New Deep Neural Nets for Fine-Grained Diabetic Retinopathy Recognition on Hybrid Color Space

Holly H. Vo and Abhishek Verma
Department of Computer Science
California State University
Fullerton, California 92834, USA
Email: hhvo@csu.fullerton.edu, averma@fullerton.edu

***Abstract*—Automatic diabetes retinopathy (DR) recognition can help DR carriers to receive treatment in early stages and avoid the risk of vision loss. In this paper, we emphasize the role of multiple filter sizes in learning fine-grained discriminant features and propose: (i) two deep convolutional neural networks - Combined Kernels with Multiple Losses Network (CKML Net) and VGGNet with Extra Kernel (VNXK), which are an improvement upon GoogLeNet and VGGNet in context of DR tasks. Learning from existing research, (ii) we propose a hybrid color space, LGI, for DR recognition via proposed nets. (iii) Transfer learning is applied to solve the challenge of imbalanced dataset.**

The effectiveness of proposed new nets and color space is evaluated using two grand challenge retina datasets: EyePACS and Messidor. Our experimental results show: (iv) CKML Net improves upon GoogLeNet and VNXK improves upon VGGNet on both datasets using the LGI color space. Additionally, proposed methodology improves upon other state of the art results on Messidor dataset for referable/non-referable screening.

***Index Terms*—diabetic retinopathy recognition; hybrid color space; convolutional neural networks; LGI; CKML Net; VNXK; transfer learning**

I. Introduction

Diabetic retinopathy (DR) is a common eye disease which affects one in three diabetes carriers in America and could lead to irreversible vision loss. Although, timely treatment can reduce the risk of severe vision loss by over 90%, DR carriers do not notice vision changes until the late stages. Meanwhile the manual process of grading a retina image consumes time and labor.

Fig. 1 shows a fundus image with labeled signs of diabetic retinopathy. Microaneurysms (MAs) are tiny bulges in blood vessels and appear as deep-red dots. Hemorrhages are small spots of blood discharge. Hard exudates are leakage of lipid and protein in the retina. Hard exudates typically emerge as bright, reflective lesions. Depending on the presence of DR signs and their complexity, a fundus image can be marked by an ophthalmologist as normal or as some specific level of DR.

There exist several researches in the DR domain to automate recognition of stages. Most of them offer complex feature engineering framework that are either DR-lesion based or ensemble of multiple feature extraction techniques. In this work, we propose novel deep learning convolutional neural networks with transfer learning on a discriminant color space for DR recognition task. The rest of the paper is organized as follows. Section II introduces two DR datasets to be used in this study and discusses their challenges. Section III reviews related work and the possible techniques to solve DR recognition. Section IV describes the proposed deep learning networks and hybrid color space. Experimental results and discussion is presented in section V. Section VI concludes the research with future direction.

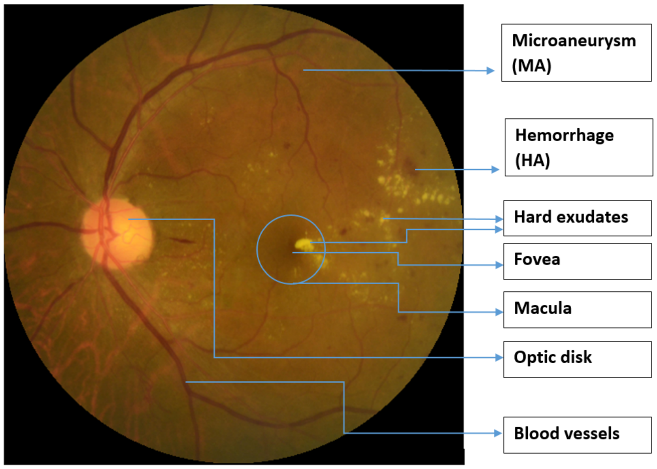

Fig. 1. Main structure of a retina image and DR signs.

II. Description of Retina Datasets

In this paper, we evaluate the proposed networks and hybrid color space on two publicly available datasets: EyePACS and Messidor.

A. EyePACS

It is a diabetic retinopathy image dataset provided by EyePACS, a free platform for retinopathy screening, through Kaggle website [10] in 2015. The dataset consists of 35,126 training images and 53,576 testing images. These images are taken from various models and types of camera; under varied conditions and stored at different high resolutions. Each image has been manually examined for the presence of DR signs by a clinician and labeled with a DR stage from 0 to 4, which corresponds to no DR, mild, moderate, severe, and proliferative DR. Training set comprises approximately of 74% images from stage 0 (no DR), 7% from stage 1 (mild), 15%

from stage 2 (moderate), 2% from stage 3 (severe) and 2% from stage 4 (proliferate DR). Class ratios of the test set are similar to those in the training set.

