+\frac{2\alpha\cdot\exp(-2\alpha s_{xy})(\exp(-2\alpha s_{xy})-1)}{(\exp(-2\alpha s_{xy})+1)^2}\right)\cdot\frac{\partial s_{xy}}{\partial\theta}\\ &= \sum_{y\neq x, y\in\chi}\frac{-4\alpha\cdot\exp(-2\alpha(f(\hat{x})-f(\hat{y})))}{\left(\exp(-2\alpha\big(f(\hat{x})-f(\hat{y})\big))+1\right)^2}\\ &\quad\cdot\left(\frac{\partial f(\hat{x})}{\partial\theta}-\frac{\partial f(\hat{y})}{\partial\theta}\right).\end{aligned} \tag{17}$$

Hence, after substituting derivation results of (15) and (17) into (14), it becomes (18) shown at the bottom of the page which is the (9).

Appendix B
Multiple Comparison Test

The reason to adopt multiple comparison test, instead of ordinary t-test in statistics here is because, there are many pairs of methods to compare. If an ordinary t-test is applied in this situation, β (Appendix C) would apply to each comparison, so the chance of incorrectly finding a significant difference when there is no real difference would increase with the number of comparisons. Multiple comparison tests can avoid this situation since it provides an upper bound on the probability in the case that any comparison will be incorrectly found significant.

Appendix C
95%

A significance level is specified for determining the cutoff value of the *t statistic*. Commonly, $\beta = 0.05$ is applied to insure that, when there is no real difference, one will incorrectly find a significant difference no more than 5%. Hence, the confidence level is $100(1-\beta)\% = 95\%$ in this case.

References

[1] The World Health Report: Life in the 21st Century—A Vision for all World Health Organization [Online]. Available: http://www.who.int/entity/whr/1998/en/whr98 en.pdf

[2] Magnitude and Causes of Visual Impairments World Health Organization [Online]. Available: http://www.who.int/mediacentre/factsheets/fs282/en/index.html.

[3] B. A. Henderson, R. Pineda, II, C. Ament, S. H. Chen, and J. Y. Kim, *Essentials of Cataract Surgery*. Thorofare, NJ: SLACK, 2007.

[4] Advances in eye research—Cataract Research to Prevent Blindness [Online]. Available: http://www.rpbusa.org/rpb/eye_info/cataract/cataract_facts.pdf

[5] T. Y. Wong, S. C. Loon, and S. M. Saw, "The epidemiology of age related eye diseases in Asia," *Br. J. Ophthalmol.*, vol. 90, pp. 506–511, 2006.

[6] A. L. Coleman and J. C. Morrison, *Management of Cataracts and Glaucoma*. Oxfordshire, U.K.: Taylor Francis, 2005.

[7] C. L. Martinyi, C. F. Bahn, and R. F. Meyer, *Slit Lamp: Examination and Photography*. Sedona, AZ: Time One Ink, 2007.

[8] S. K. West, F. Rosenthal, H. S. Newland, and H. R. Taylor, "Use of photographic techniques to grade nuclear cataracts," *Invest. Ophthalmol. Vis. Sci.*, vol. 29, no. 1, pp. 73–77, 1988.

[9] B. E. Klein, R. Klein, K. L. Linton, Y. L. Magli, and M. W. Neider, "Assessment of cataracts from photographs in the beaver dam eye study," *Ophthalmology*, vol. 97, no. 11, pp. 1428–1433, 1990.

[10] J. M. Sparrow, A. J. Born, N. A. Brown, W. Ayliffe, and A. R. Hill, "The oxford clinical cataract classification and grading system," *Int. Ophthalmol.*, vol. 9, no. 4, pp. 207–225, 1986.

[11] L. T. Chylack, J. K. Wolfe, D. M. Singer, M. C. Leske, M. A. Bullimore, I. L. Bailey, J. Friend, D. McCarthy, and S. Y. Wu, "The lens opacities classification system III," *Arch. Ophthalmol.*, vol. 111, no. 6, pp. 831–836, 1993.

[12] B. Thylefors, L. T. Chylack, K. Konyanma, K. Sasaki, R. Sperduto, H. R. Taylor, and S. West, "A simplified cataract grading system," *Ophthalmic Epidemiol.*, vol. 9, no. 2, pp. 83–95, 2002.

[13] N. J. Ferrier, "Automated identification of the anatomical features in slit lamp photographs of the lens," *Invest. Ophthalmol. Vis. Sci.*, vol. 43, pp. 435–435, 2002.

[14] S. Fan, C. R. Dyer, L. Hubbard, and B. Klein, "An automatic system for classification of nuclear sclerosis from slit-lamp photographs," in *Proc. Med. Image Comput. Computer-Assisted Intervention (MICCAI)*, 2003, pp. 592–601.

[15] D. D. Duncan, O. B. Shukla, S. K. West, and O. D. Schein, "New objective classification system for nuclear opacification," *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.*, vol. 14, no. 6, pp. 1197–1204, 1997.

$$\begin{aligned}\frac{\partial \text{S-NDCG}(x)}{\partial\theta} &= N_M^{-1}\sum_{x\in\chi}\left(-\frac{2^{r(x)}-1}{(\log_2(1+\pi'(x)))^2}\cdot\frac{1}{(1+\pi'(x))\ln 2}\right)\\ &\cdot\left(\sum_{y\neq x, y\in\chi}\frac{-4\alpha\cdot\exp(-2\alpha(f(\hat{x})-f(\hat{y})))}{\left(\exp(-2\alpha(f(\hat{x})-f(\hat{y})))+1\right)^2}\cdot\left(\frac{\partial f(\hat{x})}{\partial\theta}-\frac{\partial f(\hat{y})}{\partial\theta}\right)\right)\end{aligned} \tag{18}$$

[16] H. Li, J. H. Lim, J. Liu, T. Y. Wong, A. Tan, J. Wang, and P. Mitchell, "Image based grading of nuclear cataract by svm regression," in *Proc. SPIE-Med. Imag.*, 2008, vol. 6915, pp. 691536.1–691536.8.

