Automatic Grading of Nuclear Cataracts from Slit-Lamp Lens Images Using Group Sparsity Regression

Yanwu Xu[1,⋆], Xinting Gao[1,⋆], Stephen Lin[2], Damon Wing Kee Wong[1], Jiang Liu[1], Dong Xu[3], Ching-Yu Cheng[4], Carol Y. Cheung[4], and Tien Yin Wong[4]

[1] Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[2] Microsoft Research Asia, P.R. China
[3] School of Computer Engineering, Nanyang Technological University, Singapore
[4] Singapore Eye Research Institute, Singapore

Abstract. Cataracts, which result from lens opacification, are the leading cause of blindness worldwide. Current methods for determining the severity of cataracts are based on manual assessments that may be weakened by subjectivity. In this work, we propose a system to automatically grade the severity of nuclear cataracts from slit-lamp images. We introduce a new feature for cataract grading together with a group sparsity-based constraint for linear regression, which performs feature selection, parameter selection and regression model training simultaneously. In experiments on a large database of 5378 images, our system outperforms the state-of-the-art by yielding with respect to clinical grading a mean absolute error (ε) of 0.336, a 69.0% exact integral agreement ratio (R_0), a 85.2% decimal grading error ≤ 0.5 ($R_{e0.5}$), and a 98.9% decimal grading error ≤ 1.0 ($R_{e1.0}$). Through a more objective grading of cataracts using our proposed system, there is potential for better clinical management of the disease.

1 Introduction

Cataracts are the leading cause of visual impairment worldwide, accounting for more than 50% of blindness in developing countries. Most cataracts are age-related, though they have also been attributed to disease, trauma and congenital factors. With the global trend of aging populations, the prevalence of cataracts is expected to increase. By 2020, number of blind people is projected to reach 75 million [1].

In cataracts, the normally clear crystalline lens develops opacities which result in reduced transmission of light to the retina. There are three main types of cataracts which are defined by their location and clinical appearance: nuclear, cortical and posterior subcapsular cataracts [2]. Of these, nuclear cataracts are the most common type and will be the focus of this work. With progression, nuclear cataracts may result in the loss of vision and color discrimination, eventually leading to blindness.

Currently, cataracts are diagnosed by ophthalmologists directly using a slit-lamp microscope, or graded by clinicians who assess the presence and severity of the cataract by comparing against a set of standard reference photographs. These photographs are provided with cataract grading protocols such as the Lens Opacities Classification System III (LOCS III) [3] and the Wisconsin cataract grading system [4]. Fig. 1 shows the

⋆ Indicates equal contributions.

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 468–475, 2013.

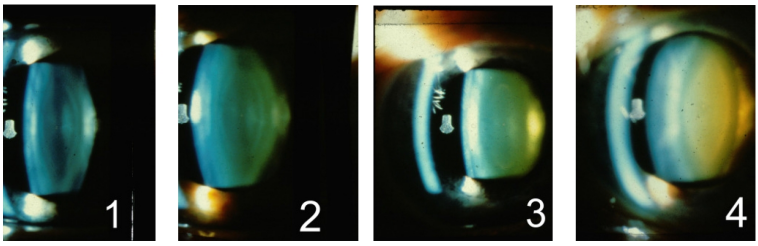

Fig. 1. Standard photographs of the Wisconsin grading system

standard photographs of the Wisconsin protocol. However, such manual assessments can be subjective, time-consuming and costly [4]. Accurate, automated grading of the presence and severity of cataracts would help to improve clinical management of the disease, as well as provide an objective basis for epidemiological studies [5].

To increase the objectivity and efficiency of cataract grading, some computer-aided systems have been proposed to grade nuclear cataracts. In the state-of-the-art method of [6], bottom-up detection and top-down modeling are combined to detect the region of interest (ROI) and lens structure robustly, with a detection rate of 95% on a population study database [7,8] of 5850 slit-lamp lens images. Within the detected lens and structure of the nucleus, twenty-one features describing the lens and nuclear regions are extracted to model nuclear cataracts. Support Vector Regression (SVR) is then applied to automatically determine the cataract grade. In comparison to the method of the Wisconsin group [9], the technique of [6] yields an improvement of 33.6% (from 0.541 to 0.359) in average grading difference with respect to ground truth.

