A Computer-Aided Diagnosis System of Nuclear Cataract via Ranking

Wei Huang¹, Huiqi Li², Kap Luk Chan¹, Joo Hwee Lim², Jiang Liu²,
and Tien Yin Wong³

¹ School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore

² Institute for Infocomm Research, Agency for Science,
Technology and Research, Singapore

³ National University of Singapore, Singapore National Eye Center and Singapore
Eye Research Institute

Abstract. A novel computer-aided diagnosis system of nuclear cataract via ranking is firstly proposed in this paper. The grade of nuclear cataract in a slit-lamp image is predicted based on its neighboring labeled images in a ranked images list, which is achieved using an optimal ranking function. A new ranking evaluation measure is proposed for learning the optimal ranking function via direct optimization. Our system has been tested by a large dataset composed of 1000 slit-lamp images from 1000 different cases. Both experimental results and comparison with several state-of-the-art methods indicate the superiority of our system.

1 Introduction

Cataract, the clouding or opacity of normally clear human lens, is the leading cause of blindness globally. Based on locations of opacity, age-related cataracts can be categorized into three types: *posterior sub-capsular cataract*, *cortical cataract* and *nuclear cataract*, among which nuclear cataract (cataract formed in the nucleus) is the most common type [1]. In clinical diagnosis, a reasonable grade of nuclear cataract is often assigned by trained ophthalmologists to each *slit-lamp image* by comparing its opacity severity with a set of standard photos [2]. This grading scheme is often subjective (inter-observer agreement is only around 65% [2]) and time-consuming. Recently, there are research efforts towards automatic diagnosis of nuclear cataract to improve its grading objectivity and save workload [3], [4]. In [3], intensities on visual axis were used as features and *linear regression* was applied to grade nuclear cataract. In [4], more diverse features and advanced technique (*support vector machine* (svm) regression) were adopted. Both systems aim to output a final grade, but they offer little help to train junior ophthalmologists. One of the most important means to gain skills in clinics is practice and learning from similar cases.

In this paper, we, computer scientists and clinicians working closely together, propose a novel computer-aided diagnosis (CAD) system, which firstly regards nuclear cataract grading as a *ranking* task. The flowchart of our system is shown

Fig. 1. Flowchart of our system

Fig. 2. Diagram of feature extraction scheme

in Fig. 1. Given a slit-lamp image to grade, ranking aims to sort the slit-lamp image together with other slit-lamp images, whose grades are known, to form a ranked images list according to their nuclear cataract severity measured by a *ranking function*. Grade of the slit-lamp image is then predicted using its neighboring images in the ranked list. In order to achieve an optimal ranking function for ranking these images, we incorporate the “*learning to rank*” technique [5], in which we propose a new ranking evaluation measure that can be directly optimized for learning ranking functions. This is different from many other traditional “*learning to rank*” methods, which adopt various complicated approaches, such as *adaptive boosting* [6], [7] or *svm* [8], to learn ranking functions, based on traditional ranking evaluation measures [9]. Our contribution lies in three aspects: (1) We firstly propose a novel CAD system that grades nuclear cataract in slit-lamp images via ranking using local features; (2) We propose a new ranking evaluation measure, and incorporate “*learning to rank*” technique to learn an optimal ranking function via direct optimization based on the new measure; (3) Our system has been tested by a large dataset composed of 1000 slit-lamp images from 1000 different cases. Experimental results have been compared with ones achieved from several state-of-the-art methods and evaluated with clinical ground truth.

2 Methodology

2.1 Local Feature Extraction from Slit-Lamp Images

Robust feature extraction is essential for any medical imaging application. Here we apply a model-based method to detect the anatomical contour of lens, and then, we extract local discriminative feature from detected regions following previously published clinical works [2], [10]. The whole process is illustrated in Fig. 2. The foreground of lens image is segmented by thresholding 20% - 30% brightest pixels in the grey image. Horizontal profile clustering and vertical profile clustering are employed to estimate lens structure as an ellipse, which is based on the largest cluster and used as the initialization of detected lens

Table 1. Description of feature extracted from each slit-lamp image

Feature Number	Description
1	mean intensity inside the lens contour
2 - 4	color information on posterior reflex in HSI color space
5	mean intensity within the central part of lens
6	intensity ratio between anterior lentil and posterior lentil

contour. After that, a modified *active shape model* (ASM) [11] is employed to further evolve the contour to better match the lens. Based on the lens structure detected by ASM, a 6-dimensional local feature vector is selected and extracted from each slit-lamp image. The specification of feature vector is described in Table 1.

