Semantic Reconstruction-Based Nuclear Cataract Grading from Slit-Lamp Lens Images

Yanwu Xu[1(✉)], Lixin Duan[2], Damon Wing Kee Wong[1], Tien Yin Wong[3], and Jiang Liu[1,4]

[1] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, Singapore
yaxu@i2r.a-star.edu.sg
[2] Amazon, Seattle, USA
[3] Singapore Eye Research Institute, Singapore, Singapore
[4] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Beijing, China

Abstract. Cataracts are the leading cause of visual impairment and blindness worldwide. Cataract grading, i.e. assessing the presence and severity of cataracts, is essential for diagnosis and progression monitoring. We present in this work an automatic method for predicting cataract grades from slit-lamp lens images. Different from existing techniques which normally formulate cataract grading as a regression problem, we solve it through reconstruction-based classification, which has been shown to yield higher performance when the available training data is densely distributed within the feature space. To heighten the effectiveness of this reconstruction-based approach, we introduce a new semantic feature representation that facilitates alignment of test and reference images, and include locality constraints on the linear reconstruction to reduce the influence of less relevant reference samples. In experiments on the large ACHIKO-NC database comprised of 5378 images, our system outperforms the state-of-the-art regression methods over a range of evaluation metrics.

1 Introduction

Cataracts are a clouding of the lens that reduces transmission of light to the retina. They may be caused by a variety of factors, including age, ultraviolet radiation, and genetics. This obstruction of light can seriously impair vision and may even progress into blindness [1]. Due to its prevalence particularly among the elderly, there is a need to screen for them in an efficient and cost-effective manner.

Most commonly, cataracts develop in the nucleus, which is the central layer of the lens. The opacification and coloration caused by nuclear cataracts is visible in cross-sectional views of the lens in slit-lamp images. Currently, nuclear cataracts are diagnosed by ophthalmologists directly using a slit-lamp microscope, or graded by clinicians who assess the presence and severity of a cataract

© Springer International Publishing AG 2016
S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 458–466, 2016.
DOI: 10.1007/978-3-319-46726-9_53

by comparing slit-lamp images against a set of protocol photographs [2–4] such as that shown in Fig. 1. However, manual assessments can be both subjective and time-consuming [4].

The need for objectivity and efficiency in nuclear cataract grading has led to the development of several computer-aided systems [5–10]. These systems generally operate with three main steps: lens structure detection, feature extraction, and severity prediction. Most prior art formulate the severity prediction as a regression problem on certain visual features. In the state-of-the-art method of [5], bag-of-features (BOF) descriptors are extracted from RGB and HSV color channels of different lens sections, and group sparsity regression (GSR) is used to jointly select the features, parameters and models for grading. In [6], 21 predefined features are extracted from different sections of the lens, and then are fed into a pre-learned RBF kernel-based Support Vector Regressor (SVR) to estimate the cataract grade.

While regression-based methods have achieved higher accuracy than other previous techniques, we observe that the dense sampling of cataract grades in the available training data allows for a more direct grading prediction. When training samples are densely distributed in the feature space, higher accuracy can be achieved through reconstruction from the samples, where class membership of an input is estimated based on reconstruction accuracy from similar instances within each class. Compared to regression, a reconstruction-based approach is less reliant on discriminative feature quality and more robust to small inter-class margins, such as those that exist for the continuous space of cataract grades. These advantages of reconstruction-based classification have been exploited for human gait recognition [11], where a large number of training samples are densely distributed in a feature space that is compact due to the relatively narrow range of gait differences.

