Automated Cataracts Screening from Slit-Lamp Images Employing Deep Learning

Zhipeng Zhang[1(✉)], Wenhui Shou[1], Dongjia Xing[1], Wenting Ma[1], Qingqing Xu[1], Wei Wang[1], Li-Qun Xu[1], Ziming Liu[2], and Ling Xu[2]

[1] China Mobile Research Institute, Beijing 100032, China
zzp_zzp2002@aliyun.com, shouwenhui@chinamobile.com
[2] Shenyang He Eye Hospital, Shenyang 110034, China

Abstract. To assess the feasibility and performance using deep learning networks to automatically detect cataracts from slit-lamp images in large-scale eye diseases screening scenarios. Two datasets were collected using, respectively, the professional Slit-Lamp Microscopes (SLM) and the portable Slit-Lamp Devices (SLD) clipped on a Smartphone, during routine eye disease screening programs in China. The former Dataset-M comprised 4891 images from 1670 subjects and the latter Dataset-D comprised 2516 images from 802 subjects. Each image was then labelled by three ophthalmologists as one of the three classes: 1) un-gradable image, 2) cataract, and 3) normal. For each dataset, two deep learning models were created: one for image quality assessment, and the other for cataracts detection, and the performance of which was assessed by the Area Under a ROC Curve (AUC) and kappa agreement. For the quality assessment models, on Dataset-M (Dataset-D), the corresponding AUC achieved were 0.929 (0.881), with kappa agreements of 0.628 (0.590) and $p < 0.001$, respectively. For the cataract detection models, the corresponding AUC were 0.997 (0.987), with kappa agreements of 0.912 (0.893) and $p < 0.001$, respectively. Furthermore, based on these models we built a practical cloud application that has been trialled in 25 real-world screening settings in China, receiving favourable feedbacks from clinicians, primary care physicians and patients alike.

Keywords: Deep learning · Mobile healthcare application · Automated cataracts screening

1 Introduction

Cataract is the second leading cause of visual impairment and the first leading cause of blindness [1]. According to the Chinese Ophthalmology Society, the prevalence of cataract was 80% among people aged between 60 and 89, and it even reached 90% among people aged over 90. Early screening plays a key role in controlling the disease progression and preventing needless cases of blindness. However, the provision of specialists eye care services are unevenly distributed in China. Primary care physicians at county and lower level hospitals and clinics are generally lack of expensive equipments and necessary expertise in screening and diagnosing common eye diseases,

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 282–291, 2021.
https://doi.org/10.1007/978-3-030-67514-1_23

resulting in a large number of patients, especially in resource-limited rural and remote regions, having little or no access to high-quality eye care services.

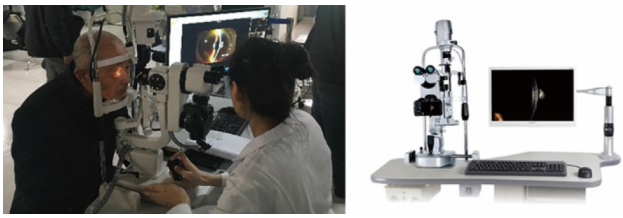
In recent years, with the cost going down and the use of ophthalmic imaging devices becoming widespread, a large volume of eye images can be available. On the other hand, machine learning technology such as deep learning is advancing very rapidly, which proved to be amenable to various medical image analysis tasks with superior performance [2–5]. Furthermore, the ubiquitous mobile broadband network and cloud computing platform provide unprecedented data acquisition, transmission, storage and processing capabilities. These driving forces make it possible to automatically detect cataracts from slit-lamp images in large-scale eye diseases screening scenarios. Compared to using traditional feature extraction and grading methods to detect cataracts [6], deep learning and convolutional neural network (CNN) methods can be more robust to noises and interferences, and are generally applicable in practical scenarios, which has been shown by Long et al. [7], in the case of diagnosing rare congenital cataracts based on slit-lamp images with diffuse light.