The first challenge of this dataset is its large variation in resolution, intensity, and quality. From examining the training set we note: image height varies from 289 to 3,456 pixels, width varies from 400 to 5,184 pixels, range for ratio of height and width is from 0.66 to 1.00. On a scale of 0-255, average image intensity ranges from 1 to 192 and mean average image intensity is 63. File size for low resolution images is 8KB and goes up to 2MB for higher resolution images. Secondly, the dataset is highly imbalanced in terms of distribution of images across classes. Most machine learning algorithms look to minimize the overall error rate, the rare classes may end up having worst performance in the learning process [7], [9]. In this particular retinopathy dataset, rare classes are the classes that carry DR signs, and improving their individual performance could have a huge positive impact in DR domain.

B. Messidor

The Messidor dataset is publicly available for studies on computer-aided diagnosis (CAD) of DR [11]. The dataset consists of 1,200 digital fundus images that are prepared by three French ophthalmology departments via a non-mydriatic digital color video 3CCD camera with a 45 degree field of view. The images are captured in one of three high resolutions: 1,440x960, 2,240x1,488, or 2,304x1,536 with 0-255 range for each color channel. Each image has been diagnosed with a DR stage from 0 to 3. Among 1,200 images, there are 546 no-DR images, 254 images from stage 1, 247 from stage 3, and 153 from stage 4. Although, DR severity distribution in this dataset does not represent the real-world population, Messidor dataset is still widely used in retinopathy studies [12], [13], [14] to evaluate performance of different DR screening methods.

For DR screening purposes, there are different ways to convert four classes Messidor to a binary dataset. One way is to label no-DR images as normal and group images of other stages as abnormal as in the works of [12], [13]. However, given the fact that the difference between normal images and images of stage 1 is the most difficult task for both CAD systems and experts, Sánchez et al. [14] grouped stages 0 and 1 of the Messidor dataset as referable images and combined stages 2 and 4 as non-referable in their screening work.

III. Background

A. Feature Engineering Frameworks

It is a clinical fact that MAs are the earliest signs of diabetic retinopathy [19], hence, there are many DR recognition frameworks that are built upon localization of segment lesions, blood vessels, optic disks, and macula one by one. Basic point operators are applied to balance and enhance local contrast; linear filters and neighborhood operators such as morphological operators, median filters, and Gaussian filters are convolved on images during preprocessing step as indicated in [14] and in survey work [20], [26]. Watershed transformation is considered in [27] to solve the issue of over segmentation due to thresholding. Other techniques such as active contour modeling and recursive region-growing technique (RRGT) are applied in the domain researches to isolate blood vessels and other interesting regions [26].

Texture extraction is another approach in DR recognition. Statistical texture extraction is based on the relationship between pixel intensity. Contrast texture is extracted together with areas of MAs and HAs in [28] for DR classification. In [13], Tang et al. proposed novel splat feature classification method to detect retinal hemorrhages. The retina color images are partitioned into non-overlapping segments, i.e., splats, under a supervised manner. Each splat is represented by information of color, spatial location, shape, and texture of the splat and its interactions with neighboring splats. Finally, an optimal subset of splat feature is selected by wrapper approach. Pires et al. [12] utilized bags-of-visual-words representation and BossaNova, and Fisher Vector to detect lesions. The probability scores obtained from lesion detectors are used to represent the retina images.

Various color spaces and individual color channels are considered in DR researches to improve effectiveness of segmentation. HSI is applied on retina images for extraction of MAs and exudates [21], and for locating fovea [22]. Green component of RGB color space is considered for extracting blood vessel structure in [23], [28]. In [25], morphological operations are performed on each channel of RGB to extract the total area and perimeter of blood vessels, HAs, and MAs. Ram et al. [24] extracted lesion pixel values on multiple color spaces such as RGB, L*u*v*, HSV and HSI.

B. Convolutional Neural Networks

In the last several years deep convolutional neural networks (CNNs) has emerged as the leading and prevailing technique for computer vision tasks. CNNs do not only win in performance but also offers an end-to-end framework, from feature extraction to classification. The annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [15] has proven to be a popular arena to introduce and evaluate performance of different network architectures on large-scale image datasets. Among those, GoogLeNet, VGGNet, and ResNet have attracted research attention for both accuracy and revolutionary ideas in architecture.

1) GoogLeNet: GoogLeNet is the winner of ILSVRC 2014 by Szegedy et al. [2] of Google. GoogLeNet is known for its improvement in computation by increasing the net in both depth and width. Its low computational cost is based on following ideas. First, optimize convolutional neural network through the use of connection sparsity [3]. Secondly, perform dimensionality reduction through 1x1 convolutional layer as proposed by Lin et al. in their model Network-in-Network [4]. For extracting more details, a max-pooling layer and multiple filters of sizes 1x1, 3x3, and 5x5 are arranged in parallel. The 3x3 and 5x5 filters are preceded by a 1x1 convolutional layer to reduce dimensionality, while the max-pooling layer is followed by a 1x1 convolutional layer to remove computational bottlenecks. The network contains 22 parameter layers in depth.

To overcome the problem of vanishing gradient in deep networks with large depth, GoogLeNet adds auxiliary classifiers to intermediate layers of the network. These auxiliary classifiers build extra small convolutional networks on top of the partial network with a high dropout ratio. During training, their loss contributes to the total loss of the network at a discount weight. Thereby, it provides extra supervision to the earlier parameter layers in the network.