For nuclear cataract grading, however, the reconstruction-based approach is ineffective when employed in a straightforward manner. In our preliminary tests, linear reconstruction of lens images after alignment and size normalization of lens sections led to grading performance much lower than the state-of-the-art ***BOF+GSR*** method [5]. This is mainly due to inadequate lens alignment. Since lenses vary in both size and shape, a non-rigid structural alignment of lenses is needed for accurate reconstruction, but is challenging to accomplish. To address this problem, we propose to model the test and reference images with a new semantic representation of lens structure that is less sensitive to slight misalignments, instead of processing in the original raw image space. In addition, we improve the accuracy of reconstruction-based cataract grade prediction by accounting for the degree of similarity between the test image and reference images. By ignoring reference images that are not well-aligned to the test image as done in the similarity ranking-based CBMIR approach [8], the alignment issue is further diminished.

Our proposed method essentially follows the manual grading protocol, as it directly compares with reference images through the alignment and reconstruction procedure, and it compares intensity/color and contrast patterns via the

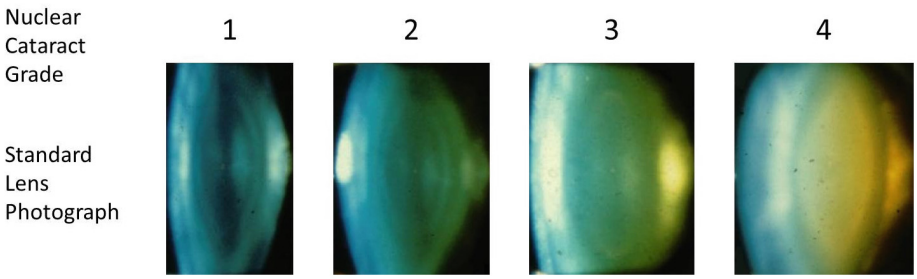

Fig. 1. Standard photographs of the Wisconsin nuclear cataract grading system. From left to right, the severity of the nuclear cataracts increases, with greater brightness and lower contrast between anatomical landmarks. In addition, the color of the nucleus and posterior cortex exhibits more of a yellow tint due to brunescence.

semantic feature representation. With this approach, our system attains higher overall performance than the state-of-the-art, and has the potential to be applied to other ocular diseases such as angle closure glaucoma detection and optical cup localization.

2 Nuclear Cataract Grading Through Semantic Reconstruction

We formulate nuclear cataract grading as a linear reconstruction problem with a similarity-weighted constraint. For a given slit-lamp test image, our algorithm follows the steps of lens structure detection, semantic feature extraction, and linear reconstruction with reference images.

2.1 Lens Structure Detection

Detection of lens structures in slit-lamp images is a well-studied problem with effective solutions [5–10]. For this purpose, we employ techniques similar to those used in [5,6]. As illustrated in Fig. 2, the lens structure detection proceeds with the following steps:

1. Using the active shape model based lens structure detection proposed in [6], each lens image is separated into three sections: anterior cortex, nucleus, and posterior cortex. The visual axis is located as well.
2. A lens cross-section is extracted around the visual axis, using a bounding box with a height of h pixels ($h = 128$ in our implementation) to obtain the central parts of the nucleus, anterior cortex and posterior cortex.
3. Features are extracted from only the nucleus and posterior cortex sections, since the anterior cortex contains no discriminant information for nuclear cataract grading [5]. This practice is also supported by clinical protocol [4], where nuclear cataracts are graded based on the intensity and visibility of nuclear landmarks and the color of the nucleus and posterior cortex.

