PILL RECOGNITION USING MINIMAL LABELED DATA

Yu Wang, Javier Ribera, Chang Liu, Sri Yarlagadda, Fengqing Zhu

School of Electrical and Computer Engineering, Purdue University

ABSTRACT

Inappropriate medication use such as wrong drug or wrong dose intake can be harmful to patients. In this work we present a method to automatically identify a pill from a single image using Convolutional Neural Network (CNN). We first localize the pill in the image by detecting the region with the highest concentration of edges. To overcome the challenge of minimal labeled training data and domain shift from the training images taken under the controlled lab environment to the consumer images taken under natural living conditions, several data augmentation techniques are applied on the Region of Interest to generate synthetic pill images for training the CNN. We adopted GoogLeNet Inception Network as our main classifier. Three GoogLeNet models with different specialties on color, shape and feature are trained on the augmented dataset. We evaluate our proposed method with a publicly available dataset provided by National Institute of Health that contains 1000 different pill classes.

Index Terms— pill recognition, convolutional neural network, image analysis

1. INTRODUCTION

Adverse Drug Events (ADE) are scenarios in which harm is caused by a medicine or inappropriate use of the drug [1]. Medication errors [2] such as wrong drug intake are a frequent cause of ADE [3]. Elderly people are more prone to misidentify medicines, which increases the risk of ADE. In the US, the Food and Drug Administration enforces under 21 C.F.R § 206 (2016) [4] that the imprint, size, shape and color must permit the unique identification of any drug product in oral dosage. The imprint is an identification code or symbol marked in the pill by means of embossing, debossing, engraving, or printing with ink. Hence, these characteristics can be used to unequivocally identify a pill, capsule, or tablet.

Online tools such as Drugs.com [5], Pillbox [6], and WebMD [7] can identify a pill by manually entering its shape, color or imprint. These web sites provide valuable information but are time consuming, require a skilled operator and are subject to human errors. In this paper, we focus on automatic pill identification systems. A possible application would allow end-users to identify an unknown pill from a single image taken from a mobile phone thus reducing user burden and human errors.

There has been significant amount of research on automatic pill identification systems. In many works, feature vectors representing each pill class are generated based on labeled data, a pill image is then matched to its closest class. In [8], the authors use color, shape, size, and texture features to classify pills or capsules. In [9], shape and color features are used to filter pill categories. Modified Stroke Width Transform (MSWT) [10] is used to segment the pill's imprint, and Two-step Sampling Distance Set (TSDS) is used to describe the imprint shape. In [11], shape, imprint, and color features are combined for pill classification. All these features are generated from the interior of a mask indicating the location of the pill. In [12], Scale Invariant Feature Transform (SIFT) and Multi-scale Local Binary Pattern (MLBP) descriptors on the gradient magnitude image are used as imprint features. However, all these methods were tested on pill images with uniform background and constant levels of noise, which allows perfect segmentation of the pill.

Recent breakthroughs of object detection in the Computer Vision area are largely powered by the explosive endeavors in Deep Learning, especially CNNs. Despite of the successes achieved in letter and digit recognition [13], CNN really caught researchers' eyes in 2012 [14] when CNNs were applied to a data set of about a million images from collected from 1,000 different classes and they topped the ImageNet Challenge by almost halving the error rates of the best competing approaches. Since then, CNN has been applied with great success to object detection, segmentation and recognition tasks [15], such as face recognition [16], self-driving cars [17], predestrian and gesture detection [18]. In [19], the authors conduct several visual classification experiments and concluded that deep learning with CNN should be regarded as the primary candidate in essentially any visual recognition task. However, image datasets are inherently biased [20]. It has been shown that performance degrades using supervised methods in proportion to the difference between the test and training input distribution [21]. Transfer learning methods were proposed to compensate dataset bias and address the problem of using Deep Learning with relatively small dataset [22, 23].