2) VGGNet: VGGNet by Simonyan et al. [5] won the second place in classification task of ILSVRC 2014. Their best 16-layer network is built upon 3x3 convolutional layers with stride of one and 2x2 max pooling layers with stride of two. Their work is significant in showing that the depth of the network is beneficial for classification accuracy. The downside of VGGNet is its high cost in terms of memory usage and computation time, thus, it is expensive to evaluate this network.

3) Residual Network (ResNet): Residual network, the winner of ILSVRC 2015, was implemented by He et al. [6] of Microsoft. The network employs shortcut paths that perform identity mapping in order to achieve the desired output. In each residual unit, an input is branched out: one goes into the function and is transformed in the "residual" branch while the "identity" branch bypasses the function. The residual network implementation heavily employs batch normalization to reduce internal covariate shift and accelerates the training [8]. However, one of the known issues with residual net is that changes to batch sizes impacts accuracy. At lower batch sizes the accuracy is low and higher batch sizes require more hardware resources. Reported accuracy utilized eight modern GPUs for a batch size of 256 images [8] to obtain competitive results. Due to limitation of available hardware resources at this time we are unable to benchmark ResNet on the retina datasets.

C. Transfer Learning on Imbalanced Datasets

As aforementioned in the dataset section, the imbalanced distribution of images across DR classes is a challenge in DR recognition because rare classes are dominated by majority classes and achieve worst accuracy in the learning process. According to Rao et al. [32], the consequence of miss-classifying a disease carrying class is overwhelmingly costly in diagnostic problems in medical domain. For DR recognition problem, classes of later DR stages are rare but very significant. Therefore, the need for a classifier that provides high accuracy for the minority class without severely penalizing the overall accuracy arises. Different sampling techniques, cost-sensitive methods, and kernel-based methods are well-known solutions to problem of imbalanced classification [7], [9].

For CNN framework, oversampling is a common and simple solution to balance data. In particular, retina images of rare classes can be rotated to increase its sample size. However, this technique adds redundancy to the dataset and may lead to overfitting [33]. On EyePACS dataset, this technique approximately increases the total sample size by 3.7 times and consequently increases the training time.

Firstly, in this work, we solve class imbalance issue by transfer learning. In particular, pre-trained weights from GoogLeNet and VGGNet that were trained on 1,000 classes ImageNet 2015 dataset [15] will be utilized in our 5-classes DR recognition task on the EyePACs dataset. ImageNet dataset includes large number of object and scene classes. According to Yosinski et al. [34], initializing almost any number of layers of a deep CNN with transfer learned features from another network has proven to boost generalization in similar target tasks. In highly imbalanced datasets, the pre-trained weights initialize the network with some discriminant abstract features. Thereby, it prevents training to converge to a local minima where over-fitting of dominant classes occurs.

Secondly, how could DR recognition task benefit from transfer learning from ImageNet? In fact, both object and scene classes are proven to be successfully identified by texture descriptors in [16], [30]. Most learned filters of the first convolutional layer in AlexNet, Network-in-Network, and GoogLeNet are similar to Gabor filters, which is a well-known technique for texture representation and discrimination. As texture features are significant in DR recognition, the two tasks are similar at the abstract feature level.

IV. Methodology

The methodology of this work is derived from GoogLeNet and VGGNet-16 and the weights these nets have learned are on ImageNet dataset. To examine DR signs on larger image size, the stride of the first convolutional layers on both GoogLeNet and VGGNet-16 is simply doubled to support the double image size without changing any other layers or parameters in these nets. Furthermore, we propose two deep neural network architectures, namely, Combined Kernels with Multiple Losses Network (CKML Net) and VGGNet with Extra Kernel (VNXK), to train and test the large EyePACS dataset on proposed LGI, which is a hybrid color space. The trained nets are then applied on Messidor dataset to extract features for DR screening.

We use Caffe [17], which is an open source deep learning software framework developed by the Berkley Vision and Learning Center. Caffe plugs in into the NVIDIA DIGITS platform [18], which is a Deep Learning GPU Training System. NVIDIA DIGITS [18] is an open source project that enables the users to design and test their neural networks for image category classification and object detection with real-time visualization. The hardware configuration of our system is one NVIDIA GeForce GTX TITAN X GPU with 12GB of VRAM. The system has two Intel Xeon processors E5-2690 v3 2.60GHz with a total of 48/24 logical/physical cores and 256 GB of main memory.

GoogleNet and CKML Net take from 30 to 40 epochs and last approximately 2-3 hours to train on EyePACS dataset. VGGNet and VNXK converge within 20 to 30 epochs and take approximately 5-7 hours. All nets are initialized with some small learning rate of around 0.001 with 2-3 steps of 0.2 or 0.3.

A. Image Pre-processing

The main task in image pre-processing is to convert input images of any size to a fixed square image size. For each input image, a circumscribing rectangle of the eye is determined by scanning pixels along its horizontal and vertical mid-lines. The image is then cropped around the center of the circumscribing rectangle to extract a square image whose dimension is the shorter side of the boundary rectangle. Next, the square image is scaled by bicubic interpolation method to a dxd image before being clipped around its center to maximize retina content in the final output square image. 256x256 images are prepared for GoogLeNet and VGGNet, and 512x512 images are for CKML Net and VNXK. Weights of new layers are transferred from equivalent layers using Python and Caffe [17] libraries.