In this work, we propose a new approach for the automatic grading of nuclear cataracts from slit-lamp images. Our contributions include the introduction of a new feature for nuclear cataract grading, and the use of a corresponding group sparsity regression (GSR) to perform feature selection, parameter selection and regression model training simultaneously. Our proposed system is able to achieve higher overall performance than previous work, and has the potential to be applied to other eye diseases.

2 Automatic Grading System for Nuclear Cataracts

Our automatic grading system for nuclear cataracts is formulated as linear regression with a group sparsity constraint. The system consists of three components: ROI and structure detection, feature extraction, and prediction.

2.1 Feature Extraction

Lens Structure Detection. Current methods for lens structure based feature detection [6,9,10,11,12] are highly effective, and in this work we employ the technique used in [6] for this purpose. With structure detection, each lens is separated into three sections: nucleus, anterior cortex, and posterior cortex. After obtaining the lens structure of each image, the central part of the lens along the visual axis is extracted and resized to 128×512. Features are extracted from each of the resized sections.

Bag-of-features Extraction. The bag-of-features (BOF) model, also known as the bag-of-words model [13], is a simplifying representation used in natural language processing to model a text by a sparse vector or histogram of word occurrences over a vocabulary. This idea has been adopted in computer vision to model an image as a sparse vector of occurrence counts over a vocabulary of local image features (codebook). The BOF model provides a location-independent global representation of local features in which properties such as intensity, rotation, scale or affine invariance can be preserved.

In this work, the local features in our BOF model are image patches that represent intensity and texture information. Each section of the resized lens image is divided into a grid of half-overlapping $s \times s$ patches each represented as an s^2-dimensional vector. After obtaining all the local patches from a set of training images, k-means clustering is used to generate the codebook from randomly selected samples, and then the BOF (*i.e.*, occurrence counts of the visual words) is obtained in a binning procedure. Since each BOF is a histogram, the clustering parameter k is referred to as the bin number. Readers may refer to [13] for more details on BOF extraction.

Image Feature Representation. Finally, for each slit-lamp image, we obtain its image feature representation $\mathbf{f}_i$ by concatenating the BOFs extracted in its three sections $S = \{S_a, S_n, S_p\}$ (*i.e.*, anterior cortex, nucleus, and posterior cortex), computed for each of six color channels $C = \{C_h, C_s, C_v, C_r, C_g, C_b\}$ (*i.e.*, HSV and RGB color channels), with a fixed patch size $s = 8$ and various bin numbers $K = \{100, 200, 400\}$. This leads to a feature dimension of $|\mathbf{f}_i| = |S| \times |C| \times \sum_{n=1}^{|K|} K_n = 18 \sum_{n=1}^{|K|} K_n = 12600$.

We refer to each BOF extracted for a given section of the lens, color channel, patch size and bin number as a *group feature*. L_1-normalization is performed such that the sum of each group feature is equal to 1, and a truncation similar to that used in SIFT feature extraction [14] is applied to reduce feature bias and noise, *i.e.*, if a bin is greater than 0.2, it is set to 0.2 and the L_1-norm is recomputed.

With this image feature representation, we wish to train a regression model for the nuclear cataract grading task; however, the large size and redundancy of this model would make training and testing inefficient. A reduced representation could potentially be used, but it is unclear which color channels are most informative for each section of the lens, and how many bins is optimal for a given channel. To address this problem, we apply a group sparsity constraint in the regression to select an effective subset of the extracted features for nuclear cataract grading.

2.2 Feature Selection and Grading Using Group Sparsity Regression

Identifying and using only the effective elements of the image feature representation can bring higher precision and speed. For a training sample with an image feature representation $\mathbf{f}_i$ consisting of g feature groups, we denote its regression value (*i.e.*, the clinician grading) as $y_i \in (0, 5]$. We adopt the linear regression model $\omega^T \mathbf{f}_i + \mu$ to predict the grading value, where ω is the weighting vector and μ denotes the bias. We minimize the following objective function:

$$\min_{\omega,\mu} \sum_{i=1}^{n} \|y_i - \omega^T \mathbf{f}_i - \mu\|^2 + \lambda \sum_{j=1}^{g} \|\omega_j\|_2, \tag{1}$$

where ω_j is the corresponding weight of the j^{th} feature group, n is the number of training samples, g is the number of groups and λ is used to control the sparsity of ω. In Eq. (1), the first term represents the regression error and the second term is an $L_{2,1}$-norm based regularizer to enforce group sparsity. As the features are intrinsically organized into groups, the $L_{2,1}$-norm based regularizer essentially selects features from only a sparse set of groups. In our experiments, we use the SLEP toolbox [15] to optimize Eq. (1).