In this paper, we employ and experiment CNNs to classify pills images with minimal labeled data. We trained our models with a publicly available dataset consisting of 2,000 im-

978-1-5090-6549-3/17 $31.00 © 2017 IEEE
DOI 10.1109/BigMM.2017.61

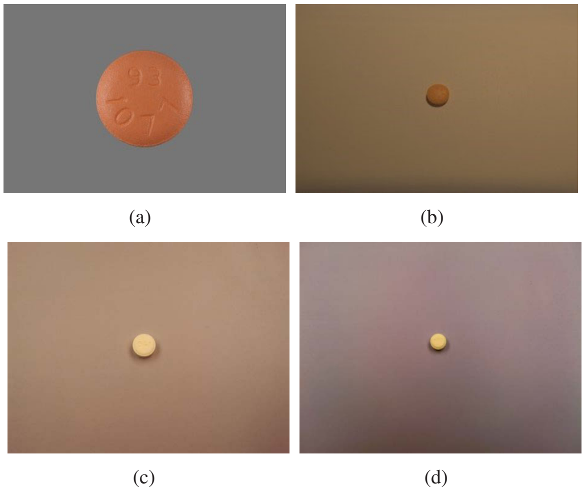

Fig. 1. (a) An example of training image. (b) Same pill as (a) in the consumer dataset. (c) and (d) are two examples of different pills in the consumer dataset.

ages of 1,000 different pills, each with one front-side and one back-side image taken under a controlled laboratory enviorment. The challenges come not only from the limited amount of training data versus large number of pill classes, but also from the subtle inter-class variations and the difference between training images and the consumer images which are taken under natural living conditions. 5000 consumer images provided by NIH Pill Challenge [24] were used to test the performance. As shown in Figure 1(a), all the training images are in the pure top-down view with uniform grey background. Figure 1(a) and (b) indicate the visual differences between a training image and one of its counterparts in the consumer dataset while Figure 1(c) and (d) imply that consumer images of different pills have very similar appearance. The consumer images are varied in noise level, scale, background, exposure, and lighting conditions. To our knowledge, there has not been other works that attempt to solve the pill recognition problem eithor using deep learning or the same pill dataset.

The remainder of the paper is organized as follows: Section 2 explains the method we employ to detect the ROI of the pill. Section 3 describes our dataset and data augmentation methods. Section 4 presents the proposed CNN structure for pill recognition. Section 5 shows the experimental results.

2. REGION OF INTEREST

Before the classification of a pill, we first extract its Region of Interest (ROI). We define ROI as a rectangular area that completely includes the pill. Ideally, a ROI should exclude as much background as possible, i.e, a rectangle circumscribing the pill. An example of a ROI is shown in Figure 2(d). As the goal of the ROI detection is only to speed up the pill classification, focus has been placed into ensuring that the ROI includes all the pill, at the expense of including some background. In addition, we choose the ROI to be squared, which the GoogLeNet Inception Network [25] we adopted in this paper requires.

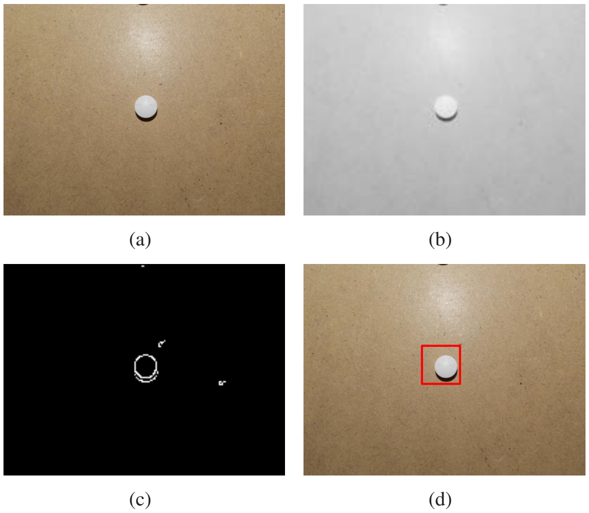

Fig. 2. (a) Original image. (b) Image after converting to gray scale and applying bilateral filter. (c) Response of Canny edge detection on (b). Note that some noise is still present. (d) Detected ROI with highest density (Equation (1)), with $P = 1$. No further steps are needed for this pill, as the ROI in (d) contains all the pill.