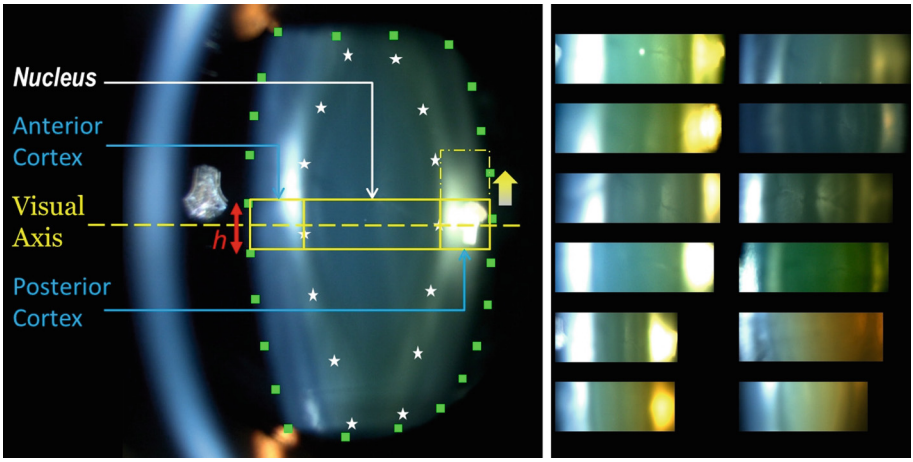

Fig. 2. Left: illustration of lens structure detection, with the initially detected lens structure (solid yellow boxes) and final detected posterior cortex (dashed yellow box). Right: examples of detected lens cross-sections. On the left are initially detected cross-sections with sharp reflections in the posterior cortex. On the right are cross-sections that were detected without the posterior cortex reflections. Since the anterior cortex will be discarded, reflections there need not be avoided. (Color figure online)

4. As illustrated on the right side of Fig. 2, bright spots may appear in the extracted posterior cortex section, due to reflections of the photographic flash. The presence of these sharp reflections may greatly reduce grading accuracy, so we avoid them by simply shifting the bounding box of the posterior cortex vertically with a step size of $h/2$ until it has a mean scaled value lower than a threshold value θ_p, where $\theta_p = 192$ in our implementation.

2.2 Semantic Feature Representation

After detection, the posterior cortex is divided into $s \times s$ ($s = 3$ in our implementation) half-overlapping grid cells, and the nucleus is partitioned into $s \times 2s$ half-overlapping grid cells. For each of the RGB and HSV color channels[1], a grid cell is represented by its mean intensity $\hat{t}$ and entropy e, defined as $e = -\sum_{l=0}^{255} p_l \log p_l$ where p_l is the probability of intensity l in the grid for a given color channel. With this data, each image is represented by a feature vector with $3 \times s \times s \times 6 \times 2 = 36s^2$ dimensions.

With the downsampling and half-overlapping grid cells, this representation becomes less sensitive to slight misalignments caused by differences in lens shape. We note that the feature vectors used in previous works such as [6], though containing features such as intensity ratios and edge strength that are useful

[1] HSV values are linearly scaled to the range $[0, 255]$ for consistency with the RGB channels.

for discriminative classification, are less suitable for dealing with the alignment problem, which is critical to the success of reconstruction-based techniques.

2.3 Similarity Weighted Linear Reconstruction (SWLR)

Suppose we have a dictionary that consists of n reference images, denoted by $\mathbf{D} = \{\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_n}\} \in \mathbb{R}^{f \times n}$ where each column d_i denotes a reference image expressed by its semantic feature vector. For a given test image expressed as $\mathbf{y} \in \mathbb{R}^{f \times 1}$, we compute the optimal linear reconstruction coefficients $\mathbf{w} \in \mathbb{R}^{n \times 1}$, $\sum_{i=1}^{n} w_i = 1$, $w_i \geq 0$, that minimize the reconstruction error $||\mathbf{y} - \mathbf{Dw}||^2$. Our objective function also includes a cost term that penalizes the use of references that are less similar to the test image. Let us denote the costs for the reference images in $\mathbf{D}$ as the vector $\mathbf{c} = \{c_1, c_2, \cdots, c_n\}^\top \in \mathbb{R}^{n \times 1}$, where c_i is the cost of using $\mathbf{d_i}$ for reconstruction. The overall cost term can then be expressed as $||\mathbf{c} \odot \mathbf{w}||^2$ where $\odot$ denotes the Hadamard product. Combining this cost term with the reconstruction error gives the following objective function:

$$\min_{\mathbf{w}} ||\mathbf{y} - \mathbf{Dw}||^2 + \lambda ||\mathbf{c} \odot \mathbf{w}||^{\mathbf{2}}, \quad s.t. \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. This objective can be minimized in closed form using the Lagrange multiplier method:

