RETHINKING RETINAL LANDMARK LOCALIZATION AS POSE ESTIMATION: NAIVE SINGLE STACKED NETWORK FOR OPTIC DISK AND FOVEA DETECTION

Shishira R. Maiya, Puneet Mathur**

University of Maryland, College Park

ABSTRACT

Automatic detection of optic disk and fovea, the two fundamental biological landmarks of the retinal system, is crucial to track the disease progression in a diabetic patient. Recent advances in this direction were mostly limited to applying CNN based networks to aggressively extract visual geometric features. In a departure from that practice, we put forward the notion of treating the landmark detection problem in human eye scans as a pose estimation problem owing to the anatomical geometrical relationship between optic disk and fovea. In this regard, we present *Naive Single Stacked Hourglass* (NSSH) network which learns the spatial orientation and pixel intensity contrast between optic disk and fovea to accurately pinpoint their locations. NSSH network significantly reduces the mean squared loss, thus outperforming all previously known techniques and establishing a state of the art in both optic disk and fovea localization tasks.

Index Terms— Biomedical imaging, retina fundus images, key-point detection, pose estimation, hourglass networks

1. INTRODUCTION

Diabetic Retinopathy (DR) is the leading cause of visual impairment in diabetic patients. Most cases pertaining to vision loss can be detected in early stages using high resolution retinal scans to spot morphological abnormalities. However, such practices are time consuming and not scalable without the presence of a highly experienced ophthalmologist. Often, the retinal fundus images may be difficult to comprehend due to presence of hemorrhages, hard exudates and non-uniform illumination. When performing diagnostic measurements on eye fundus images, the two key anatomical structures of particular interest for specialists are - Optic Disk (OD) and Fovea. Optic disk is a bright yellowish oval region that marks the beginning of the optic nerve and entry point for major blood vessels that supply the retina. Fovea is a small depression in the center of retina which is usually the darkest region in a digital fundus image.

The fovea and the optic disk are the defining elements of retina fundus coordinate systems and are essential to characterize the spatial distribution of retinal features. The medical significance of detecting the OD and fovea is that the closer a haemorrhage is to any of them, the more likely is the person to suffer from partial or complete blindness in the near future. Recently, localization and pin-point detection of the retinal landmarks have gained much popularity owing to the proliferation in early disease prediction and treatment capabilities. This work proposes a new approach to view the standard biomedical localization as a key-point detection problem where the optic disk and fovea are seen as landmarks in static images. Consequently, inspired by advances in human pose estimation techniques in the recent past, a CNN-based *Naive Single Stack Hourglass* (NSSH) network is proposed to accomplish the task of retinal key-point detection. The experimental results are computed on IDRiD Retinopathy Image Dataset[1], which is one of the most popular retinal scans dataset, far better than its predecessors in terms of size and quality. Moreover, our methodology gives state of the art performance by significantly reducing the mean squared error in OD and fovea detection as compared to any of the existing techniques.

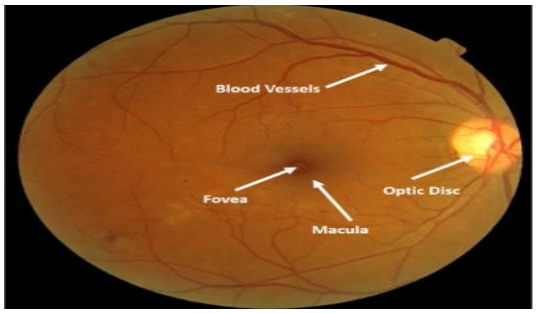


Fig. 1: Retinal scan with Optic Disk and Fovea marked [1]

2. CHALLENGES WITH PRIOR WORK

Most prior works were handicapped by one or more of the following challenges that plagued retinal bio-medical imaging either due to lack of requisite data or computational techniques. Some of them are defined below:

- **Dependence on auxiliary tasks:** A number of previous approaches performed segmentation of optic disk as a precursor to localization, similar to [2]. Hence,

[1] https://IDRid.grand-challenge.org/

building a robust segmentation pipeline became essential for key-point localization in such cases. Our method considers localization of retinal key-points disjoint from segmentation, thereby overcoming the performance bottleneck resulting from segmentation.

