IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Automatic drug pills detection based on enhanced feature pyramid network and convolution neural networks

ISSN 1751-9632
Received on 21st March 2019
Revised 9th September 2019
Accepted on 7th October 2019
E-First on 20th January 2020
doi: 10.1049/iet-cvi.2019.0171
www.ietdl.org

Yang-Yen Ou[1] ✉, An-Chao Tsai[2], Xuan-Ping Zhou[1], Jhing-Fa Wang[1,3]

[1]*Department of Electrical Engineering, National Cheng Kung University, No. 1, University Rd, East District, Tainan, Taiwan*
[2]*Department of Computer Science and Entertainment Technology, Tajen University, No. 20, Weixin Rd., Yanpu Township, Pingtung, Taiwan*
[3]*Department of Digital Multimedia Design, Tajen University, No. 20, Weixin Rd., Yanpu Township, Pingtung, Taiwan*
✉ *E-mail: ouyang0916@gmail.com*

Abstract: Drug pill detection is one of the most important tasks in medication safety. The correct identification of drug based on the visual appearance is a key step towards the improvement of medication safety. Previous studies have aimed to recognise a drug based on the front or back view of the drug under a fixed viewing angle. In cases with multiple drugs and randomly placed drugs, the previous methods have difficulties in detecting and recognising different drugs in practical applications. A convolution neural network-based detector is proposed in this work to overcome the difficulties and to assist patients in drug identification. The proposed system includes a localisation stage and a classification stage. The enhanced feature pyramid network (EFPN), is proposed for drug localisation, and Inception-ResNet v2 is used in drug classification. The proposed Drug Pills Image Database contains a collection of 612 categories of drug datasets for deep learning research in the pharmaceutical field. The proposed EFPN achieves over 96% accuracy in the localisation experiment. In the complete system evaluation, the proposed system has obtained the Top-1, Top-3, and Top-5 accuracies of 82.1, 92.4, and 94.7%, respectively.

1 Introduction

Medication errors have claimed around six to eight thousands of lives each year [1]. Patients often have difficulties in distinguishing unpackaged drugs and wrongly-medicated themselves. Moreover, statistical research has estimated that 3–7% of intended medications is not used, which amount to around five billion dollars wasted each year in the United States [2, 3]. The improvement in medication knowledge and the provision of adequate drug information to patients have become important issues in eliminating wasted medications [4]. However, drug identification based on appearances is still a difficult task for the patients.

The main challenges in drug identification are the wide varieties and similar appearances of the drugs. For example, there may be as many as 700 categories of drugs in a medical centre, many of which may appear identical to the untrained eye. In general, drug appearances may be determined by the shape, colour and imprint. Drug shapes can be further divided into several types: round, capsule, oval, barrel, 4-sided, 6-sided and so on. Other differentiation features can also be observed with respect to the drug colours and imprints. In conventional drug identification, the user tries to determine an unknown drug by manually entering the drug features on the drug website, which is a time-consuming process. In previous studies, several problems have not been effectively resolved in drug recognition [5–9], such as the random placements of drugs and the presence of multiple drugs in an image. The angle of drug rotation is also difficult to determine and normalise for each drug category. Furthermore, it is difficult to perform automatic segmentation of the pills using traditional image processing.

In our previous research [10], the drug detection system utilises the Feature Pyramids Networks (FPNs) [11] and Xception [12] for drug localisation and drug classification. In this paper, we propose the Enhanced Feature Pyramid Networks (EFPNs) by coupling the FPN with Global Convolution Network (GCN) [13], which is used to increase the accuracy of drugs localisation. The present work also replaces the Inception-ResNet v2 with the Xception for drug classification and extends the drug image database. The Drug Pills Image Database (DPID) is constructed from images collected in the Kaohsiung Veterans General Hospital (KVGH). The database contains a total of 2,429,753 images under 612 drugs categories and includes the datasets for training, validation and testing.

The main contributions of this work are summarised as follows: (i) The EFPN is proposed to solve the problem of multiple drug localisation. (ii) The convolution neural network (CNN)-based classifier is used to solve the multi-category classification task with the variables, including the variations in light angles and drug rotations. (iii) A drug database, DPID, is collocated for deep learning and contains images with various attributes such as the variations in light angles and pill rotations, as well as having multiple pills in an image.

The organisation of the paper is as follows. Previous works on drug pill detection, drug database, and CNN-based object detector are discussed in Section 2. The proposed system and overview are described in Section 3. Section 4 provides the evaluation of the proposed system and the image database. Finally, the conclusions are provided in Section 5.

2 Related work

This section provides an overview of drug pill detection, drug databases, and CNN-based object detectors.

2.1 Automatic drug pills detection and drug database

Several studies on pill recognition systems have been proposed for the classification of drugs. In 2010, Hartl [14] used the shape and colour parameters as query features for an unknown drug. Lee *et al.* [15] considered the Scale-Invariant Feature Transform and Multi-Scale Local Binary Patterns for drug matching. In 2012, Caban *et al.*[16] modified the shape distribution to examine the shape, colour and imprint of drugs and created an invariant descriptor for drug recognition. The pill's imprint was described by the weighted shape context in [17]. Chen *et al.* [9] proposed a similarity measurement for the shape and colour of drugs, where the shape classification was applied using a simple neural network, and the colours of the drugs were transferred from the RGB colour space to the HSV colour space. Yu *et al.* [8, 18] proposed a modified stroke width transform to detect traces of the impression

© The Institution of Engineering and Technology 2019

for imprint features. Suntronsuk and Ratanotayanon [19, 20] proposed a K-means based clustering to extract the imprint text for drug classification. Neto *et al.* [5] proposed an invariant feature extractor based on the shape and colour of the drugs. According to the above literary works, the shape, colour, and imprint of a drug are the key characteristics for feature extraction. Previous research works have all focused on single drug classification based on the features at the topside of the pill, and the pill regions within an image have been determined manually or by limiting the background colour to black [5–9].