$$\begin{aligned} \mathbf{w} =& \frac{1}{\mathbf{1}^\top (\hat{\mathbf{D}}^\top \hat{\mathbf{D}} + \lambda \mathbf{C}^\top \mathbf{C}) \mathbf{1}} (\hat{\mathbf{D}}^\top \hat{\mathbf{D}} + \lambda \mathbf{C}^\top \mathbf{C})^{-1} \mathbf{1}, \\ \hat{\mathbf{D}} =& (\mathbf{1} \otimes \mathbf{y} - \mathbf{D}), \end{aligned} \tag{2}$$

where $\mathbf{C} = \text{diag}(\mathbf{c})$ and $\otimes$ denotes the Kronecker product. The cost c_i is defined as the χ^2-distance between the test image $\mathbf{y}$ and the i-th reference image $\mathbf{d_i}$, *i.e.*,

$$c_i = \sum_{j=1}^{f} \frac{(y_j - d_{i,j})^2}{2(y_j + d_{i,j})}, \tag{3}$$

where $d_{i,j}$ denotes the j-th entry of reference image $\mathbf{d_i}$. We note that the inclusion of entropy in the semantic feature helps to exclude misaligned references in the SWLR algorithm, since high entropy indicates the presence of structural variations, and differences in entropy caused by misalignment are penalized by this cost function.

Finally, the test image is graded as $\mathbf{w}^\top \mathbf{g}$, where $\mathbf{g}$ denotes the corresponding cataract grades of reference images in dictionary $\mathbf{D}$.

3 Experiments

To evaluate our method, we first compare it to the state-of-the-art nuclear cataract grading methods [5,6]. We then validate its major components by comparing to versions of our technique with certain components replaced.

3.1 Experimental Setting

Dataset. All the experiments are performed on the large ACHIKO-NC dataset used in [5,6], which is comprised of 5378 images with decimal grading scores that range from 0.3 to 5.0. The scores are determined by professional graders based on the Wisconsin protocol [4]. The protocol takes the ceiling of each decimal grading score as the integral grading score, *i.e.*, a cataract with a decimal grading score of 2.4 has an integral grading score of 3. ACHIKO-NC consists of 94 images of integral grade 1, 1874 images of integral grade 2, 2476 images of integral grade 3, 897 images of integral grade 4, and 37 images of integral grade 5. All left eye images are flipped horizontally so that they can be processed in the same way as right eye images.

Evaluation Criteria. For a fair comparison to prior art, we measure grading accuracy using the same four evaluation criteria as in [5,6], namely the exact integral agreement ratio (R_0), the percentage of decimal grading errors ≤ 0.5 ($R_{e0.5}$), the percentage of decimal grading errors ≤ 1.0 ($R_{e1.0}$), and the mean absolute error (ε), which are defined as

$$\begin{aligned} R_0 &= \frac{|\lceil G_{gt} \rceil = \lceil G_{pr} \rceil|_0}{N}, & R_{e0.5} &= \frac{\left||G_{gt} - G_{pr}| \leq 0.5\right|_0}{N}, \\ R_{e1.0} &= \frac{\left||G_{gt} - G_{pr}| \leq 1.0\right|_0}{N}, & \varepsilon &= \frac{\sum |G_{gt} - G_{pr}|}{N}, \end{aligned} \tag{4}$$

where G_{gt} denotes the ground-truth clinical grade, G_{pr} denotes the predicted grade, $\lceil \cdot \rceil$ is the ceiling function, $|\cdot|$ denotes the absolute value, $|\cdot|_0$ is a function that counts the number of non-zero values, and N is the number of testing images ($N = |G_{gt}|_0 = |G_{pr}|_0$). $R_{e0.5}$ has the most narrow tolerance among the four evaluation criteria, which makes it more significant in evaluating the accuracy of grading.