In this paper, an automated system for cataracts screening from slit-lamp images was investigated, which suited for both outpatient and ambulatory scenarios. In outpatient screening scenarios, such as the county and township hospitals, cities' community health service stations and optical vision centres, and even some large rural clinics in China, the professional Slit-Lamp Microscopes (SLM) with single lens reflex (SLR) cameras are available for a technician to take one or more slit-lamp images for each eye of the patient. The images were then uploaded in real-time through 4G-LTE or Wi-Fi hotspots to our cloud platform for detailed analysis using advanced deep learning algorithms in respect of images' quality and cataract presence. On the other hand, in ambulatory screening scenarios such as in remote and rural regions, village doctors, care assistants or trained volunteers were equipped with a cheap option, a portable Slit-Lamp Device (SLD) clipped on a Smartphone, to capture slit-lamp images of the patients anywhere and anytime, and the images were transmitted using the 4G-LTE cellular network to the cloud platform. By taking advantage of this eye diseases screening system, urban and rural communities currently under-served can have equitable access to professional and quality primary eye care services, receiving a fast feedback, diagnosis and/or referral, helping to take a preventive measure or early treatment action as cataracts develop. This would improve quality of life of the patients, save unnecessary medical expenses and get them back to work, thus, relieving illness-related poverty. The feasibility and effective performance of the algorithms and system were investigated and verified.

2 Methods

The slit-lamp images that we used to train, validate and test the methods were acquired during the eye disease screening programs conducted by the He Eye Hospital Group (HEHG) in China between December 2015 and December 2017. Figure 1 shows, respectively, the two typical settings and devices used in the case of outpatients and ambulatory screening scenarios. These were eye hospitals, optical vision centres,

community halls and village clinics in Cities of Shenyang, Dalian, Jinzhou, Yingkou, Huludao, and so on across China Northeast Liaoning Province.



(a)

(b)

Fig. 1. Eye disease screening scenarios and typical slit-lamp devices (a) Outpatient screening scenario and SLM with SLR camera (b) Ambulatory screening scenario and SLD

The images were collected using either professional SLM-KD4, a SLM with SLR camera (Made by Chongqing Ruiyu Instruments and Equipment Co. Ltd, Chongqing, China), or portable SLD prototypes (Built by the HEHG, Shenyang, China) clipped on a Smartphone. Up to six images were captured for each eye by an operator through shifting the focus positions to the front (diffuse light) and side (slit light) of the anterior lens capsule as well as the posterior lens capsule (slit light). Depending on different ophthalmic screening programs, the acquisition of images could have taken place in a darkroom, half-darkroom, or indoor room, all without pupil dilation.

Given that the images taken by the two different types of devices and cameras were largely different in resolution, field of view, brightness, contrast, and other interference factors (Fig. 2), two independent datasets, were created as shown in Table 1. Dataset-M contained 4 891 SLM photographed images and Dataset-D contained 2 516 SLD photographed images.

Each image in the above two datasets was labelled independently, by three ophthalmologists each having at least six years of clinical experience in the HEHG, as one

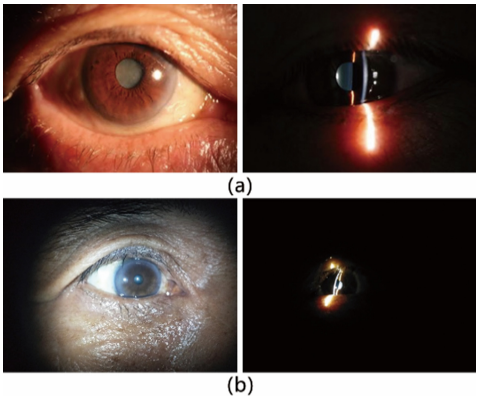


Fig. 2. Examples of slit-lamp images (a) SLM photographed images (b) SLD photographed images

Table 1. Summary of image characteristics and available demographic information in the two datasets