- **Difficulty in generalization of visual features:** [3] developed a mathematical morphology based nonlinear image processing pipeline to localize fovea. [4, 5, 6] employed edge detection, entropy filters, Hough transforms, pyramidal decomposition and spatial geometric distance for the same task. [7] showed how histogram matching of pixel densities helped build a robust localizer for optic disc in retinal images. A major drawback in such works is that they are not resilient to domain adaptation and fail when tried on images from different sources. Alternatively, our method extracts key-point features in an unsupervised fashion at different scales.
- **Asynchronous detection of OD and fovea** [8] performed independent detection of optic disk, while [9, 10] focused on fovea localization in isolation. Independent landmark localization berefts the model from utilizing the spatial information between the key-points leading to poor performance. NSSH network does simultaneous key-point detection, extendable to newer landmarks in the future.
- **Lack of interpretability:** [11] utilized a CNN-based deep neural network to segment optic disk. [12] built an ensemble based model that incorporated several independent visual features. [13] used a fully-connected DenseNet architecture for segmentation of optic disk. [14] utilized relation-networks to exploit the geometric relationships between fovea and optic disk. Application of such deep learning methods critically ignores the explainability of learnable anatomical geometric relationships in medical key-points. Moreover, these models are difficult to train due to humongous number of learnable parameters. Our method tries to employ a minimalistic model with significantly lesser parameters that outputs easily interpretable confidence heatmaps to make the task of OD and fovea detection more transparent for medical practitioners.

3. METHODOLOGY

3.1. Problem Formulation

In this paper, we propose to view the task of retinal landmark detection as a key-point estimation problem for localizing optical disk and fovea in a diabetic retinopathy image . Given a retinal scan image I, we define the optic disk and fovea points by a set P = $\{p_i, p_j\}$, where the coordinate vectors p_i and p_j represent the 2D location (x, y) in the image. Let the location of each key point be parameterized by image position and orientation $[x_i, y_i, \theta_i]$. The spatial morphology of OD and fovea can be described in terms of their geometric locality $\phi(p_k)$ and illumination intensity $\rho(p_k)$. Let the feature vectors extracted from the oriented key point patches at location p_i and p_j be given by Equations 1 and 2.

$$\vec{\phi}(p_i, p_j) = {\beta_i}^T \vec{O}(p_i) - {\beta_j}^T \vec{O}(p_j) \tag{1}$$

$$\vec{\rho}(p_i, p_j) = {\alpha_i}^T \vec{I}(p_i) - {\alpha_j}^T \vec{I}(p_j) \tag{2}$$

where $\vec{O}(p_i, p_j)$ and $\vec{I}(p_i, p_j)$ represent the spatial orientation and pixel intensity contrast vectors between OD and fovea, respectively. α_k and β_k are the fine tuned parameters learnt from the data to account for noise and illumination variations in the RGB channel. The output of the network is a set of low resolution heatmaps where, for a given heatmap, the network predicts the probability of occurrence of either optic disk or fovea given the other's presence at each and every pixel. Contradicting evidence and anatomic impossibility are implicitly learnt at different scale resolutions by maintaining precise local information while considering and then reconsidering the overall coherence of the features. A landmark-estimation function is defined to take in the input image and learn the spatial orientation and pixel intensity contrast to distinctively mark the fovea and OD points as given by Equation 3 and 4 respectively. The landmark estimator algorithm relies on heatmap regression to output the likelihood of expected key-point at each pixel location (${H_k}^2 D$), where $\emptyset$ is the space of all possible pixel locations, σ controls the standard deviation of the heatmaps, E is the joint expectation of OD and fovea locations and g_k denotes the ground truth annotation of the landmarks.

$${H_i}^{2D}(p_i) = E(-\frac{||p - g_i||}{\sigma^2}) \forall p \in \emptyset - p_j \tag{3}$$

$${H_j}^{2D}(p_j) = E(-\frac{||p - g_j||}{\sigma^2}) \forall p \in \emptyset - p_i \tag{4}$$

3.2. Naive Single Stack Hourglass Network

Inspired by [15], we propose the *Naive Single Stack Hourglass Network* (NSSH) which incorporates three salient design decisions. **Hourglass Geometry:** The first-half of the NSSH network acts as an encoder which is performs coarse feature extraction. The second half is supposed as a decoder consisting of deconvolutional layers stacked to recover the fine-grained details of the input from the decoder outputs. The proposed network has an upsampling layer for each corresponding pooling layer, thus following an hourglass geometry. **Convolutional Layer Stacking:** Stacking convolution layers followed by repeated pooling and upsampling at each resolution is known to preserve spatial information across scales. NHHS has a single stack structure for optimal performance. Addition of more stacked blocks led to degraded performance. **Replacing ResNet blocks with FC-ConvNets:** Traditionally, deep ResNet blocks are used in

stacked hourglass networks to provide residual learning effect. However, upon empirical analysis, such deeper residual layers did not significantly improve performance in our case. Rather, the ResNet blocks increased the model training time, number of trainable parameters and the model's tendency to overfit. Hence, a conscious design choice was to replace the residual blocks with simple fully-connected convolutional layers that helped to capture a larger spatial context. The unit convolutional blocks are added at the end to perform heatmap regression as discussed in Section 3.1.