2.2 Surrogate-NDCG and Ranking Function Learning

In this section, we aim to sort slit-lamp images to form a ranked images list according to their nuclear cataract severity. We incorporate the “learning to rank” technique here, to learn an optimal ranking function, which is used to rank images. In order to learn the function, we propose a new ranking evaluation measure, named *surrogate-NDCG* (*Normalized Discounted Cumulative Gain*), which can be directly optimized for learning ranking functions.

NDCG is a traditional position-based ranking evaluation measure, which can handle multiple-level relevance judgements [9]. It was chosen here because nuclear cataract grades annotated by ophthalmologists in our application are also multiple ordinal values. The definition of NDCG is as follows [9]:

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

where, x is a slit-lamp image and χ is the set of images to be ranked; $r(x)$ and $\pi(x)$ are the grade and position of image x in the ranked images list sorted by decreasing nuclear cataract severity, respectively; N_n is a normalization term denoting the maximum of DCG, which can be achieved when all images are sorted in a perfect order of decreasing severity. 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 $\pi(x)$. Hence, many traditional state-of-the-art “learning to rank” methods [6], [7], [8] incorporate complicated techniques for learning ranking functions with indirect optimization based on NDCG.

To solve the problem, we first approximate position $\pi(x)$ as follows:

$$\pi(x) = 1 + \sum_{y \neq x, y \in \chi} \mathbf{1}\{s_{xy} < 0\}, \quad s_{xy} = s_x - s_y = f(\theta, \hat{x}) - f(\theta, \hat{y}) \quad (2)$$

where, s_x is the *score* of slit-lamp image x computed from ranking function $f(\theta, \hat{x})$, which is of a linear form ($f(\theta, \hat{x}) = \theta \hat{x}$) in this work; θ and $\hat{x}$ are parameters of $f(\theta, \hat{x})$ to learn and extracted feature of image x , respectively. $\mathbf{1}\{s_{xy} < 0\}$

is an *indictor function*, whose value is positive when $s_{xy} < 0$ and negative otherwise. Hence, when the score of image x is smaller than y 's ($s_{xy} = s_x - s_y < 0$), $\mathbf{1}\{s_{xy} < 0\}$ becomes positive and $\pi(x)$ becomes larger, which matches the fact that images with lighter symptom (represented by smaller score s_x) should be ranked in the rear of a ranked images list (larger position $\pi(x)$) in a descending order of nuclear cataract severity.

Furthermore, we overcome the *step characteristics* of indictor function $\mathbf{1}\{s_{xy} < 0\}$ by approximating it via a continuous *hyperbolic tangent function*. In this way, we can achieve a new continuous approximated position $\hat{\pi}(x)$:

$$\hat{\pi}(x) = 1 + \sum_{y \neq x, y \in \chi} \mathbf{1}\{s_{xy} < 0\} \simeq 1 + \sum_{y \neq x, y \in \chi} \frac{\exp(-2\alpha s_{xy}) - 1}{\exp(-2\alpha s_{xy}) + 1} \quad (3)$$

where, $\alpha > 0$ is a scaling constant. Hence, our new continuous and differentiable ranking evaluation measure *surrogate-NDCG* ($\widehat{NDCG}$) can be achieved:

$$\widehat{NDCG} = N_n^{-1} \sum_{x \in \chi} \frac{2^{r(x)} - 1}{\log_2(1 + \hat{\pi}(x))} \quad (4)$$