Characteristics	Dataset-M	Dataset-D
No. of images	4 891	2 516
Patient demographics		
No. of individuals	1 670	802
No. of female individuals	1 034	509
Age, mean (SD), y	68.0 (10.7)	62.8 (11.6)
Label distribution		
Label 1	1 216	855
Label 2	1 326	1 246
Label 3	2 349	415

of the three classes: 1) un-gradable image, 2) cataract, and 3) normal. Figure 3 shows a few common examples of un-gradable images, denoted as 'Label 1': no eyeball, no lens, defocused lens, excessive width of slit light, bright ambience, slit light with less than 30-degree angle, and reflective pupil area. Examples of the gradable images, marked as either 'Label 2' (cataracts with lens opacity) or 'Label 3' (normal with clear lens), are shown in Fig. 4. The three ophthalmologists had to discuss to reach a consensus when the labels were in disagreement. After labelling, the distributions of image labels among the two datasets were shown in Table 1.

Deep learning was used to create two models for each dataset, one for image quality assessment, and the other for cataracts detection. In the image quality assessment task, all images were empirically resized to a predefined 650 $\times$ 500 pixels, which was based on prior experiments on suitable image sizes; The pixel values were normalized to between 0 and 1. Then, the images with Label 1 were considered as positive samples and images with Label 2 and 3 were negative samples. These images were randomly

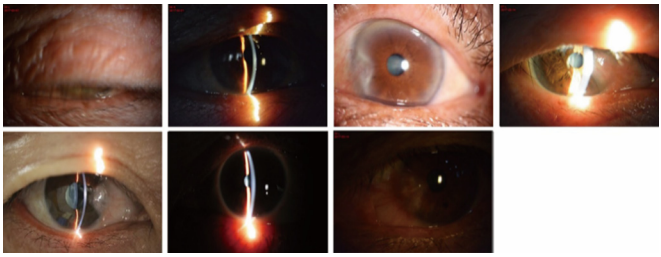


Fig. 3. Examples of un-gradable slit-lamp images

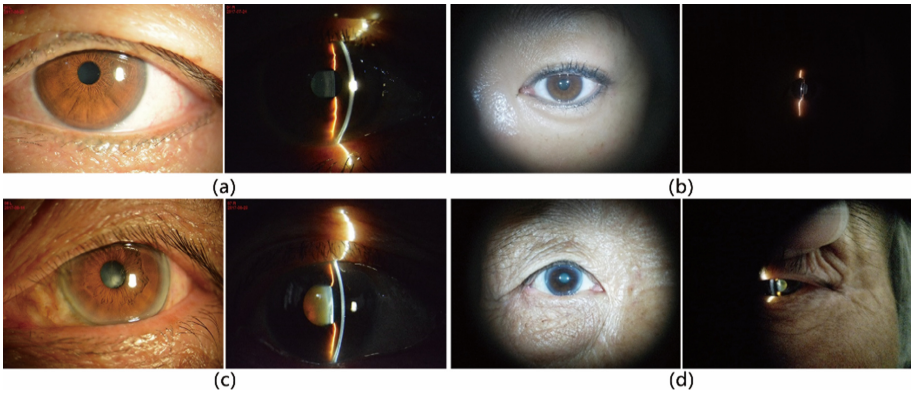


Fig. 4. Examples of normal and cataract (a) Normal – SLM photographed images (b) Normal – SLD photographed images (c) Cataract – SLM photographed images (d) Cataract – SLD photographed images

divided into three parts proportionally according to 8:1:1 as shown in Table 2: a) For training, 80% of the data was used to train the Inception-v3 network [8] (proved to be superior to other popular networks in our experiments) pre-trained on ImageNet dataset [9]; b) For validation, 10% of the data was used to determine the appropriate number of training epochs to avoid the model's over-fitting; and c) For testing, the remaining 10% of the data was used to assess the performance of the final model to unseen data. In the training phase, two separate models for the Dataset-M and Dataset-D were fine-tuned and validated, respectively, incorporating the standard practice of data augmentation. To counter the effect of the data imbalance in positive and negative samples, weighting factors were introduced in the loss function based on binary cross entropy to equally penalize under- or over-represented classes in the training set.