B. Proposed hybrid color space: LGI

The proposed hybrid color space LGI is derived from three color spaces: L*a*b*, RGB, and $I_1I_2I_3$. These color spaces are commonly used in DR recognition for feature extraction. In L*a*b*, the luminance channel L outperforms the other two chrominance channels in DR recognition using local binary descriptors [35]. The green channel performs best in RGB [23], [28], [35]. In $I_1I_2I_3$, the first channel performs better than other two channels as it holds most chrominance and luminance information [35]. Thus, LGI is simply composed of the most discriminant channel in DR recognition from each aforementioned color space. Except, for L channel that represents luminance in the range from 0 to 100, other channels are in 0-255 range.

C. Proposed Nets

The 7x7 filters in the first convolutional layer of GoogLeNet that was trained on ImageNet dataset resemble the Gabor filters, which is a popular technique for texture representation and discrimination. Another thing to note is that the DR signs in retina datasets are fine-grain and vary highly in granularity. Therefore, by increasing the range of Gabor-like filter scales can capture more discriminant features in DR recognition. We propose the following nets:

1) Combined Kernels with Multiple Losses Network (CKML Net): CKML Net is inspired from GoogLeNet. In the first part, from the first convolutional layer up to *pool2/3x3-s2* layer, are replaced by three parallel branches. Each branch starts with a convolutional layer of 64 KxK filters and with stride of 4, where K $\in \{7, 11, 15\}$ and replicates GoogLeNet structure up to *inception_3a/1x1* layer to obtain an 28x28x64 output blob, named *conv2_1.K* in CKML Net. In each branch, *conv2_1.K* is connected to an auxiliary classifier with discount weight of 0.3 to control the discrimination of features learned in the branch. A layer, named *inception_mk*, that is composed of concatenating *conv2_1.K* layers of three branches, replaces the original *pool2/3x3-s2* layer of GoogLeNet. All GoogLeNet layers below its *pool2/3x3-s2* are replicated in CKML Net. Fig. 2 shows differences between GoogLeNet and CKML Net.

For each branch K, where $K \neq 7$, filters of its first convolutional layer are created by scaling the original 7x7 filters to KxK with bicubic interpolation method. To prepare parameters for other layers of each branch, first layer's initialized filters and auxiliary classifier of each branch are separately plugged into the original pre-trained GoogLeNet and re-learned on EyePACS dataset. The learned parameters of each branch are then copied to its corresponding layers in CKML Net. Training process for GoogLeNet and CKML Net takes between 30 to 40 epoch.

2) VGGNet with Extra Kernel (VNXK): VNXK is inspired mainly from VGGNet, additionally it combines ideas from GoogLeNet and ResNet. It is a challenging task to extend VGGNet by stacking additional layers or fork branches for two main reasons: (i) it takes much longer training time than GoogLeNet and (ii) the net itself is too heavily weighted and extending it with more layers or branches greatly increases the number of parameters. VNXK is created by converting the early part of VGGNet to a residual unit. In ResNet, the residual unit is defined as:

$$y = F(x, \{W_i\}) + W_s x \tag{1}$$

where $F(x, \{W_i\})$ represents multiple convolutional layers and W_s is only used for matching dimension of two components [7]. In VNXK, W_s is replaced by the first convolutional layer of GoogLeNet, and $F(x, \{W_i\})$ represents all layers from the beginning up to *pool1* of the original VGGNet. The parameters of W_s are copied from pre-trained GoogLeNet. Stride of two is applied to the first layer of $F(x, \{W_i\})$ and stride of four is applied to W_s to support double size images as explained in Section IV-A. Fig. 3 shows differences between VGGNet and VNXK.

V. Experimental Results and Discussion

Experiments are organized in two sets, each toward particular goal. First set of experiments focus on training and testing proposed convolutional neural networks, CKML Net and VNXK, on the hybrid color space LGI on EyePACs dataset. Second set of experiments apply the proposed nets that were fine-tuned in the previous experiments on EyePACS dataset onto Messidor dataset to extract features for screening evaluation. Thereby, transfer learning is applied from EyePACS to Messidor dataset.