Our method to directly optimize $\widehat{NDCG}$ for learning ranking functions is listed in Table 2. The key step is to compute gradient of $\widehat{NDCG}$ with respect to θ ($\nabla \widehat{NDCG}(\theta)$) in step 6. After applying the *chain rule*, the gradient is as follows:

$$\begin{aligned} \nabla \widehat{NDCG}(\theta) &= \frac{\partial \widehat{NDCG}}{\partial \theta} = N_n^{-1} \sum_{x \in \chi} \frac{\partial \frac{2^{r(x)} - 1}{\log_2(1 + \hat{\pi}(x))}}{\partial \hat{\pi}(x)} \cdot \frac{\partial \hat{\pi}(x)}{\partial \theta} \\ &= N_n^{-1} \sum_{x \in \chi} \left(-\frac{2^{r(x)} - 1}{(\log_2(1 + \hat{\pi}(x)))^2} \cdot \frac{1}{(1 + \hat{\pi}(x)) \ln 2} \right) \cdot \\ &\quad \left(\sum_{y \neq x, y \in \chi} \frac{-4\alpha \cdot \exp(-2\alpha(f(\theta, \hat{x}) - f(\theta, \hat{y})))}{[\exp(-2\alpha(f(\theta, \hat{x}) - f(\theta, \hat{y}))) + 1]^2} \cdot \left(\frac{\partial f(\theta, \hat{x})}{\partial \theta} - \frac{\partial f(\theta, \hat{y})}{\partial \theta} \right) \right) \quad (5) \end{aligned}$$

Detailed derivation is omitted here. The formula of step 7 in Table 2 shows the actual implementation of maximizing NDCG by minimizing the negative of it. For inputs in Table 2, we also apply a series of n positive scaling constants $\{\alpha_1, \dots, \alpha_n\}$ as candidates of α in Equation 3. Therefore, $(n \times T)$ ranking functions $f(\theta, \hat{x})$ can be learned as outputs from Table 2. In a following validation step, NDCG value (Equation 1) of each outputted ranking function is calculated from its ranked images list, which is achieved by applying the specified ranking function to rank validation images. The one with the highest NDCG value will be chosen as the optimal ranking function $f_{opt}(\theta, \hat{x})$.

2.3 Automatic Diagnosis of Nuclear Cataract

Using the optimal ranking function $f_{opt}(\theta, \hat{x})$ achieved from section 2.2, a slit-lamp image x to grade can be sorted together with other slit-lamp images with

Table 2. Learning ranking functions via direct optimization on $N\widehat{DCG}$

Inputs	1. m training sets of slit-lamp images x and their grades $r(x)$
	2. Iteration times: T
	3. Learning rate: η
	4. n scaling constants candidates $\{\alpha_1, \dots, \alpha_n\}$
Algorithm	
	1. For $\alpha \in \{\alpha_1, \dots, \alpha_n\}$
	2. Initialize parameters θ of ranking function $f(\theta, \hat{x})$ as θ_0
	3. For $t = 1$ to T
	4. Set $\theta = \theta_{t-1}$
	5. For $i = 1$ to m , apply <i>gradient descent</i> approach:
	6. Feed $(x, r(x))$ to Equation 5 for computing $\nabla N\widehat{DCG}(\theta)$
	7. Update parameter $\theta = \theta - \eta \cdot \nabla N\widehat{DCG}(\theta)$
	8. End for 5
	9. Set $\theta_t = \theta$
	10. End for 3
	11. End for 1
Outputs	$(n \times T)$ learned ranking functions $f(\theta, \hat{x})$ with $(n \times T)$ learned θ

known grades, to form a ranked images list in the order of nuclear cataract severity. After that, grade g_{x_i} of the slit-lamp image x located at position i in the ranked images list can be predicted by using both scores computed from ranking function $f_{opt}(\theta, \hat{x})$ and grades $g_{x_{i-1}}$ and $g_{x_{i+1}}$ of its neighboring images in the ranked images list as follows:

$$g_{x_i} = \begin{cases} g_{x_{i+1}} & \text{if } g_{x_{i+1}} = g_{x_{i-1}} \\ g_{x_{i+1}} + \frac{s_{x_i} - s_{x_{i+1}}}{s_{x_{i-1}} - s_{x_{i+1}}} \times (g_{x_{i-1}} - g_{x_{i+1}}) & \text{if } g_{x_{i-1}} > g_{x_{i+1}} \\ g_{x_{i-1}} + \frac{s_{x_{i-1}} - s_{x_i}}{s_{x_{i-1}} - s_{x_{i+1}}} \times (g_{x_{i+1}} - g_{x_{i-1}}) & \text{if } g_{x_{i-1}} < g_{x_{i+1}} \end{cases} \quad (6)$$