Testing Method. To examine generalization ability, we follow the repeated test settings in [5], *i.e.*, in each round, 100 training samples are randomly selected from all the 5378 images, with 20 images for each grade, and the remaining 5278 images are used for testing. In training, optimal parameters are selected for each method by cross-validation, where half of the images (50 images with 10 per grade) are used as the dictionary, the other half used for testing, and the set of parameters with the smallest average ε is chosen. The result of each round is obtained by testing the remaining 5278 images using all the 100 images as the dictionary together with the determined optimal parameters.

3.2 Comparison to State-of-the-art Regression Methods

We first compare our method to the state-of-the-art techniques, namely ***BOF+GSR*** [5] and ***RBF ϵ-SVR*** [6], using the same dataset, experimental

Table 1. Cataract grading performance vs. state-of-the-art regression based methods

Method	R_0	$R_{e0.5}$	$R_{e1.0}$	ε
Proposed SF+SWLR	**0.696±0.008**	**0.871±0.007**	**0.991±0.001**	**0.332±0.006**
BOF+GSR [5]	0.682±0.004	0.834±0.005	0.985±0.001	0.351±0.004
RBF ϵ-SVR [6]	0.658±0.014	0.824±0.016	0.981±0.004	0.354±0.014
SF+RBF ϵ-SVR	0.645±0.018	0.826±0.020	0.875±0.010	0.449±0.018
Our improvement over [5]	2.05 %	4.44 %	0.61 %	5.41 %

setting and reporting methods. The results are listed in Table 1, where ***SF*** refers to the proposed semantic feature and ***SWLR*** refers to the proposed similarity weighted linear reconstruction method. According to the results, our method is shown to surpass [5,6] in all four evaluation criteria.

In Table 1, our method is also compared to the application of ***RBF ϵ-SVR*** on the proposed semantic feature. The results show that the proposed feature has less discriminative power than the features used in [5,6]. This is not unexpected, since our semantic feature is designed for more robust alignment rather than discriminative power. In addition, our method has an extra advantage over [6] in that a more detailed segmentation of the lens is not needed for extracting discriminative features.

In summary, reconstruction and regression are two significantly different approaches to solve the cataract grading problem. The proposed ***SWLR*** selects more relevant sample *images* for *each individual testing image* to perform grading, while ***GSR*** selects more discriminative *feature vector entries* over *all training images* and assumes that good prediction can be obtained with these features on all the test images.

3.3 Comparison to Alternative Versions of Our Method

To validate the components of our technique, we compare it to alternative versions without the similarity based regularizer (referred to as ***LR***), and by applying ***SWLR*** on different feature sets. The results are given in Table 2, and the following observations can be made:

- Comparing ***SWLR*** to ***LR*** shows that the similarity constraint is helpful for selecting more relevant/representative reference images for each individual test image. With better reconstruction, the performance is improved. We note that applying ***LR*** on only the k-nearest neighbours (k-NN) also does not yield performance as high as ***SWLR***, since k cannot be fixed to a value that is suitable for all test images. By contrast, ***SWLR*** can adaptively determine a set of proper reference images to reconstruct each individual test image.
- Comparing ***SWLR*** using different feature sets shows that though some features may have greater discriminative power, they are less suitable in the context of reconstruction-based classification.

Table 2. Cataract grading performance using different reconstruction techniques and features

Method	Feature	R_0	$R_{e0.5}$	$R_{e1.0}$	ε
Proposed SWLR	***Proposed SF***	**0.696±0.008**	**0.871±0.007**	**0.991±0.001**	**0.332±0.006**
LR	*Proposed SF*	0.685±0.009	0.846±0.010	0.986±0.002	0.348±0.009
Proposed SWLR	*BOF* [5]	0.586±0.035	0.758±0.031	0.815±0.018	0.484±0.019
Proposed SWLR	[6]	0.655±0.021	0.773±0.027	0.801±0.014	0.406±0.017