The ROI of a pill is obtained as follows: First, the image is converted to gray scale, and scaled to a fixed width of 200 pixels while keeping the width/height ratio. The scaling allows the ROI detection process to complete in around one second.

A bilateral filtering is applied to remove some background noise or texture, while keeping the edges of the pill. An example is shown in Figure 2(b). Next, Canny edge detection [26] is applied to obtain a binary mask indicating edge pixels. Figure 2(c) shows an example of this binary mask.

To detect the region of the image with highest concentration of edges, a bank of 20 squared ROIs are slided along the image. The minimum ROI size that we want is 256×256 pixels. Also, the ROI cannot be bigger than the input image. Thus, the length of the side of these squares, denoted as S, uniformly ranges between 256 and M pixels, where M is the minimum of the height and width of the input image. We define the edge density of a candidate ROI as

$$D = \frac{E}{S^{\frac{1}{P}}} \tag{1}$$

where E is the number of pixels contained inside the ROI that are classified as edge by the the Canny edge detector. $P > 0$ is a constant. In order to make sure that the ROI includes the

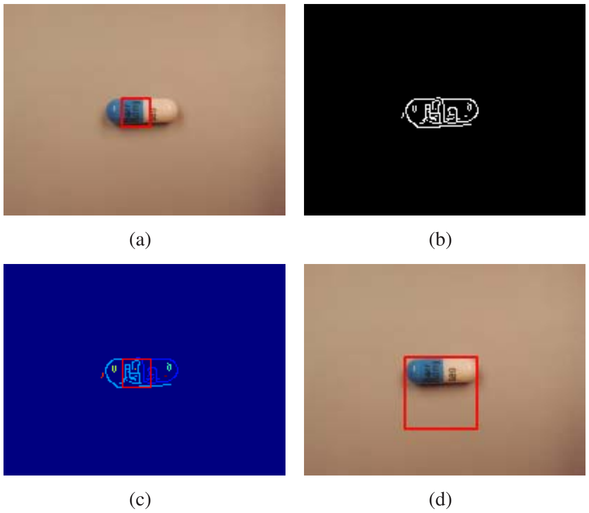
(a) (b) (c) (d)

Fig. 3. (a) Original image. The ROI with highest concentration of edges with $P = 1$ is shown with a red square. (b) Canny edge detection on the original image after applying bilateral filter. (c) Connected components of edge map. Each color shows a different connected edge. (d) Final detected ROI after the expansion.

pill, we set the value of P very high ($P = 4$) at the expense of occasionally include some unnecessary background inside the ROI. The spatial standard deviation of the bilateral filter is experimentally chosen to be 0.5. In the Canny edge detection, the lower and higher thresholds are experimentally set at 0.1 and 0.3, respectively.

The density D of all the candidate ROIs is computed, and the ROI with highest edge density is selected. A higher value of P tends to make the selected ROI larger.

In some cases, parts of the pill may remain outside the ROI (Figure 3(a)). To remediate it, we want to expand the ROI until it completely contains the pill. In order to do this, a connected components algorithm [27] is performed on the result of the edge map. Figure 3(b) shows the edges detected by the Canny operator. Figure 3(c) shows the connected components of the edges of Figure 3(b), and the region with highest density of edges. If the ROI does not fully contain the pill, some connected edges may cross the ROI. We expand the size of the selected ROI until no internal edge is connected to the outside of the ROI. Note that the ROI is expanded both horizontally and vertically because the GoogLeNet Inception Network expects squared images as input. The method does not generally position the pill in the center of the ROI. This is because different ROI positions may coincidentally have the same edge density (E). Figure 3(d) shows the ROI after the expansion.

