Diabetic Retinopathy Identification and Severity Classification

Sagar Honnungar, Sanyam Mehra, Samuel Joseph
{sagarh, sanyam, josamuel}@stanford.edu

***Abstract*—Manual examination of retina images for the diagnosis of diabetic retinopathy is a time consuming and error prone process, requiring identification of inconspicuous anomalies like micro-aneurysms and exudates. In this work, we explore machine learning techniques for automatic identification and severity classification of diabetic retinopathy from retina images. The presented approach involves image pre-processing, feature extraction using the bag of visual words model, and a multi-class classifier to classify the image into different DR stages. We have considered SURF, LBP and HoG features for constructing the bag of visual words. For the multi-class classification, we have implemented multinomial logistic regression, SVM and random forests.**

I. Introduction

DIABETIC Retinopathy (DR) is one of the most frequent causes of visual impairment in developed countries and is the leading cause of new cases of blindness in the working age population. Altogether, nearly 75 people go blind every day as a consequence of DR [6]. Effective treatments for DR require early diagnosis and continuous monitoring of diabetic patients, but this is a challenging task as the disease shows few symptoms until it is too late to provide treatment.

Currently, diagnosis of DR is performed by manual evaluation of retinal images by expert clinicians who identify presence of lesions in the eye such as micro-aneurysms (red lesions), hemorrhages and exudates (bright lesions). This turns out to be a slow and demanding process. Further, the expertise and equipment required for such evaluation may be lacking in many areas with a large DR affected population.

Because of the aforementioned reasons, it is evident that an automated system for DR detection of color fundus images could have a huge impact in making timely treatment accessible to more patients. Towards this end, we have explored machine learning techniques to automatically detect the severity of DR using information from retina images in this work. Our approach is to use a bag of words model with local feature descriptors such as SURF, LBP and HoG to extract important features from the image, which are fed into a multi-class classifier that predicts the severity of DR in the eye as Class 0 (No DR), Class 1 (Mild DR) and Class 2 (Severe DR). Fig. 1 shows example images of each class.

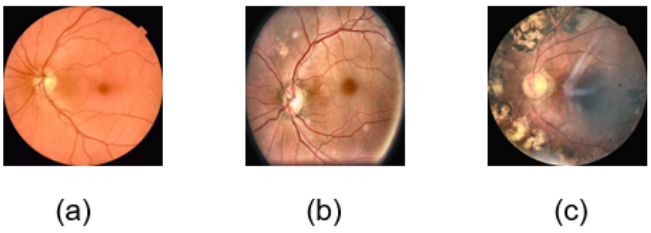

Fig. 1. Example images of different stages of DR: (a) No DR (b) Mild DR (c) Severe DR. These images also show the variance in image characteristics in the dataset—the images have different dimensions, positioning of retina and color profile

Note that only a highly trained and experienced doctor can classify these images accurately – even the authors could not classify most images accurately with their eyes. Hence this is a highly challenging image classification problem, when compared to other image classification problems like classifying animals or identifying sign language.

II. Related Work

Kaggle conducted a competition on the same problem in 2015 [2]. All the top-performing Kaggle teams used sophisticated neural network models that require many days of training on high-end GPUs. [5], [3], [4]. Instead of trying to replicate those methods, we focused on trying to come up with simpler innovative method that can give comparable results. We focus on pre-processing the images and feature engineering using traditional image processing techniques and utilise classifiers to solve the multi-class classification problem.