The solution for ω indicates which group features are included in the final feature $\mathbf{x}_i$, *i.e.*, the j^{th} group of features is selected when $\|\omega_j\|_2 > 0$. The lower dimension of the final feature $\mathbf{x}_i$ leads to faster feature extraction and grading in the testing phase when compared to using the higher-dimensional image feature representation.

3 Experiments

In this section, we first compare our method with the state-of-the-art nuclear cataract grading method [6], and then evaluate the effectiveness and robustness of the group sparsity regression by comparing to other related regression methods while using the same feature.

All the experiments are performed on the large ACHIKO-NC dataset [6], comprised of 5378 images with decimal grading scores that range from 0.1 to 5.0. The scores are determined by professional graders based on the Wisconsin protocol [4], with higher decimal scores indicating greater severity of the cataract, *e.g.*, a 3.1 is judged to be a bit more severe than that of standard 3 in Fig. 1. 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 3.1 has an integral grading score 4. 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. Since the unbalanced data distribution of ACHIKO-NC may skew a learned prediction model towards middle grade estimates, we set the training sample size of each grade to 20 as done in [6].

3.1 Evaluation Criteria

In this work, we use the same four evaluation criteria as in [6] to measure grading accuracy, namely the exact integral agreement ratio (R_0), the ratio of decimal grading errors ≤ 0.5 ($R_{e0.5}$), the ratio of decimal grading errors ≤ 1.0 ($R_{e1.0}$), and the mean absolute error (ε), which are defined as

$$\begin{aligned} R_0 &= \frac{|\lceil G_{gt} \rceil = \lceil G_{pr} \rceil|_0}{N}, & R_{e0.5} &= \frac{\left||G_{gt} - G_{pr}| \leq 0.5\right|_0}{N}, \\ R_{e1.0} &= \frac{\left||G_{gt} - G_{pr}| \leq 1.0\right|_0}{N}, & \varepsilon &= \frac{\sum |G_{gt} - G_{pr}|}{N}, \end{aligned} \tag{2}$$

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.

3.2 Comparison to the State-of-the-art [6] and Professional Grading

We first compare our method to the state-of-the-art [6] using the same dataset, experimental setting and reporting methods. Results are listed in Table 1, where the performance of [6] are values reported in their paper. Our method is shown to surpass [6] in all four evaluation criteria. We additionally performed a comparison of their features to our BOF group features by employing our BOF group features with their RBF ϵ-SVR regression. Table 1 shows an improvement with our features on $R_{e0.5}$ and $R_{e1.0}$, but some decrease in performance on R_0 and ε. It also indicates that our group sparsity regression outperforms their RBF ϵ-SVR regression when BOF group features are used.

More in-depth comparisons, such as multiple repeated tests with random training samples, could not be conducted since the features of [6] are not described in enough detail for accurate reproduction. Although we cannot conclude from these experiments that our BOF group features are superior to those of [6], it is interesting to note that their features could potentially be added to the BOF features in our method and processed with our group sparsity regression.

Our method has important advantages over [6]. It requires less detailed segmentation of the lens, and our BOF features are less sensitive to segmentation accuracy. By contrast, some features in [6] rely on accurate further segmentation of the nucleus into three specific parts, which in practice can be difficult to achieve. Another advantage of our method is its ability to identify and use discriminative features from a broad feature set through group sparsity regression, in contrast to [6] which performs regression on a set of predefined features. In this way, our method can consider a variety of potentially useful features without introducing noise if the features are found to be redundant or less effective. The performance improvements of our method can largely be attributed to these differences.

A plot of our automatic grading against that of a professional grader is exhibited in Fig. 2, which demonstrates that our method is in close agreement with the grader. The selected discriminant feature groups are illustrated in Fig. 3, which indicates that:

1. $K = 400$ bins is not necessary for this problem and may introduce noise, so it was not selected for any groups.
2. Features extracted from the anterior cortex have no discriminative power and can be ignored in testing. This suggests that features in [6] that are dependent on the anterior cortex may be less discriminative and could introduce noise.
3. Features extracted from the posterior cortex and nucleus have discriminative power, which is consistent with protocol grading criteria. In the protocols [3,4], nuclear cataracts are to be graded based on the intensity and visibility of the nuclear landmarks, and the color of the nucleus and posterior cortex.