3. DATA AUGMENTATION

For each selected ROI, we use CNN to identify pills in the consumer images. Training CNNs generally requires a large amount of labeled data. However, the training set only contains 2 images per pill class, which are the front and back views of a pill. As mentioned in Section 1, training images are acquired in a laboratory setting with a consistent top-down view of the pill on the same grey background. On contrary, consumer pill images are acquired in natural environments, where lighting condition, shooting angle, scale of a pill and background vary dramatically. The large visual difference between the training and consumer images, and limited training data impose a big challenge for recognizing pills in the consumer dataset. For our experiments, we used 1,000 consumer images as the validation set and the rest 4,000 images as the testing set.

Data augmentation has been shown to improve the performance of neural networks [28, 14]. Obviously, the training dataset represents an extremely biased subset of the whole pill image space. Nevertheless, we would like the network to learn the important features that are invariant for the pill categories, rather than the artifact of the training images. We propose to expand the training dataset by combining each pill region extracted from the given training image with backgrounds learned from a random subset of the consumer dataset. Here, we list several augmentation techniques used to generate synthetic pill images:

1. **Color casting:** we perform color casting to alter the intensities of the RGB channels in training images. A random scaling parameter p_r ranged from 0.7 to 0.95 is assigned to each color channel. When the red channel is altered, without loss of generality, the new intensity value, R_n, should be, $R_n = p_r R$, where R_n is the randomized parameter and R is the original pixel value. Color casting is used to simulated pill images in various lighting conditions.

2. **Projective distortion:** as we observed from the consumer dataset, most of the images are not in top-down views. Opposed to lab environment, objects in the consumer images have projective distortion due to camera shooting angles. To simplify the simulation of projective distortion, we assume a constrained homogeneous matrix, H and apply it to the pill region:

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a & b & 1 \end{bmatrix} \qquad (2)$$

where a and b are random numbers from -0.0002 to 0.0002.

3. **Gaussian filter, median filter:** consumer images usually have lower image quality compared to the training

images, especially for the pill region. Gaussian filters with random standard deviations from 0.2 to 0.8 and median filters with random radii are applied to simulate the blurriness in the consumer images.

4. **Random scale and position:** after we extracted the pill region from one training image, we randomly downscale the pill region and combine it with the background image at the randomized position, which is also dependent on the pill scale. In our implementation, we make sure that the synthetic image contains the whole pill with such random scale and position.
5. **Fixed rotation:** in the training dataset, all the pills are aligned horizontally with respect to their pivotal axis while in the consumer dataset, pills can be placed either horizontally or vertically with very slight tilted angle in the 2D plane of the surface where the pills are set. Thus, we use fixed rotations of 90 degrees, 180 degrees and 270 degrees.
6. **Background learned from the validation set:** from the validation set of the consumer images, we picked those images with relatively high resolutions and used the top left corners of the images (600 by 600 pixels) as the candidate background patches. For each training image, we extract the pill region, then apply all the augmentation techniques from 1-5 and finally combine the pill region with various background patches from the candidate set.

In the current implementation, we generated roughly 1500 synthetic images per pill category using the combination of all the techniques above. Figure 4 shows some examples of the synthetic pill images. Noticeably, the circular pill in Figure 4(a) is no longer a perfect circle due to the projective distortion. The imprint in Figure 4(b) is more subtle due to the random filters. Note that we regard the front and back images of a pill as the same pill category. Thus, we have around 1,500,000 training images in total.

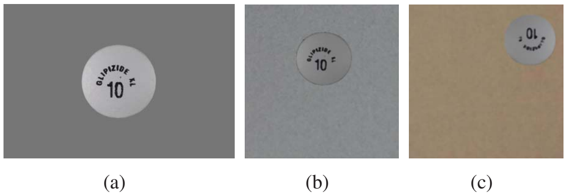

Fig. 4. (a) The original image, (b)(c) Synthetic pill images of (a).