A. Performance of Proposed Nets and LGI Color Space on EyePACS Dataset

First set of experiments are conducted for EyePACs dataset on two color spaces: RGB and LGI for six nets: GoogLeNet, VGG, GoogLeNet/2xs, VGG/2xs, CKML Net, and VNXK. First, the original GoogLeNet and VGGNet are fine-tuned on 256x256 images. Experiments are then repeated for GoogLeNet/2xs and VGGNet/2xs on 515x512 images. Note that the high resolution original images are down-sampled to 512x512 or 256x256. Recall that GoogLeNet/2xs and

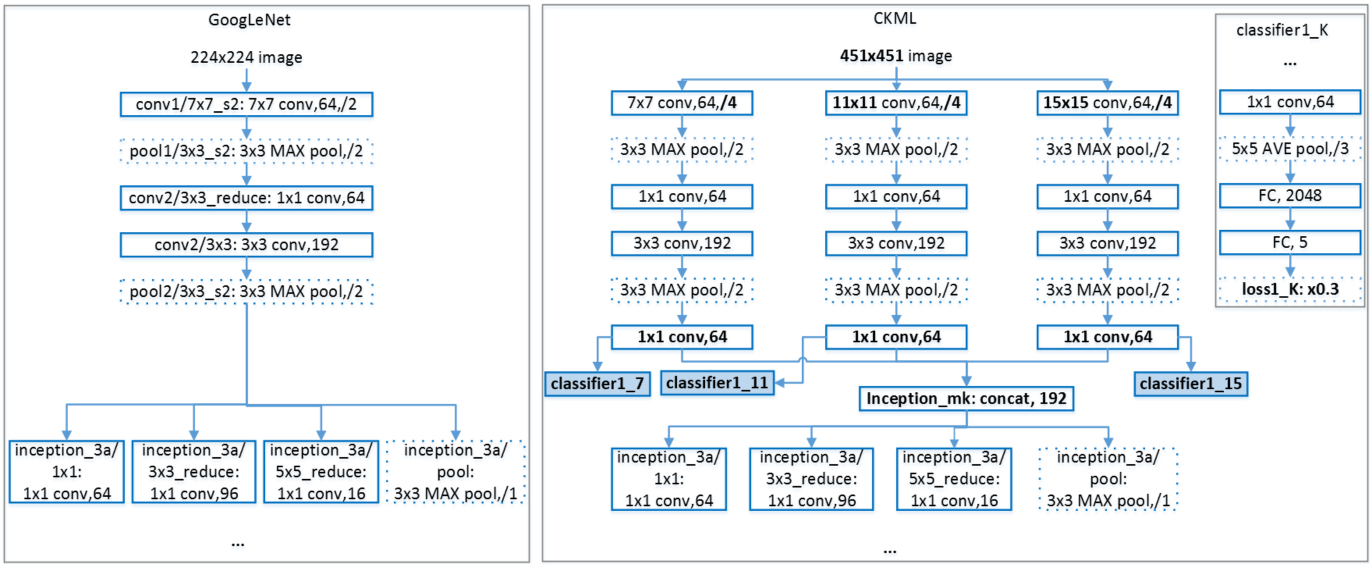
GoogLeNet
224x224 image
conv1/7x7_s2: 7x7 conv,64,/2
pool1/3x3_s2: 3x3 MAX pool,/2
conv2/3x3_reduce: 1x1 conv,64
conv2/3x3: 3x3 conv,192
pool2/3x3_s2: 3x3 MAX pool,/2
inception_3a/ 1x1: 1x1 conv,64
inception_3a/ 3x3_reduce: 1x1 conv,96
inception_3a/ 5x5_reduce: 1x1 conv,16
inception_3a/ pool: 3x3 MAX pool,/1
...
CKML
451x451 image
7x7 conv,64,/4
11x11 conv,64,/4
15x15 conv,64,/4
3x3 MAX pool,/2
1x1 conv,64
3x3 conv,192
3x3 MAX pool,/2
1x1 conv,64
classifier1_7
classifier1_11
classifier1_15
Inception_mk: concat, 192
inception_3a/ 1x1: 1x1 conv,64
inception_3a/ 3x3_reduce: 1x1 conv,96
inception_3a/ 5x5_reduce: 1x1 conv,16
inception_3a/ pool: 3x3 MAX pool,/1
...
classifier1_K
...
1x1 conv,64
5x5 AVE pool,/3
FC, 2048
FC, 5
loss1_K: x0.3

Fig. 2. Difference in network architecture between GoogLeNet [2] (left) and proposed CKML Net (right).

Table I. Classification Accuracy (%) on RGB and LGI Color Spaces and Training Time of Various Network Architectures for EyePACS Dataset

Input Image	Model	Image Data Size	Train (hrs)	RGB	**LGI**
256x256	GoogLeNet	224	0.95	79.71	80.22
	VGGNet	224	5.18	82.34	*81.16*
512x512	GoogLeNet/2xs	453	1.64	81.61	82.17
	CKML Net	451	1.68	**82.30**	**82.88**
	VGGNet/2xs	449	6.38	82.93	83.38
	VNXK	449	7.08	**83.12**	**83.63**

VGGNet/2xs are respectively derived from GoogLeNet and VGGNet by doubling the stride of each first convolutional layer to support double image size. Finally, the proposed nets, CKML Net and VNXK, are experimented on 512x512 images. After being trained, each net is tested on the test set: on the same color space and same image size, which corresponds to the training set. For each experiment, the overall test accuracy is recorded in Table I. Column "image data size" indicates the actual image data size that is automatically resized in data layer by each net.

We observe that randomly initialized GoogLeNet and VGGNet give around 73.5% accuracy on validation. The validation set is 20% of total dataset less test set. Pre-trained GoogLeNet and VGGNet surpass the accuracy of 73.5% within a couple of epochs. Table I shows that VGGNet strongly surpasses GoogLeNet on RGB color space by approximately 2.5% and on LGI color space by 0.9%.