On a four-core $2.4GHz$ PC with $24GB$ RAM, our method takes 20.45s on average to process an image, with 4.23s for feature extraction and 10^{-5}s for prediction. This processing speed slightly exceeds the 25.00s per image of [6], which takes 8.76s for feature extraction and 0.02s for prediction.

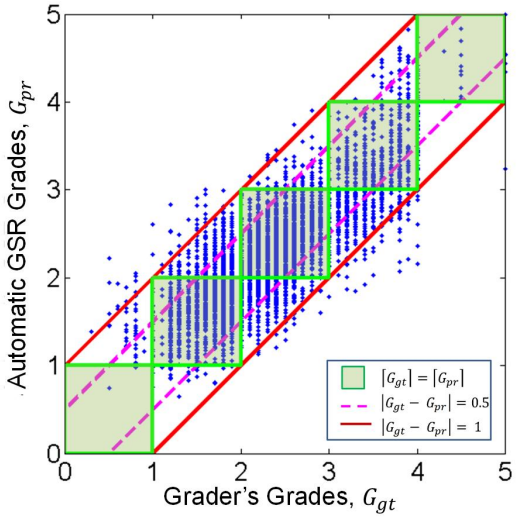

Fig. 2. Visualization of our automatic grading vs. that of a professional grader

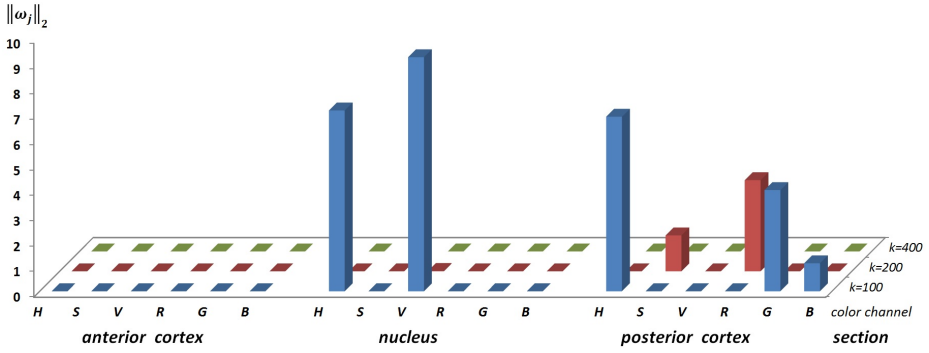

Fig. 3. Visualization of the selected feature groups

Table 1. Performance comparisons for nuclear cataract grading methods

Method	R_0	$R_{e0.5}$	$R_{e1.0}$	ε
Proposed	**0.690**	**0.852**	**0.989**	**0.336**
BOF+RBF ϵ-SVR	0.604	0.789	0.978	0.388
RBF ϵ-SVR [6]	0.654	0.778	0.975	0.355
Our improvement over [6]	5.5%	9.5%	1.4%	-5.4%

3.3 Comparison to other Regression Methods

We also compare our group sparsity regression method to linear SVR [16] and RBF kernel based ϵ-SVR [6], using our BOF group features, to verify that the group sparsity

Table 2. Grading performance comparisons with different regression methods

Method	R_0	$R_{e0.5}$	$R_{e1.0}$	ε
Proposed	**0.682±0.004**	**0.834±0.005**	**0.985±0.001**	**0.351±0.004**
Linear SVR[16]	0.667±0.010	0.815±0.015	0.980±0.004	0.363±0.011
RBF ϵ-SVR [6]	0.615±0.013	0.799±0.012	0.980±0.002	0.375±0.011

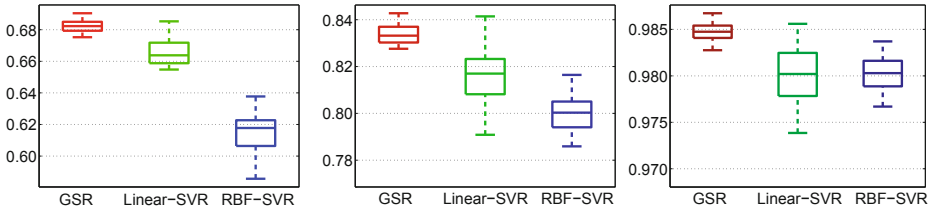

Fig. 4. Performance comparison in terms of R_0, $R_{e0.5}$ and $R_{e1.0}$ (from left to right)