4. DEEP NETWORKS

We adopted GoogLeNet Inception Network [25] as our primary classifier, because of its good performance in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [29]. All the source code related to Deep Network was developed in the Caffe framework. To obtain faster convergence and incorporate various visual features, we applied fine-tuning on three GoogLeNet models pretrained on ILSVRC data [29] to enforce different specialties, namely GoogLeNet-color model, GoogLeNet-shape model and GoogLeNet-feature model. GoogLeNet-color model predicts the color of a pill while GoogLeNet-shape model is in charge of determining the shape. GoogleNet-feature model is a 1000-category classifier trained directly on the augmented dataset, which is intended for general features, such as color, texture and imprint of a pill. The reason for using three models is to reinforce the basic features (color and shape) as pixel information flows deeper in the network, which also conforms to the idea of ensembling.

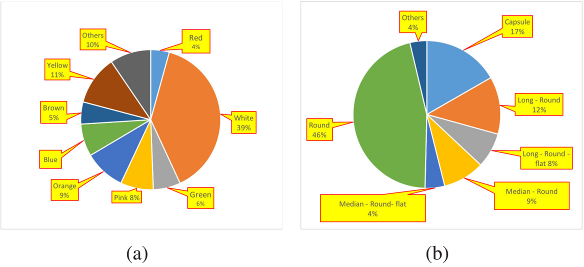

Fig. 5. (a) The distribution of color classes in the training set. (b) The distribution of shape classes in the training set.

To train GoogleNet-color model and GoogleNet-shape model, we used clustering techniques to generate several superclasses and applied crowdsourcing to refine those superclasses. As shown in Figure 5, both color and shape distributions of the training set are highly unbalanced. White and circular pills are the majority in the population. In Section 4.1 and 4.2, we explain the clustering techniques in detail. Section 4.3 discusses how we combine the decisions from multiple models.

4.1. Color Clustering

From observation, each pill only has one or two colors (excluding the imprint color). For each training image, we extracted two most dominant colors from the pill region and formed a 1 by 6 feature vector to represent the image:

$$[aR_1, aG_1, aB_1, bR_2, bG_2, bB_2]$$

where R_n, G_n and B_n represent the mean RGB value of the n^{th} most dominant color and a, b represents the ratio of the corresponding color segment occupied in the whole pill region. To obtain two dominant color segments, we applied K-means where K was set to 2. After obtaining 2000 feature vectors for all the training images, we apply K-means to the feature vectors to create initial color clusters, where

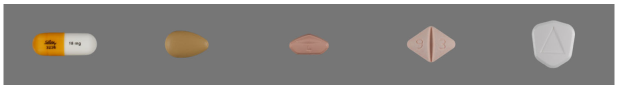

Fig. 6. From left to right, examples of capsule, bullet, free-form, diamond and shield.

K was set to 40. The initial color clusters were manually refined. Ten colors are used for the grouping: blue, brown, green, orange, pink, purple, red, white, yellow and near white. For most of these pills, the back of the pill has at least one color that is the same as the color of the front of the pill. For those pills where the front and back have different colors, their images were labeled as "weird". Other pill images are labeled as one or a combination of the ten colors. The color labels are listed as follows: blue, blue-blue, blue-brown, blue-green, blue-orange, blue-pink, blue-red, blue-white, blue-yellow, brown, brown-orange, brown-white, brown-yellow, green, green-black, green-purple, green-white, green-yellow, near white, orange, orange-red, orange-white, orange-yellow, pink, pink-purple, pink-red, pink-yellow, purple, purple-white, red, red-white, red-yellow, weird, white, white-yellow, yellow.