3 Experiments and Discussion

3.1 Data Description

Our CAD system has been tested by a large dataset composed of 1000 slit-lamp images from 1000 different cases including both normal and pathological conditions achieved from a population-based study, the *Singapore Malay Eye Study* (SiMES) [12]. All images were captured using a Topcon DC-1 digital slit-lamp camera with FD-21 flash attachment. The slit-lamp images were saved as 24-bit color images with the size of 2048×1536 pixels. Clinical ground truth were provided by ophthalmologists using *Wisconsin cataract grading system* [2].

3.2 Ranking Experiments and Statistical Analysis

After feature extraction, the 1000 images are divided into 5 subsets for *five-fold cross validation*; there are 3 training sets, 1 validation set and 1 testing set respectively. We empirically set $m = 30$, $T = 200$, $\eta = 0.01$ and

Fig. 3. An example of ranking the same 20 slit-lamp images (ranking errors are highlighted in red within brackets; the clinical ground truth is 4: 1st-6th images; 3: 7th-11th images; 2: 12th-19th images; 1: 20th image)

$\alpha \in \{0.1, 0.2, 0.5, 1, 1.5, 2, 2.5, 5, 7, 10\}$ as inputs in Table 2. There are three state-of-the-art “learning to rank” methods for ranking comparison, including *RankBoost* [6], *AdaRank* [7] and *RankingSVM* [8]. Related parameters setting are equivalent to ours. For these three methods, 4 subsets beside the testing set were used for training as there were no validations included. An example of ranking the same 20 testing images is shown in Fig. 3. The number below each image is its clinical ground truth. In this testing set, 4 represents the severest symptom of nuclear cataract, while 1 is the lightest. It can be seen that our method achieved the best ranking performance (only 2 ranking errors) among all methods (RankBoost: 16 errors; AdaRank: 9 errors; RankingSVM: 4 errors).

To evaluate our method quantitatively, we utilized ranking evaluation measure NDCG (Equation 1), and results are listed in Table 3.2. We can easily see that RankingSVM and our method achieve better performance than RankBoost and AdaRank, although comparison between RankingSVM and our method is not clear. Thus, a statistical test is conducted to further discern which method is superior from statistical point of view. After performing one-way *analysis of variance* (ANOVA), the *p-value* is nearly 0, which casts serious doubts on the null hypothesis that RankingSVM and our method share the same NDCG means. Hence, a series of post-hoc *multiple comparison tests* are applied and partial results are listed in Table 3.2. The estimated NDCG mean of our method is

Table 3. NDCG of all methods on five-fold cross validation test (mean±variance)

Fold	Our Method	RankBoost	AdaRank	RankingSVM
1	0.9770±0.0013	0.8068±0.0046	0.9078±0.0028	0.9769±0.0013
2	0.9693±0.0018	0.8077±0.0033	0.9105±0.0026	0.9597±0.0022
3	0.9169±0.0028	0.7421±0.0110	0.8779±0.0081	0.9206±0.0038
4	0.9729±0.0009	0.8075±0.0051	0.9523±0.0011	0.9714±0.0014
5	0.9396±0.0025	0.7536±0.0064	0.9064±0.0049	0.9410±0.0019

Table 4. Multiple comparison test results of our method against other methods

Method I	Method II	Estimated Mean Diff. (I-II)	95% Confidence Int.
Our Method	RankBoost	0.1708	[0.1401, 0.2016]
Our Method	AdaRank	0.0454	[0.0147, 0.0762]
Our Method	RankingSVM	0.0017	[-0.0291, 0.0325]

0.0017 higher, and a 95% confidence interval is [-0.0291, 0.0325], which shows that for less than 50% cases (about 44.8% if the confidence interval is uniformly distributed), RankingSVM performs better. For the majority, ours is superior.