In the cataract detection task, the main processing flow was similar to that of the previous one, including data pre-processing, dataset partition, model training and testing, though the images with Label 2 were considered as positive samples and images with Label 3 negative samples, as shown in Table 3. Considering that the total

number of negative samples in Dataset-D was limited (the last row of Table 3), to ensure that the number of images for testing is adequate, we repartitioned the samples to increase the number of testing images.

Table 2. Dataset partition for developing the models for image quality assessment (Positive samples – Label 1; Negative samples – Label 2 & Label 3)

Dataset	Label	Training	Validation	Test	Total
Dataset-M	Positive	974	121	121	1 216
	Negative	2 943	366	366	3 675
Dataset-D	Positive	685	85	85	855
	Negative	1 329	166	166	1 661

Table 3. Dataset partition for developing the models for cataracts detection (Positive samples – Label 2; Negative samples – Label 3)

Dataset	Label	Training	Validation	Test	Total
Dataset-M	Positive	1 881	234	234	2 349
	Negative	1 062	132	132	1 326
Dataset-D	Positive	989	128	129	1 246
	Negative	277	38	100	415

3 Statistical Analysis

The performance of the models on the two test datasets using the consensus of three ophthalmologists as the reference standard was shown in Table 4. For the two image quality assessment models, the AUC were 0.929 (95% CI 0.904–0.951) on Dataset-M and 0.881 (95% CI 0.833–0.925) on Dataset-D, with kappa agreements of 0.628 ($p < 0.001$) and 0.590 ($p < 0.001$), respectively. This shows very strong classification performance of the two image quality models, while the kappa agreements around 0.6 mean a substantial consistency between the test results and the consensus of three ophthalmologists. For the two cataract detection models corresponding to the two datasets, the AUC were 0.997 (95% CI 0.993–0.999) and 0.987 (95% CI 0.975–0.996), with kappa agreements of 0.912 ($p < 0.001$) and 0.893 ($p < 0.001$), respectively. Thus, the classification performance of the two cataract detection models were superb and the consistency proved to be almost perfect.

Table 5 and Fig. 5 compare the performance of the models against that of three individual ophthalmologists, i.e. AJ - junior ophthalmologist, BS - senior ophthalmologist and CS - senior ophthalmologist, for the two datasets, respectively. Except that one of the three ophthalmologists (CS) had a better performance than the model in the image quality assessment task on the Dataset-M, in most cases, the accuracy and kappa agreement of models were superior to ophthalmologists. For quality assessment task, the models can be most valuable to reduce the rate of undetected poor quality

Table 4. The performance indicators (sensitivity, specificity, accuracy, AUC, kappa agreement) of the models for image quality assessment (IQA) and cataracts detection (CPD), using the consensus of three ophthalmologists as the reference standard

Tasks	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Accuracy, % (95% CI)	AUC (95% CI)	Kappa (95% CI)	P-value
IQA on Dataset-M	83.5 (76.5, 89.6)	85.2 (81.5, 89.0)	84.8 (81.5, 88.1)	0.929 (0.904, 0.951)	0.628 (0.548, 0.703)	<0.001
IQA on Dataset-D	80.0 (71.3, 88.2)	81.3 (75.2, 87.2)	80.9 (75.7, 85.7)	0.881 (0.833, 0.925)	0.590 (0.485, 0.689)	<0.001
CPD on Dataset-M	96.2 (93.5, 98.3)	95.5 (91.7, 98.6)	95.9 (93.7, 97.8)	0.997 (0.993, 0.999)	0.912 (0.860, 0.952)	<0.001
CPD on Dataset-D	95.3 (91.3, 98.5)	94.0 (89.0, 98.1)	94.8 (91.7, 97.4)	0.987 (0.975, 0.996)	0.893 (0.832, 0.947)	<0.001