The original VGGNet is the only net that degrades more than 1% on LGI. As 3x3 filters on its first layer resemble color blobs rather than texture blobs. Moreover, it is hard to capture various texture patterns in such small filters. VGGNet seems to be highly constraint to color information, thus, applying its learned features on a different color space degrades its performance. However, all other nets, including VGGNet with double stride and VNXK, gain approximately 0.5% on the proposed hybrid LGI color space.

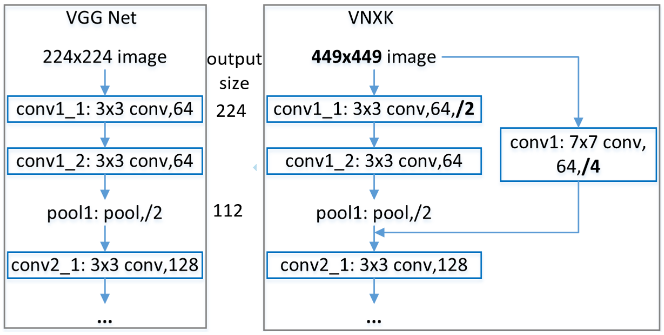

Fig. 3. Difference in network architecture between VGGNet [5] (left) and proposed VNXK (right).

The simple trick on doubling the stride for double image size adds approximately 0.5% on VGGNet/2xs. As the technique increases GoogLNet performance by 2% on RGB and 1% on LGI, it seems 7x7 Gabor-like filters is closer to the sizes and texture patterns of DR signs on 453x453 retina images. Thus, it is obvious that CMKML further improves GoogLeNet/2xs because the extra filter branches, 11x11 and 15x15, can capture larger DR signs.

VNXK surpasses VGGNet/2xs on RGB and LGI by 0.2% or more. Combination of VNXK and LGI color space is the best model in this experiment set. VNXK/LGI contributes a total gain of 1.3% upon the original VGGNet on RGB. Although CKML Net/LGI cannot outperform VNXK and VGGNet/2xs, it improves upon GoogLeNet/RGB by 3.2%. Moreover, CKML Net/LGI can be more efficiently evaluated and extended than VNXK and VGGNet variants because it is

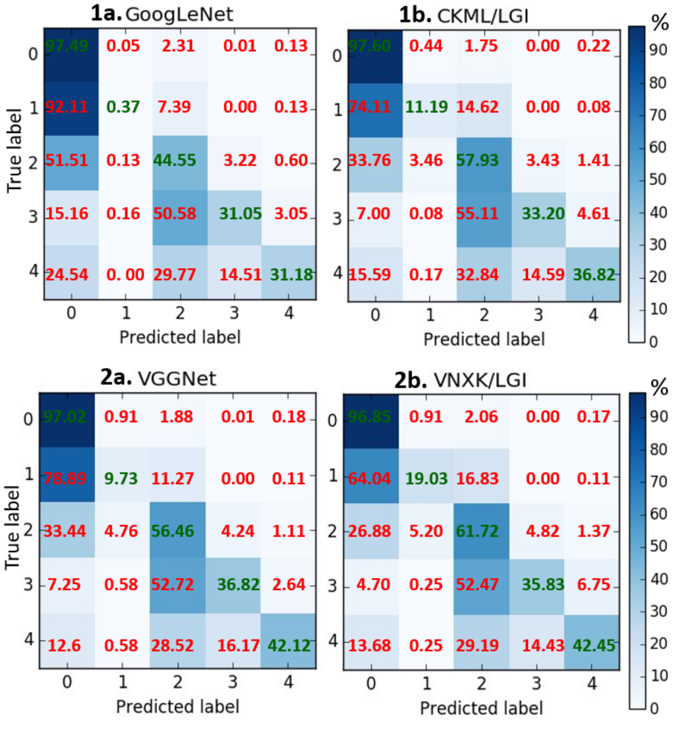

Fig. 4. Confusion matrices of classification accuracy on EyePACS dataset: 1a. GoogLeNet, 1b. CKML Net/LGI, 2a. VGGNet, and 2b. VNXK/LGI.

several times faster than VGGNet. For this reason we focus on comparing the accuracy of VNXK with VGGNet and its variants and CKML Net with GoogLeNet and its variants.