[17] H. Li, J. H. Lim, J. Liu, P. Mitchell, A. Tan, J. Wang, and T. Y. Wong, "A computer-aided diagnosis system of nuclear cataract," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1690–1698, Jul. 2010.

[18] M. Patasius, V. Marozas, D. Jegelevicius, and A. Lokosevicius, "Ranking of color space components for detection of blood vessels in eye fundus images," in *Proc. Eur. Conf. Int. Fed. Med. Biol. Eng.*, 2009, pp. 464–467.

[19] E. Arimura, C. Matsumoto, S. Okuyama, S. Takana, S. Hasbimoto, and Y. Sbimomura, "Retinal contraction and metamorphopsia scores in eyes with idiopathic epiretinal membrane," *Invest. Ophthalmol. Vis. Sci.*, vol. 46, no. 8, pp. 2961–2966, 2005.

[20] M. Miura, A. Elsner, M. Osako, K. Yamada, T. Agawa, M. Usui, and T. Iwasaki, "Spectral imaging of the area of internal limiting membrane peeling," *Retina*, vol. 25, no. 4, pp. 468–472, 2005.

[21] S. Agarwal, C. Cortes, and R. Herbrich, in *Learning to Rank Workshop at NIPS 2005—Overview* [Online]. Available: http://web.mit.edu/shivani/www/ Ranking-NIPS-05/

[22] K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," in *Proc. ACM Special Interest Group Inf. Retrieval (SIGIR)*, 2000, pp. 41–48.

[23] R. Baeza-Yates and B. Ribeiro-Neto, *Morden Information Retrieval*. Reading, MA: Addison Wesley, 1990.

[24] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preference," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, 2003.

[25] J. Xu and H. Li, "AdaRank: A boosting algorithm for information retrieval," in *Proc. ACM SIGIR*, 2007, pp. 391–398.

[26] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM Special Interest Group Knowledge Discovery Data Mining (SIGKDD)*, 2002, pp. 133–142.

[27] R. Herbirch, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Adv. Large Margin Classifiers*, pp. 115–132, 2000.

[28] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 145–152.

[29] K. Crammer and Y. Singer, "Pranking with ranking," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 641–647.

[30] E. F. Harrington, "Online ranking/collaborative filtering using the perceptron algorithm," in *Proc. ICML*, 2003, pp. 250–257.

[31] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Proc. NIPS*, 2002, pp. 937–944.

[32] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. ICML*, 2005, pp. 89–96.

[33] Y. B. Cao, J. Xu, T. Y. Liu, H. Li, Y. L. Huang, and H. W. Hon, "Adaptive ranking svm to document retrieval," in *Proc. ACM SIGIR*, 2006, pp. 186–193.

[34] Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. ICML*, 2007, pp. 129–136.

[35] T. Qin, X. D. Zhang, M. F. Tsai, D. S. Wang, T. Y. Liu, and H. Li, "Query-level loss function for information retrieval," *Inf. Process. Manage.*, vol. 44, no. 2, pp. 838–855, 2007.

[36] M. F. Tsai, T. Y. Liu, T. Qin, H. H. Chen, and W. Y. Ma, "FRank: A ranking method with fidelity loss," in *Proc. ACM SIGIR*, 2007, pp. 383–390.

[37] A. Jeffery, *Handbook of Mathematical Formulas and Integrals*, 2nd ed. New York: Academic, 2000.

[38] A. W. Foong, S. M. Saw, J. L. Loo, S. Shen, S. C. Loon, M. Rosman, T. Aung, D. T. Tan, E. S. Tai, and T.-Y. Wong, "Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study," *Ophthalmic Epidemiol.*, vol. 14, no. 1, pp. 25–35, 2007.

[39] H. Li and O. Chutatape, "Boundary detection of optic disk by a modified ASM method," *Pattern Recognit.*, vol. 36, no. 9, pp. 2093–2104, 2003.

[40] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed. Reading, MA: Addison-Wesley, 1992.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[42] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, SVM and Kernel Methods Matlab Toolbox [Online]. Available: http:// asi.insa-rouen.fr/enseignants~arakotom/toolbox/index.html

[43] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, pp. 81–93, 1938.

[44] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[45] T. Qin, T. Y. Liu, W. Lai, X. D. Zhang, D. S. Wang, and H. Li, "Ranking with multiple hyperplanes," in *Proc. ACM SIGIR*, 2007, pp. 279–286.

[46] T. Joachims, SVM Lght—An implementation of support vector machine in C [Online]. Available: http:// svmlight.joachims.org

[47] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Pacific Grove, CA: Duxbury, 2007.

[48] P. McCullagh, "Regression models for ordinal data," *J. R. Stat. Soc., Series B (Methodology)*, vol. 42, no. 2, pp. 109–142, 1980.

[49] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," *Inf. Retrieval*, 2009.

[50] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD*, 2006, pp. 217–226.