Naive Single Stack Hourglass Network has four downsampling and upsampling steps. All convolutional layers in downsampling and upsampling steps have filter size of 3 x 3. At each max pooling step, the network branches off and applies more convolutions at the original pre-pooled resolution. After reaching the lowest resolution, the network begins the top-down sequence of upsampling and combines features across scales. The 2 x 2 max pooling is used to halve the size of the feature maps, and the nearest neighbor interpolation is used to double the size of the feature maps in the upsampling steps. The maximum feature maps in convolutional layers are fixed to 256 after trying several versions of the model with 64, 128, 256, 512 and 1024 feature maps. After the last upsampling layer, a single 3 x 3 convolution and two 1 x 1 convolution is performed to generate network outputs. Lastly, a 1 x 1 convolution is applied to the outputs to match the number of input feature maps to the number of channels. This is followed by another 1 x 1 convolution for output heatmap generation.

3.3. Loss Function

For training the key-point estimation function, we use Mean Squared Error (MSE) based loss function taking into account both optic disk and fovea landmarks. Let ${M_k}^{2D}$ represent the predicted 2D Gaussian confidence map for each k^{th} annotations. Thus, the confidence maps ${M_i}^{2D}$ and ${M_j}^{2D}$ for optic disk and fovea respectively as given by Equation 5 and 6.

$$ {M_i}^{2D}(p_i) = \frac{1}{2\pi\nu} \exp\left(\frac{-[(<\vec{p}-\vec{p_i},\vec{x}>)^2 + (<\vec{p}-\vec{p_i},\vec{x}>)^2]}{2\nu}\right) \tag{5} $$

$$ {M_j}^{2D}(p_j) = \frac{1}{2\pi\nu} \exp\left(\frac{-[(<\vec{p}-\vec{p_j},\vec{y}>)^2 + (<\vec{p}-\vec{p_j},\vec{y}>)^2]}{2\nu}\right) \tag{6} $$

where ν denotes square of spatial variance, and $<\vec{u},\vec{v}>$ indicates the inner product of vectors $\vec{u}$ and $\vec{v}$. The MSE loss is then formulated as given by Equation 7, with ${H_k}^{2D}(p_k)$ as the ground truth confidence map and ${M_k}^{2D}(p_k)$ and predicted confidence map.

$$ \mathscr{L} = \frac{1}{2}(\gamma||{H_i}^{2D} - {M_i}^{2D}||_2^2 + (1-\gamma)||{H_j}^{2D} - {M_j}^{2D}||_2^2) \tag{7} $$

We use SGD with RMSProp as the optimizer for the NSSH model by back-propagating the mean squared errors on training data through batch normalization.γ is a hyperparameter to adjust the weights corresponding to the loss of each landmark. Through cross validation, it was observed that keeping equal weightage for optic disk and fovea loses gives the most optimal loss convergence.

Method	Euclidean Distance Error	
	Optic Disk	Fovea
ISBI - 2018 Challenge (Rank 1)[2]	25.61	45.89
ResNet-18	80.48	115.12
ResNet-50	60.32	95.45
Relation Network Regressor [14]	26.12	43.46
NSSH Network (proposed)	**14.21**	**35.45**

Table 1: Results of NSSH network and baselines

4. EXPERIMENTATION SETTINGS

4.1. Dataset

The dataset used for experimentation is the Indian Diabetic Retinopathy Image Dataset (IDRiD) [16]. The database consists of 516 images with center pixel locations of optic disc and fovea. It is divided into train, validation and test set of 383 (75%), 30 (5%) and 103 (20%) respectively. The dataset was distributed as a part of "Diabetic Retinopathy: Segmentation and Grading Challenge" workshop at organized at ISBI-2018. The input image size is 4288 x 2848 which was resized to 1024 x 1024. To avoid overfitting and improve generelization, data augmentation of flips and rotations was applied followed by color normalization for standardizing illumination across images. The augmentation regime was kept minimal to avoid inadvertent distortions in natural anatomical geometrical relationships between optic disk and fovea.