Improvements of the proposed nets upon their counterparts in terms of overall accuracy and individual class accuracy is indicated in their confusion matrices in Fig. 4. Results show that multiple kernel sizes is an important factor in improving performance of CNNs. Additionally, proposed nets improve upon their counterparts for each of the rarer DR stages 1-4. This is extremely valuable in terms of being able to more accurately identify the DR signs. Furthermore, the mean average class accuracy for CKML Net is 47.3% compared to 40.9% for GoogLeNet, which is a significant increase. Mean average class accuracy for VNXK is 51.2% compared to 48.4% for VGGNet

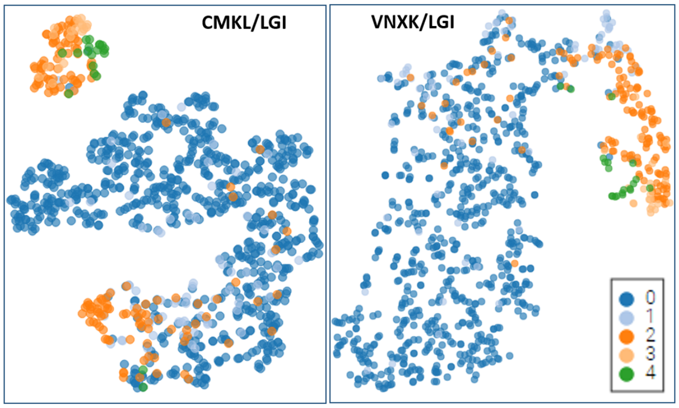

Fig. 5. t-SNE [31] visualization of test image set from 5-DR classes on EyePACS dataset. Clustering results are generated by taking 5 dimensional features of a given image from the last fully connected layer of CKML Net/LGI (left) and VNXK/LGI (right).

Table II. Comparison of the Classification Performance (ROC) of Proposed Methods with Other Methods for DR Screening on Messidor Dataset

Screening	Method	ROC	Sens.	Spec.	Acc.
Normal/ Abnormal	Sánchez et al. (2011) [14]	0.876			
	Tang et al. (2013) [13]	0.870			
	CKML Net/LGI	0.862	0.916	0.803	0.858
	VNXK/LGI	0.870	0.882	0.857	0.871
Referable/ Nonreferable	Pires et al. (2015) [12]	0.863			
	CKML Net/LGI	**0.891**	0.893	0.900	0.897
	VNXK/LGI	**0.887**	0.900	0.892	0.893

Sens. - Sensitivity; Spec. - Specificity; Acc. - Accuracy

Additionally, we use t-Distributed Stochastic Neighbor Embedding (t-SNE) [31] for visualization of results. t-SNE is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. Fig. 5 shows t-SNE visualization of test image set from 5-DR classes on EyePACS dataset. Clustering results are generated by taking 5 dimensional features of a given image from the last fully connected layer of CKML Net/LGI and VNXK/LGI. Results indicate that classes are better separated in VNXK/LGI than CKML Net/LGI.

Both t-SNE visualization and confusion matrices reinforce that identifying mild DR (stage 1) is the most difficult task in DR recognition. Although, mild DR class has 2.5 times more training images than severe and NDPR stages, it performs far worse than those other rarer classes. Based on the confusion matrix, small filter VGGNet performs far better than GoogLeNet on mild DR stage. Furthermore, multiple filter sizes in both VNXK and CKML Net outperform the original nets with single filter size by approximately 10%. This result is encouraging and it seems there is space for some extra filters of middle sizes to further improve performance of this class.

B. Performance of Proposed Nets and LGI Color Space on Messidor Dataset

The second set of experiments re-evaluates the proposed nets on the proposed LGI color space for DR screening on Messidor dataset. Although the Messidor might not reflect the proper distribution of DR-stages in real-world, it is commonly used to evaluate different methods on DR screening. As Messidor is too small for convolutional neural networks, the proposed net CKML Net/LGI and VNXK/LGI with the parameters that was learned for 5-class DR recognition in the previous experiment set are applied on Messidor dataset to extract features. Binary classification on extracted features is done by linear SVM.

Five dimensional feature vectors are extracted from the last fully connected layer of each proposed net. To be compatible with referable/non-referable DR screening of [12], after extracting features from Messidor, 10-fold training and testing are conducted for referable/non-referable classification. For normal/abnormal screening task, extracted feature from EyePACS are used for training. The trained SVM is tested on the extracted features of the entire Messidor dataset. Table II shows receiver operating characteristic (ROC), sensitivity, specificity, and accuracy of each experiment and compares results with prior work.

For normal/abnormal screening, our nets perform close to prior works [14], [13]. Normal/abnormal screening could be improved by improving accuracy of mild DR stage as discussed in previous experiment set. For referable/non-referable screening, CKML Net/LGI and VNXK/LGI achieve 0.891 and 0.887 respectively for ROC, and they both surpass the recorded ROC in [12] for Messidor.

VI. Conclusion and Future Work

In this paper, we have proposed two convolutional neural networks: CKML Net and VNXK with multiple filter sizes, which are an improvement upon GoogLeNet and VGGNet. Additionally, we have introduced hybrid color space, LGI, for DR recognition via convolutional neural networks. The proposed nets and the hybrid color space prove their discriminating power for DR recognition both on EyePac and Messidor datasets for DR screening.

We also applied transfer learning in existing nets and proposed nets to overcome the issue of highly imbalanced class sizes. There exist other techniques such as cost-sensitive and kernel-based classifiers that could be combined with transfer learning to further boost the performance of our current methodology. Thus, exploring these techniques is the future direction of our work.

References

[1] Y. LeCun, K. Koray, and F. Clement, “Convolutional networks and applications in vision,” *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, Paris, France, 2010.

[2] C. Szegedy et al., “Going deeper with convolutions,” *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015.