4.2. Shape Clustering

To cluster the training images into different shape categories, we first define the shape descriptor as an 8-bin normalized histogram, which is calculated by counting the number of pixels on each arc for every 45 degree starting at the intersection of the vertical line through the centroid of the pill and the pill contour. Then, we used K-means to generate 15 rough clusters and then manually corrected those images which were grouped into the wrong set. Similar to color clustering, refining clusters may require light load crowdsourcing for several rounds to ensure the clusters are perfect. Here, we list the shape categories: oval, long-round, capsule, rectangle, median-round-flat, shield, free-form, round, pentagon, long-round-flat, median-round, triangle, trapezoid, diamond and bullet (egg). Some examples are illustrated in Figure 6.

4.3. Classification and Decision Fusion

Given a testing image, the squared ROI is first obtained as discussed in Section 2. Next, the ROI is fed into three GoogLeNet models. Basically, every GoogLeNet model outputs the class predictions and corresponding confidence scores. Confidence scores from all three models are normalized to the range of 0 to 1 thanks to a customized softmax function added at the end of each network. The softmax function is defined as:

$$S_j^i = e^{\frac{y^i}{0.1max(y)}} \left(\sum_i e^{\frac{y^i}{0.1max(y)}} \right)^{-1} \tag{3}$$

where j represents different GoogLeNet model, y is the output from GoogLeNet and y^i is the i^{th} element of the array, y. For each prediction of a pill label, it has three confidence scores: $S_{feature}$, S_{color} and S_{shape}. Inspired by the unsupervised adaption and late fusion with linear interpolation discussed in [30], we combine the confidence scores in the following way:

$$S_{final} = S_{feature} + w_1 S_{color} + w_2 S_{shape} \tag{4}$$

where w_1 and w_2 represent the level of trust we place on the color and shape information reinforcement. In Section 5, we discuss how to set w_1 and w_2 and their effect on the classification performance.

5. EXPERIMENTAL RESULTS

The proposed method was evaluated using the dataset provided by the National Library of Medicine (NLM) under the Pill Image Recognition Challenge [31]. On January, 2016, the NIH announced the Pill Image Recognition Challenge [24], with the objective to encourage the development of software for pill image recognition. The provided dataset consists of two image sets. One is composed of 5000 consumer-quality pill images of varying noise level, scale, background, exposure, and lighting conditions. This image set is called consumer set. The other image set is called reference/training set and contains 2000 images of 1000 different pills, each with a front and a back image.

Region of Interest

In order to make sure that the ROI includes the pill, we cross validate P (see Equation 1) in a 5-fold manner and finally set $P = 4$ at the expense of occasionally include some unnecessary background inside the ROI. In this case, only in 23 of the 5000 images, the ROI does not completely include the pill, yielding an accuracy of 99.5 %. The spatial standard deviation of the bilateral filter is experimentally chosen to be 0.5. In the Canny edge detection, the lower and higher thresholds are experimentally set at 0.1 and 0.3, respectively. Since the ROI extraction is not the focus of this paper and the proposed method already reaches promosing result, we did not try to compare it to other localization methods, like Selective Search [32], region-based CNN [33].

Pill Recognition

We use two metrics to evaluate the performance of the proposed classifiers, namely Top N classification accuracy and Mean Average Precision (MAP) score. MAP score is used in the Pill Image Recognition Challenge as the performance measure. MAP score requires that a consumer-quality image has an unique rank for each reference image. It means that

Table 1. The baseline accuracy using only GoogLeNet-feature model

Accuracy	Trained from scratch	Fine-tuned
MAP	0.21	0.2933
Top 1	11.7%	19.52%
Top 5	23.1%	46.73%
Top 10	35.7%	56.34%
Top 20	46.3%	67.28%

the front and back image of the same pill class in the training dataset should have different ranks. We define a class rank matrix RC generated using the confidence scores for the pill classes. $RC^{i,j}$ ranks how well the consumer-quality image i is matched to the pill class j and $RC^{i,j} \in [1, 1000]$. $RC^{i,j} = 1$ means the consumer image i has the highest confidence score with the pill class j.