Table 5. The performance of the models vs. that of three individual ophthalmologists for image quality assessment and cataracts detection tasks (AJ – junior Ophthalmologist, BS – senior Ophthalmologist, CS – senior ophthalmologist)

Indicators		Image quality assessment		Cataracts detection	
		Dataset-M	Dataset-D	Dataset-M	Dataset-D
Sensitivity, %	Model	83.5	80.0	96.2	95.3
	AJ	39.7	38.8	40.2	81.4
	BS	43.8	38.8	88.9	83.7
	CS	66.1	41.2	26.1	82.9
Specificity, %	Model	85.2	81.3	95.5	94.0
	AJ	95.1	95.8	98.5	56.0
	BS	92.6	98.2	68.9	63.0
	CS	94.3	98.2	96.2	42.0
Accuracy, %	Model	84.8	80.9	95.9	94.8
	AJ	81.3	76.5	61.2	70.3
	BS	80.5	78.1	81.7	74.7
	CS	87.3	78.9	51.4	65.1
Kappa	Model	0.628	0.590	0.912	0.893
	AJ	0.410	0.397	0.315	0.383
	BS	0.411	0.431	0.593	0.476
	CS	0.639	0.455	0.174	0.260

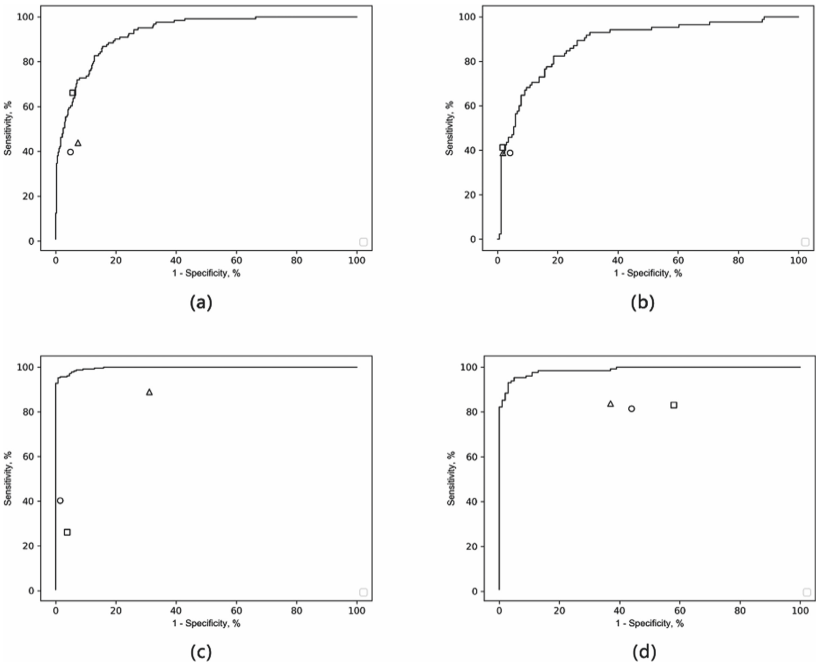


Fig. 5. The performance of the models (ROC curve) and individual ophthalmologists (AJ – circle, BS – triangle, CS – square, as indicated) for image quality assessment and cataracts detection tasks when tested on the two datasets. (a) Image quality assessment on Dataset-M (b) Image quality assessment on Dataset-D (c) Cataracts detection on Dataset-M (d) Cataracts detection on Dataset-D

images in both Dataset-D and Dataset-M. For cataracts detection task, more cataracts cases in both Dataset-D and Dataset-M could be identified accurately by the model than individual ophthalmologists. Overall, the models had achieved high performance in analyzing diverse slit-lamp images obtained in primary care settings, and could be reliably applied to real-world screening.