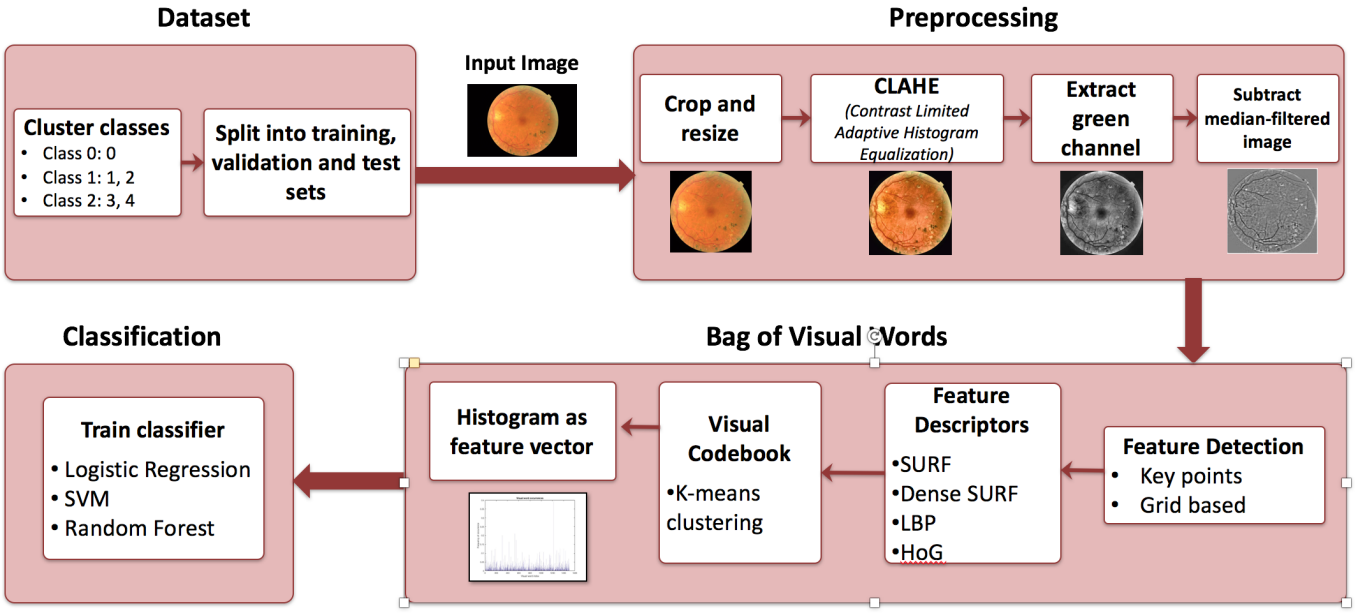

Fig. 2. Entire methodology pipeline.

Other than the well-known Kaggle competition, there has been other work done on DR detection/classification [8], [9]. However, most of these approaches have utilised smaller datasets (for example, [9] uses 94 images), or other metadata (like patient history).

III. Dataset

We use the dataset provided by Kaggle [2]. It contains images belonging to 5 different classes – normal (0), mild (1), moderate (2), severe (3) and proliferative (4). For the purpose of evaluating our algorithm, we cluster these into three categories:

- Class 0 – This category corresponds to healthy eyes (0).
- Class 1 – This category corresponds to moderate retinopathy (1 and 2). In other words, these patients do not have a very bad prognosis.
- Class 2 – This category corresponds to patients who have severe retinopathy (3 and 4) and require immediate medical attention.

IV. Methodology

The proposed method uses the bag of words approach to classify the images into different stages of DR. The pipeline involves image pre-processing, feature extraction, codebook generation and classification. Each of these steps is described in detail below.

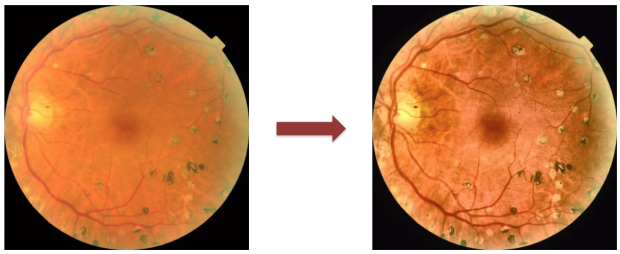

Fig. 3. CLAHE applied to a Category 3 image. It can be seen that the contrast in the image has increased which makes the exudates more visible.

A. Image Pre-processing

The dataset has images taken using different types of cameras, which results in differences in visual appearance and other image parameters. According to the source, there is also noise in both the images and labels. Some of the images may contain artifacts, be out of focus, underexposed, or overexposed.

Because of these factors, it is essential that we pre-process and standardize the dataset to a certain extent before extracting any information from the images for classification.

Firstly, we crop out the black background in each of the images and ensure that the eye is centered and occupies the maximum area in the image. After cropping, we also resize the images to a uniform size of 448×448 pixels.

Secondly, we observed that the contrast tends to

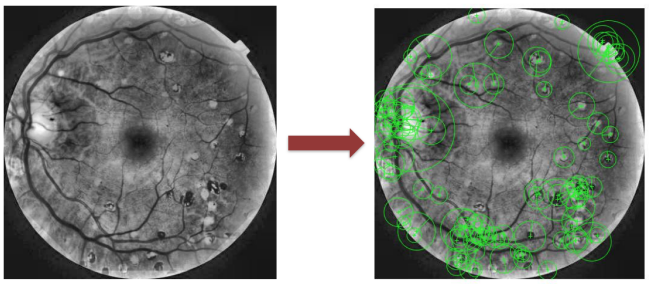

Fig. 4. Feature extraction based on keypoints applied on a pre-processed image. Note that the exudates are being captured as keypoints.

diminish toward the edge of the images. Contrast enhancement is therefore vital to clearly distinguish and identify important features indicative of DR. For this, we use Contrast-Limited Adaptive Histogram Equalization (CLAHE) [1]. CLAHE operates on small regions of the image and improves their contrast by transforming the intensity through localised histogram equalization. After performing histogram equalization in small regions, the neighboring regions are combined using bilinear interpolation. Figure 3 shows application of CLAHE to one of the images in the dataset.