We define the front image rank matrix $RF^{i,j} = 2RC^{i,j} - 1$ and back image rank matrix $RB^{i,j} = 2RC^{i,j}$, where $RF^{i,j}, RB^{i,j} \in [1, 2000]$. $RF^{i,j} = 1$ and $RB^{i,j} = 2$ means the consumer image i is the best match to the pill class j.

MAP score can be then defined as follows:

$$S_{MAP} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\left(\frac{1}{RF^{i,k_i}} + \frac{2}{RB^{i,k_i}}\right)\right] \tag{5}$$

where S_{MAP} is the MAP score, N is the total number of consumer images, k_i is the ground-truth pill class that matches with the consumer image i. Here $N = 5000$. $S_{MAP} \in (0, 1]$ and $S_{MAP} = 1$ means all the consumer images are matched to the correct pill class. Note that the groundtruth table of the consumer dataset is also provided on the Pill Image Recognition Challenge website [31].

Each pill class on average has 5 consumer images. However, the distributions of pill colors and shapes are unbalanced as illustrated in Figure 5. As a result, there are lots of white circular pills in the consumer dataset. Statistically, 45.6% of the population are circular pills and 35.6% are pure white. With the shadows and noises in the image, it is hard for human to distinguish two white circular pills.

In the following experiments, we regard the GoogLeNet-feature model as our baseline. It basically has two branches: one is trained using only the augmented data and the other is fine-tuned based on the ILSVRC pre-trained model [25]. Table 1 shows the baseline accuracy of the GoogLeNet-feature model. Obviously, fine-tuning the ILSVRC pre-trained model proves to produce much better results, especially when the training data is extremely limited.

To finetune w_1 and w_2, we split the consumer dataset in a 20/80 manner, ie. we used 1,000 consumer images as the validation set and the rest 4,000 images as the testing set. Furthermore, we cross-validated the parameters by splitting the validation set into 5 folds. w_1 and w_2 converged to (0, 0.1) and (0.1, 0.2) respectively after 20 iterations of experiment.

Table 2. MAP scores for weights of confidence scores (color model weight: w_1, shape model weight: w_2). The best score is shown in bold.

MAP		w_2					
		0.1	0.12	0.14	0.16	**0.18**	0.2
w_1	0.02	0.321	0.322	0.323	0.323	0.324	0.324
	0.04	0.323	0.324	0.325	0.325	0.326	0.325
	0.06	0.324	0.325	0.326	0.326	**0.328**	0.327
	0.08	0.324	0.325	0.326	0.326	0.327	0.327
	0.1	0.323	0.324	0.325	0.325	0.326	0.326

Table 3. The accuracy for Top N predictions of pill class (Maximum accuracy is shown in bold).

Accuracy (%)	Top 1	Top 5	Top 10	Top 20
w1=0, w2=0	19.52	46.73	56.34	67.28
w1=0.18, w2=0.02	22.26	48.16	59.22	67.56
w1=0.18, w2=0.04	22.64	48.80	59.32	68.04
w1=0.18, w2=0.06	**22.86**	**49.06**	**59.92**	**69.52**

Finally, the proposed classifier was tested using the testing set. When only use the GoogLeNet-feature model, the MAP score is 0.2933. It increases to 0.327 when incorporating the color and shape models. As listed in Table 2, the MAP score peaks at $w_1 = 0.06, w_2 = 0.18$.

The accuracy for Top 1-20 prediction is also evaluated with different combinations of w_1 and w_2. Some of the best results are shown in Table 3. Compared with the baseline accuracy where $w_1 = w_2 = 0$, the Top N accuracy is consistently 2-3% better and nevertheless the Top 1 result is improved by $(22.86 - 19.52)/19.52\% = 16.8\%$ relatively. The proposed method of data augmentation and decision fusion exhibits a promosing result with minimal labeled data.