4.2. Network Training

The Naive Single Stack Hourglass Network was trained on Pytorch using RMSProp optimizer for optimal model convergence. The GPU used for running the experiments was NVIDIA 2080Ti, with batch size 4 and training time of 14 hours on average. The learning rate was kept at 25 x 10^{-5} for 100 epochs.

5. RESULTS AND DISCUSSIONS

Figure 2 highlights several test examples predicted by NSSH network. (●) depicts the ground truth landmarks while (●) demarcates the predicted key-points. The detection of optic disk and fovea was evaluated through mean Euclidean distance metric which is given by $\frac{1}{M}\sum_{M}^{i=1}(x_{predicted} - x_{groundtruth})^2 + (y_{predicted} - y_{groundtruth})^2$. Table 1 shows the comparison of our proposed networks with baseline

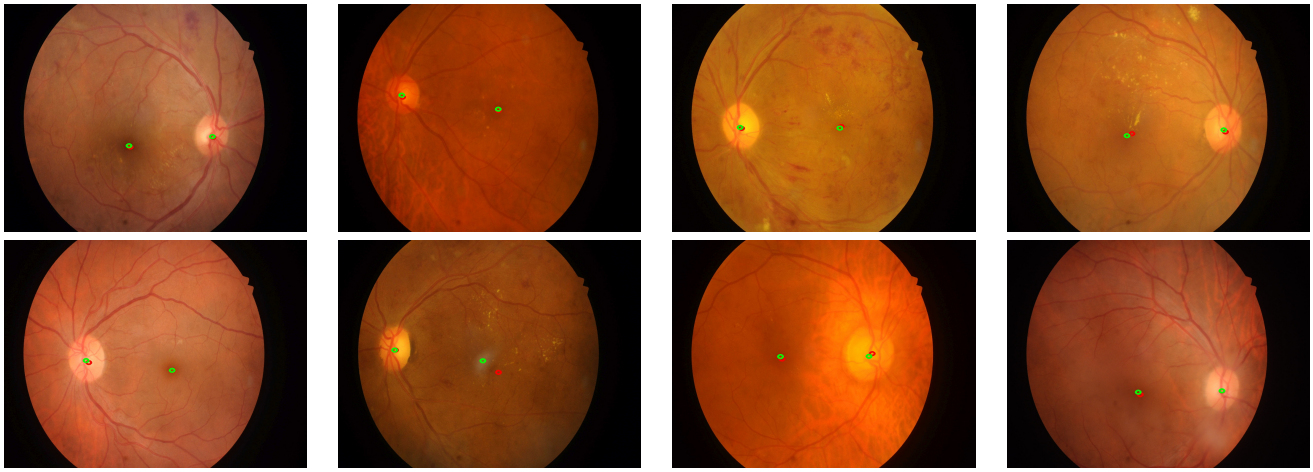
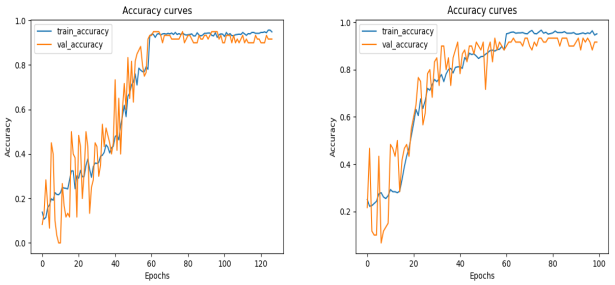


Fig. 2: Ground truth (green) and predicted (red) key points of Optic Disk and Fovea detected by NSSH network.

Method	Euclidean Distance Error		# of Parameters (in millions)	Training Time (hr)
	Optic Disk	Fovea		
Naive single Stacked Hourglass Network	**14.21**	**35.45**	**3.58**	**14**
Naive Double Stacked Hourglass Network	15.10	44.69	6.72	20
ResNet-50 Single Stacked Hourglass Network	15.05	318.10	34.0	36

Table 2: Ablation analysis of stacking and layering structures



(a) Single Stacked Hourglas (b) Double Stacked Hourglass

Fig. 3: Training statistics for Naive Single and Double Stacked Hourglass Networks

ResNet models and state of the art systems [14]. NSSH network outperforms the Relation Network Regressor put forth by [14] by a reduction of 45.6% and 18.4% in euclidean distance error for OD and fovea detection, respectively. Thus, it has been established that approaching retinal landmark detection from a pose estimation perspective substantially outweighs all previous strategies experimented in this domain. Moreover, we present an ablative analysis to understand the design choices in NSSH network. Table 2 shows variants of the proposed NSSH framework. In one of the variations, two stacked modules were used to study the impact of increasing stacking on model performance. It was observed that this led to a decrease in MSE for both optic disk and fovea. A plausible reason can be the rise in number of model parameters that tend to overfit sooner which took more time to train. This is evident from Figure 3 where the single and double stacked architectures perform identically, except that the former converges faster. Alternatively, the original ResNet version of proposed network was tried as given by [15]. This model was 10 times as bulky as the NSSH network, took 1.5 times more time to train and still performed poorly on fovea detection. Interestingly, it was found that OD detection is comparatively easier for all versions of hourglass models, while fovea detection emerges as a non-trivial task.