constraint reduces feature noise and thus increases accuracy. Testing is conducted over twenty rounds. To examine generalization ability (scalability), we followed the training/testing sample ratio in [6], *i.e.*, in each round, 100 training samples were randomly selected from all the 5378 images, with 20 images for each grade, and the remaining 5278 images were used for testing. In training, optimal parameters were selected for each method by cross-validation, where half of the images (50 images with 10 per grade) were used to train a regression model, the other half used for testing, and the set of parameters with the highest average accuracy was chosen. All 100 images were then used to train the new model with the optimal parameters, and the result of each round is obtained by testing the remaining 5278 images using this new model. The performance of the three regression methods is shown in Fig. 4 and also listed in Table 2 in terms of mean value and standard deviation of R_0, $R_{e0.5}$, $R_{e1.0}$ and ε over the twenty rounds. From these results, the following observations can be made:

1. Comparing the proposed method to linear SVR shows that the group sparsity constraint is helpful to reduce feature noise and thus improve performance.
2. Comparing linear SVR to RBF ϵ-SVR shows that a linear kernel is better than an RBF kernel for the proposed high dimensional feature, which can be expected since RBF is more suitable for low dimensional features. In addition, RBF ϵ-SVR is much slower in both training and testing, while the other two methods are more efficient.

4 Conclusions

For nuclear cataract grading from slit-lamp lens images, we have proposed a regression-based framework with BOF group features and a group sparsity constraint for joint feature selection, parameter selection and regression model training. In tests on the *ACHIKO-NC* dataset comprised of 5378 images, our system achieves a 69.0% exact

agreement ratio (R_0) against clinical integral grading, a 85.2% decimal grading error $\leq$ 0.5 ($R_{e0.5}$) and a 98.9% decimal grading error $\leq$ 1.0 ($R_{e1.0}$), which represents significant improvements over the state-of-the-art method [6]. In future work, we plan to elevate performance by using new features or by introducing other domain-specific knowledge on this problem.

References

1. IAPB Report - State of the World Sight (2010), http://www.iapb.org/resource/iapb-report-state-world-sight-2010
2. Asbell, P.A., Dualan, I., Mindel, J., Brocks, D., Ahmad, M., Epstein, S.: Age-Related Cataract. The Lancet 365(9459), 599–609 (2005)
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 Ophthalmology 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. Thylefors, B., Chylack Jr., L.T., Konyamia, K., Sasaki, K., Sperduto, R., Taylor, H.R., West, S.: A Simplied Cataract Grading System – The WHO Cataract Grading Group. Ophthalmic Epidemiology 9(2), 83–95 (2002)
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. on Biomed. Eng. 57, 1690–1698 (2010)
7. Foong, A., Saw, S., Loo, J., Shen, S., Loon, S., Rosman, M.: Rationale and Methodology for a Population-based Study of Eye Diseases in Malay People: The Singapore Malay Eye Study (SiMES). Ophthalmic Epidemiology 14, 25–35 (2007)
8. Tan, A.C.S., Wang, J.J., Lamoureux, E.L., Wong, W., Mitchell, P., Li, J., Tan, A.G., Wong, T.Y.: Cataract Prevalence Varies Substantially with Assessment Systems: Comparison of Clinical and Photographic Grading in a Population-Based Study. Ophthalmic Epidemiology 18(4), 164–170 (2011)
9. 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)
10. 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)
11. Duncan, D.D., Shukla, O.B., West, S.K., Schein, O.D.: New Objective Classification System for Nuclear Opacification. Journal of Optical Society of America 14, 1197–1204 (1997)
12. Khu, P.M., Kashiwagi, T.: Quantitating Nuclear Opacification in Color Scheimpflug Photographs. Invest. Ophthalmol. Vis. Sci. 34, 130–136 (1993)
13. Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: CVPR, vol. 2, pp. 524–531 (2005)
14. Lowe, D.G.: Distinctive Image Features from Scale-invariant Keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
15. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Arizona State University (2009), http://www.public.asu.edu/~jye02/Software/SLEP
16. Chang, C., Lin, C.: LIBSVM: A Library for Support Vector Machines. ACM Trans. on Intel. Sys. and Tech. 2(3), 27:1–27:27 (2011)