6. CONCLUSIONS

In this paper, we proposed a novel pill recognition system that benefits from deep learning and addressed the challenge of minimal labelled data and domain adaption. We have evaluated our method on a real dataset that contains pill images with noise, different backgrounds, poor lighting conditions, various resolutions, and points of view. We have achieved the MAP score of 0.328 for this dataset by implementing data augmentation and an ensembly of GoogLeNet models. Future extensions of this work will include developing dynamic weights for multiple CNN models, and extend the data augmentation techniques with a wider range of image manipulations such as JPEG distortions or creating artificial shadows. The ROI detection will also include the segmentation of the pill to completely ignore the background. Finally, we plan to develop a mobile app and a web service based on the proposed method for automatic pill recognition.

7. REFERENCES

[1] J. R. Nebeker, P. Barach, and M. H. Samore, "Clarifying adverse drug events: A clinician's guide to terminology, documentation, and reporting," *Annals of Internal Medicine*, vol. 140, no. 10, pp. 795–801, May 2004.

[2] "What Is a Medication Error?" http://www.nccmerp.org/about-medication-errors.

[3] M. Rockville, "Reducing and preventing adverse drug events to decrease hospital costs," *Research in Action, Agency for Healthcare Research and Quality*, no. 1, March 2001.

[4] "Imprinting of solid oral dosage from drug products for human use, 21 C.F.R § 206 (2016)," http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title21/21cfr206_main_02.tpl.

[5] "Drugs.com," https://www.drugs.com/pill_identification.htm.

[6] "Pillbox," https://pillbox.nlm.nih.gov.

[7] "WebMD," http://www.webmd.com/pill-identification/.

[8] R.-C. Chen, Y.-K. Chan, Y.-H. Chen, and C.-T. Bau, "An automatic drug image identification system based on multiple image features and dynamic weights," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 5, pp. 2995–3013, May 2012.

[9] J. Yu, Z. Chen, and S. i. Kamata, "Pill recognition using imprint information by two-step sampling distance sets," *Proceedings of the International Conference on Pattern Recognition*, pp. 3156–3161, August 2014, Stockholm, Sweden.

[10] Z. Chen and S. Kamata, "A new accurate pill recognition system using imprint information," *Proceedings of the SPIE International Conference on Machine Vision*, vol. 9067, pp. 906 711–906 711–5, April 2013, London, United Kingdom.

[11] J. J. Caban, A. Rosebrock, and T. S. Yoo, "Automatic identification of prescription drugs using shape distribution models," *Proceedings of the IEEE International Conference on Image Processing*, pp. 1005–1008, September 2012, Orlando, FL.

[12] Y.-B. Lee, U. Park, A. K. Jain, and S.-W. Lee, "Pill-id: Matching and retrieval of drug pill images," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 904 – 910, September 2012.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings of Advances in Neural Information Processing Systems*, pp. 1097–1105, December 2012.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, June 2014, Columbus, OH.

[17] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," *arXiv preprint arXiv:1202.2160*, 2012.

[18] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 648–656, 2015.

[19] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 512 – 519, June 2014, Columbus, OH.

[20] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1521–1528, June 2011, Providence, RI.

[21] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, pp. 137–144, 2007.

[22] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2252–2259, November 2011, Barcelona, Spain.

[23] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1785–1792, June 2011, Providence, RI.

[24] "Announcement of Requirements and Registration for Pill Image Recognition Challenge, 81 F.R § 2876 (2016)," pp. 2876 –2879, https://federalregister.gov/a/2016-00777.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[26] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, November 1986.

[27] P. A. Devijver, *Connected components in binary images: the detection problem*. New York, NY, USA: John Wiley & Sons, Inc., 1984.

[28] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *arXiv preprint arXiv:1501.02876*, vol. 7, no. 8, 2015.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[30] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell, "One-shot adaptation of supervised deep convolutional models," *arXiv preprint arXiv:1312.6204*, 2013.

[31] "Pill Image Recognition Challenge," 2016, https://www.nlm.nih.gov/news/nlm-pill-image-challenge-2016.html.

[32] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.