6. CONCLUSION

This work summarizes the proposed Naive Single Stacked Hourglass network as an excellent advancement for the detection of the optic disc and fovea in retinal fundus images. We demonstrate that the pose estimation algorithms can be reformulated to locate key-points in biomedical images, with promising improvements in performance metrics. Moreover, the stacked hourglass model is robust to scale and illumination distortions, faster to train and more interpretable due to its ability to learn hierarchies of features at different scales. In future, the same architecture can be utilized in other challenging biomedical imaging tasks to develop clinical applications.

7. REFERENCES

[1] Baidaa Al-Bander, Waleed Al-Nuaimy, Bryan M Williams, and Yalin Zheng, "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc," *Biomedical Signal Processing and Control*, vol. 40, pp. 91–101, 2018.

[2] Syed S Naqvi, Nayab Fatima, Tariq M Khan, Zaka Ur Rehman, and M Aurangzeb Khan, "Automatic optic disk detection and segmentation by variational active contour estimation in retinal fundus images," *Signal, Image and Video Processing*, pp. 1–8, 2019.

[3] J Benadict Raja and CG Ravichandran, "Automatic localization of fovea in retinal images based on mathematical morphology and anatomic structures," *International Journal of Engineering and Technology*, vol. 6, no. 5, pp. 2171–2183, 2014.

[4] Arturo Aquino, Manuel Emilio Gegundez, and Diego Marin, "Automated optic disc detection in retinal images of patients with diabetic retinopathy and risk of macular edema," *International Journal of Biological and Life Sciences*, vol. 8, no. 2, pp. 87–92, 2012.

[5] Laszlo Kovacs, Rashid Jalal Qureshi, Brigitta Nagy, Balazs Harangi, and Andras Hajdu, "Graph based detection of optic disc and fovea in retinal images," in *4th International Workshop on Soft Computing Applications*. IEEE, 2010, pp. 143–148.

[6] José Pinão and Carlos Manta Oliveira, "Fovea and optic disc detection in retinal images with visible lesions," in *Doctoral Conference on Computing, Electrical and Industrial Systems*. Springer, 2012, pp. 543–552.

[7] Amin Dehghani, Hamid Abrishami Moghaddam, and Mohammad-Shahram Moin, "Optic disc localization in retinal images using histogram matching," *EURASIP Journal on Image and Video Processing*, vol. 2012, no. 1, pp. 19, 2012.

[8] Aliaa Abdel-Haleim Abdel-Razik Youssif, Atef Zaki Ghalwash, and Amr Ahmed Sabry Abdel-Rahman Ghoneim, "Optic disc detection from normalized digital fundus images by means of a vessels' direction matched filter," *IEEE transactions on medical imaging*, vol. 27, no. 1, pp. 11–18, 2007.

[9] M Niemeijer, MD Abramoff, and B Van Ginneken, "Automated localization of the optic disc and the fovea," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 3538–3541.

[10] A Basit and Muhammad Moazam Fraz, "Optic disc detection and boundary extraction in retinal images," *Applied optics*, vol. 54, no. 11, pp. 3440–3447, 2015.

[11] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool, "Deep retinal image understanding," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 140–148.

[12] Rashid Jalal Qureshi, Laszlo Kovacs, Balazs Harangi, Brigitta Nagy, Tunde Peto, and Andras Hajdu, "Combining algorithms for automatic detection of optic disc and macula in fundus images," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 138–145, 2012.

[13] Baidaa Al-Bander, Bryan Williams, Waleed Al-Nuaimy, Majid Al-Taee, Harry Pratt, and Yalin Zheng, "Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis," *Symmetry*, vol. 10, no. 4, pp. 87, 2018.

[14] Sudharshan Chandra Babu, Shishira R Maiya, and Sivasankar Elango, "Relation networks for optic disc and fovea localization in retinal images," *arXiv preprint arXiv:1812.00883*, 2018.

[15] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[16] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, pp. 25, 2018.