[3] S. Arora et al., “Provable bounds for learning some deep representations,” *Int. Conf. on Machine Learning (ICML)*, Beijing, China, 2014.

[4] M. Lin, C. Qiang, and Y. Shuicheng, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[6] K. He et al., “Deep residual learning for image recognition,” arXiv preprint arXiv:1512.03385, 2015.

[7] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 9, pp 1263-1284, 2009.

[8] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *Int. Conf. on Machine Learning (ICML)*, Lille, France, 2015.

[9] Y. Song, M. Louis-Philippe, and D. Randall, “Distribution-sensitive learning for imbalanced datasets,” *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 2013.

[10] (2015, February 17). [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[11] MESSIDOR, TECHNO-VISION, “MESSIDOR: methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology,” Available on: http://messidor. crihan. fr/index-en. php, Accessed: October 9, 2014.

[12] R. Pires et al., “Beyond lesion-based diabetic retinopathy: a direct approach for referral,” *IEEE J. of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1-1, 2015.

[13] L. Tang et al., “Splat feature classification with application to retinal hemorrhage detection in fundus images,” *IEEE Trans. on Medical Imaging*, vol. 32, no. 2, pp. 364-375, 2013.

[14] C. I. Sánchez et al., “Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data,” *Investigative Ophthalmology and Visual Science*, vol. 52, no. 7, pp. 4866-4871, 2011.

[15] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. on Computer Vision*, 2015

[16] A. Sinha, S. Banerji, and C. Liu, “New color GPHOG descriptors for object and scene image classification,” *Machine Vision and Applications*, vol. 25, no. 2, pp. 361-375, 2014.

[17] Y. Jia et al., “Caffe: Convolutional architecture for fast feature embedding,” arXiv preprint arXiv:1408.5093, 2014.

[18] NVIDIA DIGITS Software. (2015). Retrieved April 23, 2016, from https: //developer.nvidia.com/digits.

[19] B. Antal and A. Hajdu, “An ensemble-based system for microaneurysm detection and diabetic retinopathy grading,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1720-1726, 2012.

[20] O. Faust et al., “Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review,” *J. Medical Syst.*, vol. 36, no. 1, pp. 145-157, 2012.

[21] J. Lachure et al., “Diabetic retinopathy using morphological operations and machine learning,” *IEEE Int. Advance Computing Conf. (IACC)*, Bangalore, India, 2015.

[22] C. Sinthanayothin et al., “Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images,” *British J. Ophthalmology*, vol. 83, no. 8, pp. 902-910, 1999.

[23] U. R. Acharya et al., “Computer-based detection of diabetes retinopathy stages using digital fundus images,” *Proc. Institution of Mechanical Engineers, Part H: J. Engineering in Medicine*, vol. 223, no. 5, pp. 545-553, 2009.

[24] K. Ram and S. Jayanthi, “Multi-space clustering for segmentation of exudates in retinal color photographs,” *Annu. Int. Conf. of the IEEE Eng. in Medicine and Biology Society (EMBC)*, Minneapolis, MN, 2009.

[25] W. L. Yun et al., “Identification of different stages of diabetic retinopathy using retinal optical images,” *Inform. Sci.*, vol. 178, no. 1, pp. 106-121, 2008.

[26] M. R. K. Mookiah et al., “Computer-aided diagnosis of diabetic retinopathy: A review,” *Comput. in Biology and Medicine*, vol. 43, no. 12, pp. 2136-2155, 2013.

[27] T. Walter and J. Klein, “Segmentation of color fundus images of the human retina: Detection of the optic disc and the vascular tree using morphological techniques,” *Int. Symp. on Medical Data Analysis (ISMDA)*, Madrid, Spain, 2001.

[28] J. Nayak et al., “Automated identification of diabetic retinopathy stages using digital fundus images,” *J. Medical Syst.*, vol. 32, no. 2, pp. 107-115, 2008.

[29] C. Liu and H. Wechsler, “Robust coding schemes for indexing and retrieval from large face databases,” *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 132-137, 2000.

[30] S. Banerji et al., “Novel color LBP descriptors for scene and image texture classification,” *Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, Las Vegas, NV, 2011.

[31] L.J.P. van der Maaten, “Accelerating t-SNE using Tree-Based Algorithms,” *J. of Machine Learning Research*, vol. 15, pp. 3221-3245, Oct. 2014.

[32] R. Rao, S. K. Bharat, and N. R. Stefan, “Data mining for improved cardiac care,” *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp: 3-10, 2006.

[33] D. Mease, A. J. Wyner, and A. Buja, “Boosted classification trees and class probability/quantile estimation,” *J. of Machine Learning Research*, vol. 8, no. , pp. 409-439, May 2007.

[34] J. Yosinski et al., “How transferable are features in deep neural networks?,” *Annu. Conf. on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014.

[35] H. H. Vo and A. Verma, “Discriminant color texture descriptors for diabetic retinopathy recognition,” *IEEE Int. Conf. on Intelligent Computer Communication and Processing (Track: Computer Vision) (ICCP)*, Cluj-Napoca, Romania, 2016. (to appear)