The deep learning algorithms were integrated with an end-to-end system for large-scale outpatient and ambulatory cataracts screening scenarios. The system consists of a SLM or SLD, a mobile APP running on the android smartphones or tablets, and a cloud-based platform integrating the corresponding models for both Dataset-D and Dataset-M. Slit-lamp images and demographic information such as subjects' gender, age, and personal medical history were acquired and uploaded from the APP to the platform through 4G-LTE or Wi-fi network. By taking advantages of suitable technologies like Graphics Processing Unit (GPU) server cluster and Redis message queue, the system was able to meet the real-world requirements of high concurrency and low latency. Tested on one single GPU server (Alibaba Cloud Computing Co. Ltd., Hangzhou, China), it took 0.3 s to process and analyse a total of six images from one subject. A cluster of 8 GPU servers could support up to 2000 screening settings with a response time of 10 s. The system has been trialled in 25 real-world screening settings

in 9 cities of Liaoning Province, China, receiving favourable feedbacks from clinicians, primary care physicians and patients alike.

4 Discussion

To our knowledge, this is the first time that an automated cataracts screening system from slit-lamp images employing deep learning has been developed and validated for large-scale outpatient and ambulatory screening scenarios. Compared to other relevant work applying machine learning techniques to detecting cataracts such as nuclear cataracts [6] and rare congenital cataracts [7] based on slit-lamp images, the image quality assessment and cataracts screening solution proposed in this study has the following characteristics and advantages:

(1) For the benefit of the general public, the target eye disease is age-related cataracts with high prevalence and incidence rate, including nuclear cataracts, cortical cataracts, and posterior subcapsular cataracts.
(2) Besides supporting outpatient ophthalmic screening in settings such as hospitals and eye clinics, the portable solution involving SLD enables large-scale ambulatory screening in rural areas and sparsely populated remote regions, given a widespread cellular coverage available, thus bringing specialist eye care service to the most venerable communities.
(3) Although the images were taken by operators with different knowledge levels using multiple imaging equipment (SLM or SLD) during various acceptable ambient light conditions, the performance of the cataracts detection models tested on gradable images remains unaffected by using sufficient image data to fine-tune deep learning CNNs and data augmentation methods.
(4) Under the close collaboration between scientific researchers and ophthalmologists, the standards of anterior ocular segment image acquisition and quality assessment were innovatively discussed and determined.
(5) The effectiveness of the deep learning algorithms and system has been validated; hence, the system will be ultimately integrated into routine clinical processes and tackle the challenge of real world application scenario.

Further research is necessary to expand the amount and multisource of slit-lamp images to ensure the data heterogeneity, use deep learning algorithms to further grade the severity of lens opacity, and detect other common ocular surface diseases such as corneal disease and pterygium. We have conducted relevant research and the preliminary results are very promising.

Acknowledgements. We acknowledge the help of all the ophthalmologists in HEHG for collecting slit-lamp images during the eye disease screening programs, and Jun Li and Xinghuai Xue in HEHG for image labelling.

References

1. Pascolini, D., Mariotti, S.P.: Global estimates of visual impairment: 2010. Br. J. Ophthalmol. **96**, 614–618 (2012)
2. Esteva, A., et al.: Corrigendum: dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
3. Hoochang, S., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**, 1285–1298 (2016)
4. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**, 2402 (2016)
5. Zhang, L., et al.: Automatic cataract detection and grading using Deep Convolutional Neural Network. In: IEEE ICNSC 2017, pp. 60–65 (2017)
6. Huang, W., Chan, K.L., Li, H., Lim, J.H., Liu, J., Wong, T.Y.: A computer assisted method for nuclear cataract grading from slit-lamp images using ranking. IEEE Trans. Med. Imaging **30**, 94–107 (2010)
7. Long, E., et al.: An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat. Biomed. Eng. **1**, 0024 (2017)
8. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE CVPR 2016, pp. 2818–2826 (2016)
9. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: ImageNet: a large-scale hierarchical image database. In: IEEE CVPR 2009, pp. 248–255 (2009)