Thirdly, we extract only the green channel from the image for further processing as it has been observed that the lesions are most observable show the largest contrast in the green channel and hence are most easily identifiable in this channel.

Finally, we subtract out the background image estimated by a median filter. This step ensures that common prominent features present in all eye images that are not indicative of DR, are subdued and only the relevant differentiating features are accentuated. [10]

B. Feature Extraction

Accurately capturing local information about the lesions present in images with DR such as micro-aneurysms, hemorrhages and hard exudates requires careful selection of features. We have used three different feature descriptors which capture the local information in different ways, and assessed their relative performance in our experiments.

1) SURF features: Speeded Up Robust Features (SURF) is a popular algorithm developed for detecting and describing local features of images. It captures keypoints in the image and provides a "feature description" of the image using local features of these keypoints, also known as keypoint descriptors. The algorithm selects keypoints using the Hessian blob detector and the feature descriptors are obtained using the sum of the Haar wavelet response around the interest point. For each interest point, SURF descriptor vector of length 128 is obtained. So the dimension of SURF descriptors per image is: number of interest points $\times$ 128.

2) HoG features: Histogram of Oriented Gradients (HoG) is another popular feature descriptor for object detection. It counts the number of occurrences of different gradient orientations in localized parts of the image. The image is divided into blocks of size 32×32 and a histogram of size 31 is computed for each block. Hence the dimension of HoG descriptors for each image is: number of blocks $\times$ 31

3) LBP features: Local Binary Patterns (LBP) have been found to be powerful as local descriptors for texture classification. An LBP descriptor is a string of bits, one for each neighborhood pixel, where each bit is 1 or 0 depending on whether the corresponding neighborhood pixel has greater intensity than the central pixel. LBP features are captured from local image patches of size 32×32, similar to HoG, but the feature length for each patch is 58. Hence the feature matrix for each image is of size: number of patches $\times$ 58.

C. Visual Codebook Generation

After capturing the features from each image, a dictionary is constructed using K-means clustering from the pool of all image features. This generates a visual codebook containing words which are the cluster centroids resulting from K-means clustering. Then, each image feature is quantized to the nearest word in the codebook. This results in a histogram for each image which counts the number of features which are closest to each visual word in the codebook. The resultant image histogram is the feature vector which we use as input for our classifiers.

D. Classification

Once the distinctive features (image histograms generated from the visual codebook) have been extracted from the fundus images, the stage is set to train a classifier that can accurately identify different classes of DR severity using these features. For

TABLE I
TRAIN AND TEST ACCURACIES WITH DENSE SURF AND VOCABULARY SIZE 100

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.85	0.68
Random Forest	0.90	0.73
SVM	0.83	0.72

TABLE II
CONFUSION MATRIX WITH DENSE SURF, SVM AND VOCABULARY SIZE 100

Known \ Predicted	Class 0	Class 1	Class 2
Class 0	0.62	0.26	0.12
Class 1	0.38	0.48	0.14
Class 2	0	0.03	0.97

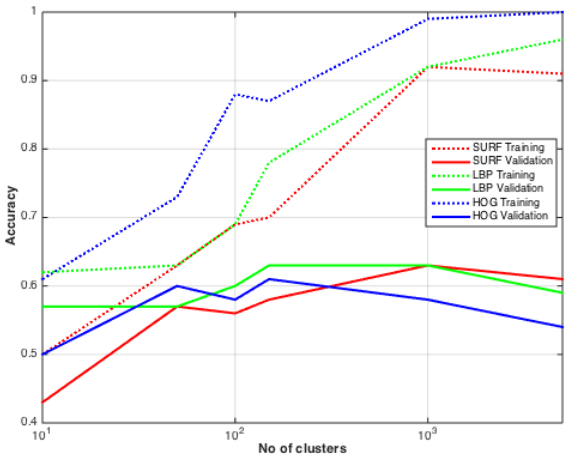

Fig. 5. Performance of SURF, LBP and HOG with SVM for varying vocabulary size (number of clusters).

this multi-class classification problem, we have used three different classifiers:

1) Multinomial Logistic Regression: The first model which we tried as a baseline was logistic regression. Since we have more than two classes, we consider the natural extension of logistic regression which is multinomial logistic regression, that uses the softmax function for predicting the probabilities of each class. It was implemented using mnrfit() and mnrval() functions in MATLAB.

2) Support Vector Machines: The optimal margin classifier with L1 regularization is used:

$$min \frac{1}{2}||w||^2 + C\sum_i \zeta_i \quad (1)$$

$$\text{s.t. } y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i \quad (2)$$

The value of C is tuned using cross-validation. We have implemented SVM with a linear kernel:

$$K(x, z) = x^T z \quad (3)$$

3) Random Forests: This is an ensemble learning method which uses decision trees. Random Forests build a number of decision trees from new datasets that were sampled with replacement from the original one (bootstrapping) and predict a class by taking the trees majority vote. They also restrict the features considered in every split to a random subset of a certain size.