3.4 Discussion

Similarity Metric. We compared our ***SWLR*** method using different similarity metrics, namely χ^2 distance and Gaussian distance, defined as $\exp(||\mathbf{y} - \mathbf{d_i}||^2/\sigma^2)$ where σ is a parameter that accounts for imaging noise. It was observed that χ^2 distance is more effective than Gaussian distance for the proposed feature representation, with metrics of (R_0, $R_{e0.5}$, $R_{e1.0}$, ε) for χ^2 distance being (0.696±0.008, 0.871±0.007, 0.991±0.001, 0.332±0.006) and for Gaussian distance being (0.688±0.009, 0.868±0.007, 0.990±0.001, 0.337±0.007).

Processing Speed. On a four-core 2.4 GHz PC with 16 GB RAM, our method takes 17.73 s on average to process an image, with 1.36 s for feature extraction and only 0.001 s for prediction because of the small dictionary size. This processing speed slightly exceeds the 20.45 s per image of [5], which takes 4.23 s for feature extraction and 0.00001 s for prediction. It is also faster than the 25.00 s per image of [6], which spends 8.76 s for feature extraction and 0.02 s for prediction.

4 Conclusion

For grading the severity of nuclear cataracts from slit-lamp lens images, we proposed a reconstruction-based approach with a new semantic feature representation and a similarity weighted regularizer. In tests on the *ACHIKO-NC* dataset comprised of 5378 images, our approach achieves significant improvements over the state-of-the-art regression based methods [5,6]. In future work, we plan to elevate performance by introducing a feature selection mechanism and investigating other similarity metrics.

References

1. Kanski, J.J.: Clinical Ophthalmology – A systematic Approach. Elsevier Butterworth-Heinemann, Edinburgh (2007)
2. Thylefors, B., Chylack Jr., L.T., Konyama, K., Sasaki, K., Sperduto, R., Taylor, H.R., West, S.: A simplified cataract grading system. Ophthalmic Epidemiol. **9**(2), 83–95 (2002)
3. Chylack, L., Wolfe, J., Singer, D., Leske, M.C., Bullimore, M.A., Bailey, I.L., Friend, J., McCarthy, D., Wu, S.Y.: The lens opacities classificatin system III. Arch Ophthalmol. **111**(6), 831–836 (1993)

4. Klein, B., Klein, R., Linton, K., Magli, Y., Neider, M.: Assessment of cataracts from photographs in the beaver dam eye study. Ophthalmology **97**, 1428–1433 (1990)
5. Xu, Y., Gao, X., Lin, S., Wong, D.W.K., Liu, J., Xu, D., Cheng, C.Y., Cheung, C.Y., Wong, T.Y.: Automatic grading of nuclear cataracts from slit-lamp lens images using group sparsity regression. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 468–475. Springer, Heidelberg (2013)
6. Li, H., Lim, J.H., Liu, J., Mitchell, P., Tan, A., Wang, J., Wong, T.: A computer-aided diagnosis system of nuclear cataract. IEEE Trans. Biomed. Eng. **57**, 1690–1698 (2010)
7. 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)
8. Huang, W., Li, H., Chan, K.L., Lim, J.H., Liu, J., Wong, T.Y.: A computer-aided diagnosis system of nuclear cataract via ranking. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 803–810. Springer, Heidelberg (2009)
9. Duncan, D.D., Shukla, O.B., West, S.K., Schein, O.D.: New objective classification system for nuclear opacification. J. Opt. Soc. Am. **14**, 1197–1204 (1997)
10. Khu, P.M., Kashiwagi, T.: Quantitating nuclear opacification in color scheimpflug photographs. Invest. Ophthalmol. Vis. Sci. **34**, 130–136 (1993)
11. Xu, D., Huang, Y., Zeng, Z., Xu, X.: Human gait recognition using patch distribution feature and locality-constrained group sparse representation. IEEE Trans. Image Process. **21**(1), 316–326 (2012)