3.3 Nuclear Cataract Grading from Ranking

To evaluate automatic diagnosis performance of our CAD system, we used the same 1000 slit-lamp images dataset for grading test. Grades in our system are computed from Equation 6 based on ranking results achieved from section 3.2. The mean error between predicted grades of our system and ophthalmologists' clinical ground truth is only 0.541. The grading accuracy measuring grading errors within one integer grade (clinically important) is 91.4%. Experimental results show that our CAD system is promising in grading nuclear cataract of slit-lamp images from a large dataset.

4 Conclusion

A novel computer-aided diagnosis (CAD) system of nuclear cataract via ranking is firstly proposed in this paper. Grade of a slit-lamp image is predicted using its neighboring images in a ranked images list, which is achieved using a learned ranking function via direct optimization based on a new proposed ranking evaluation measure. Our system has been tested by a large images dataset composed of 1000 slit-lamp images from 1000 different cases. Results show that our system performs better than several state-of-the-art methods. Our system can also be utilized as a training tool for junior clinicians to learn diagnostic decision from similar images (neighboring images in the ranked images list), which have already been diagnosed by senior clinicians. It can be used in research to analyze difference diagnosis with similar symptoms as well.

Acknowledgments. We would like to acknowledge Singapore A*STAR SBIC SiRIAN grant for providing the clinical images and ophthalmologists from Singapore Eye Research Institute for valuable clinical inputs. We want to express our gratitude to Prof. Paul Mitchell, Prof. Jie Jin Wang and Ms Ava Tan from the University of Sydney, Australia for providing technical inputs and providing the ground truth of cataract grading.

References

1. World Health Organization (WHO): Magnitude and Causes of Visual Impairments (2002), <http://www.who.int/mediacentre/factsheets/fs282/en/index.html>
2. Klein, B., Klein, R., Linton, K., Magli, Y., Neider, M.: Assessment of Cataracts from Photographs in the Beaver Dam Eye Study. *Ophthalmology* 97(11), 1428–1433 (1990)
3. Fan, S., Dyer, C.R., Hubbard, L., Klein, B.: An automatic system for classification of nuclear sclerosis from slit-lamp photographs. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2878, pp. 592–601. Springer, Heidelberg (2003)
4. Li, H., Lim, J., Liu, J., Wong, T.-Y., Tan, A., Wang, J., Paul, M.: Image Based Grading of Nuclear Cataract by SVM Regression. In: SPIE Proceeding of Medical Imaging, vol. 6915, 691536–691536(8) (2008)
5. Agarwal, S., Cortes, C., Herbrich, R.: Proceeding of NIPS Workshop on Learning to Rank (2005)
6. Freund, Y., Iyer, R., Schapire, R., Singer, Y.: An Efficient Boosting Algorithm for Combining Preference. *Journal of Machine Learning Research* 4, 933–969 (2003)
7. Xu, J., Li, H.: AdaRank: A Boosting Algorithm for Information Retrieval. In: ACM SIGIR, pp. 391–398 (2007)
8. Joachims, T.: Optimizing Search Engines using Clickthrough Data. In: ACM SIGKDD, pp. 133–142 (2002)
9. Jarvelin, K., Kekalainen, J.: IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR, pp. 41–48 (2000)
10. Chylack, L., Wolfe, J., Singer, D., Leske, M.: The Lens Opacities Classification System III. *Archives of Ophthalmology* 111, 831–836 (1993)
11. Li, H., Chutatape, O.: Boundary detection of optic disk by a modified ASM method. *Pattern Recognition* 36(9), 2093–2104 (2003)
12. Foong, A., Saw, S., Loo, J., et al.: Rationale and Methodology for a Population-based Study of Eye Diseases in Malay People: The Singapore Malay Eye Study (SiMES). *Ophthalmic Epidemiology* 14(1), 25–35 (2007)