V. RESULTS AND DISCUSSION

We have 400 images of each category (class 0, class 1 and class 2). We used 40% of the images as our training set for generating the bag of visual words. Then we used hold-out cross-validation to tune the number of clusters in the K-means clustering algorithm, which corresponds to the vocabulary size of our codebook. We found that $K = 100$ is the optimal number of clusters based on this analysis. Further, we also found that all three feature descriptors (SURF, LBP and HoG) gave similar accuracy results. Fig. 5 shows the results of this experiment. Hence, for the performance analysis of different classifiers, we have considered SURF features with a vocabulary size of 100 codewords.

Table I shows the accuracy of different classifiers on the training and test set. For SVM, the optimal value for C was found using cross-validation to be 8. Similarly, the hyperparameters for the Random Forest model were tuned as follows based on validation errors: number of trees $= 10$ and minimum leaf size $= 8$.

Table II shows the confusion matrix for SVM using a vocabulary size of 100 and SURF features. It is observed that class 2 is classified correctly with very high accuracy (97%) whereas class 1 is misclassified as class 0 quite often. This suggests that we can ascertain cases of severe DR without manual intervention successfully using the proposed method. This could therefore substantially reduce the number of images requiring manual screening. Misclassifying class 0 as class 1 or 2 is incorrect, but not harmfully consequential for the patient. On the other hand, misclassifying class 2 as class 1 or class 0 is extremely unacceptable. The results

obtained are in line with this understanding, as is also observable from the confusion matrix.

VI. Conclusion and Future Work

In this project, we developed a classification method using the bag of words model for automatically diagnosing the severity of diabetic retinopathy. We were able to achieve high accuracy in detecting cases of severe DR and believe that this method could be very useful as an initial screening step to augment and speed up the manual process of detection and also increase confidence and reduce misclassification due to human error.

In addition, we were specifically looking at methods other than deep neural networks to satisfy our curiosity and for learning purposes. For image classification problems like this, convolutional neural networks (CNNs) have been shown to perform very well. To extend the project without using CNNs, other image processing techniques like Moat Operator [7] and recursive region growing segmentation (RRGS) algorithm can be used for better detection of exudates and other lesions. To improve classification accuracy without utilizing neural networks, we could combine different classifiers through an ensemble approach.

VII. Acknowledgements

The authors would like to thank Mike Chrzanowski (Baidu) and Darvin Yi (Stanford) for introducing us to the nuances of the problem and for helping us in making a great start. The authors also thank the project TA Bo Wang for his feedback and suggestions throughout the course of the project. Lastly we thank Prof. Andrew Ng and Prof. John Duchi for their guidance and support.

References

[1] Adaptive histogram equalization - wikipedia. https://en.wikipedia.org/wiki/Adaptive_histogram_equalization. (Accessed on 12/16/2016).

[2] Diabetic retinopathy detection — kaggle. https://www.kaggle.com/c/diabetic-retinopathy-detection. (Accessed on 12/16/2016).

[3] Diabetic retinopathy winners interview: 4th place, julian & daniel — no free hunch. http://blog.kaggle.com/2015/08/14/diabetic-retinopathy-winners-interview-4th-place-julian-daniel/. (Accessed on 12/16/2016).

[4] Diagnosing diabetic retinopathy with deep learning — deepsense.io. https://deepsense.io/diagnosing-diabetic-retinopathy-with-deep-learning/. (Accessed on 12/16/2016).

[5] Machine learning for diabetic retinopathy detection — alexander rakhlin — linkedin. https://www.linkedin.com/pulse/machine-learning-diabetic-retinopathy-detection-alexander-rakhlin. (Accessed on 12/16/2016).

[6] Bálint Antal and András Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-Based Systems*, 60:20–27, 2014.

[7] Jyothis Jose and Jinsa Kuruvilla. Detection of red lesions and hard exudates in color fundus images. *International Journal Of Engineering And Computer Science*, 3:8583–8588.

[8] Kwang Baek Kim. Extraction of canine cataract object for developing handy pre-diagnostic tool with fuzzy stretching and art2 learning. *International Journal of Fuzzy Logic and Intelligent Systems*, 16(1):21–26, 2016.

[9] R Priya and P Aruna. Diagnosis of diabetic retinopathy using machine learning techniques. *Journal on Soft Computing*, 3(4):563–575.

[10] Ibrahim Sadek, Désiré Sidibé, and F Meriaudeau. Automatic discrimination of color retinal images using the bag of words approach. In *SPIE Medical Imaging*, pages 94141J–94141J. International Society for Optics and Photonics, 2015.