Zeng *et al.* [6] proposed the MoblieDeepPill, which used the Histogram of Oriented Gradient and Support Vector Machines (SVMs) for single drug localisation and the multi-CNN models were applied to collect pill characteristics. Wang *et al.* [7] used three GoogLeNet Inception models with different effects on the colour, shape, and feature, and decision fusion was used to combine those models. However, in [7], the positions of pill images were limited to rotations by 90°, 180°, and 270°. In the practical applications of a drug pill detection system, the aforementioned studies would have issues requiring further development, such as the determination of randomly positioned and rotated pills, as well as the number of pills in each detection task.

The drug images used in most studies have been collected from various websites. Some literary works [15–17, 20] used prescribed drug images from online pill databases or websites, such as 'drug.com', 'pharmer.org', the U.S Drug Enforcement Administration Office of Forensic Sciences, the U.S. National Library of Medicine and so on. Several studies constructed pill datasets for the experiments [8, 9, 18]. In [9], 526 drug images in 263 categories were acquired. Yu *et al.* [8, 18] collected 12,500 drug images in 2500 classes. Research works in [5–7] used the National Library of Medicine Image Database, NIH NLM PIR [21], which is a public database containing front and back views of 1000 drugs. The database contains 2000 high-quality images and 3000 consumer images. Although several drug databases have been collected in previous works, the overall coverage of the databases is insufficient for deep learning, especially with regard to the changes in lighting, rotation and drug angles for the different pill classes. The summary of the drug datasets is shown in Table 1.

2.2 CNN-based object detectors

The region-CNN (R-CNN) is an example of typical object detectors [22–26]. The R-CNN method uses selective search to extract region proposals and the linear SVM is used for classifications [22]. Faster R-CNN uses the Region Proposal Network rather than the selective search and utilises anchor boxes to solve the scale-variant problem [25]. The ROI-wise Reverse Reweighting Network [27] is built upon the Faster-RCNN and consists of a multi-layer pooling operation and ROI-wise reverse reweighting for different types of marking. Cheng *et al.* combined the additional multi-angle anchors with the RPN to address the rotational variations [28]. The FPN [11] was added to the backbone for multiscale localisation in the Mask R-CNN [26]. RefineDet [29] included the anchor refinement and the object detection modules, and the transfer connection block was proposed to increase the feature map for object detection.

Yuan *et al.* [30] proposed the Vertical Spatial Sequence Attention (VSSA) Network for traffic sign detection, wherein the multi-resolution feature learning module and the VSSA module were provided to enhance the semantic of small-size objects, and to gain more context information on the object. The rotation-invariant layer and a Fisher discriminative layer were added to the existing object detection CNN to solve the issues of rotation variations, within-class diversity, and between-class similarity in [31, 32].

Visual Geometry Group Net (VGG-16/ VGG-19) [33] and deep residual network (ResNet-50/ ResNet-101) [34] are the common backbones for object detection. Both networks have prevailed in classification tasks of computer vision. Nevertheless, the VGG and the ResNet are unable to serve as classifiers for drug classification, but the ResNet is still a good FPN backbone for drugs localisation.

3 Proposed system

The system architecture of automatic drug pill detection is shown in Fig. 1. A two-stage architecture is adopted in the proposed system, which includes the localisation and classification stages. The drug locations are determined through the EFPN in the localisation stage. The drugs are classified as the foreground in the drug localisation stage, regardless of the shape, colour, and texture of the drugs. Both the coordinates and sizes of the drugs are estimated in the first stage. The proposed EFPN, which provides the localisation in different scales, is constructed for object locations estimation. Two fully convolutional models, regression and classification, are adopted to determine the coordinates and sizes of the drugs for object prediction. Non-maximum suppression (NMS) is applied to eliminate redundant bounding boxes.

Although the EFPN localisation model can detect the drug regions, the drug similarity and rotation still cause difficulties in the EFPN classification model. The rotational angle is difficult to define. Furthermore, the texture and imprinted text are inconsistent between different drugs. The CNN-based classifier effectively processes the texture details on the object's surface according to the training data and data augmentation. Therefore, the Inception-ResNet v2 is used for drug classification instead. The details are discussed in the following.

3.1 Enhanced feature pyramid network for object localisation

The proposed EFPN has been built based on the ResNet-50 backbone. Five convolutional layers, C1, C2, C3, C4, and C5, have been included in the ResNet-50. Three of the convolution layers, C3, C4, and C5 have been chosen as the baseline for the pyramid layers. The layers of the ResNet provide the hierarchical features, but the shallow layers are weak in providing the feature representation. Therefore, Lin *et al.* [11] have proposed the FPN for merging the semantics with the hierarchal features.

According to the experimental results on drug localisation, objects with larger ratios are often undetectable. The FPN is insufficient for drug localisation. The EFPN generates two layers by the GCN [13] to increase the receptive field of FPN. The details

Table 1 Summary of the drug databases, including the number of classes, the number of trainings/validations/testing datasets

Database	Reference	Classes	Training	Validation	Testing
NIH NLM PIR	[5–7, 21]	1000	2000	3000	—
captured by themselves	[8, 18]	2500	12,500	—	—
captured by themselves	[9]	263	526	—	—
captured by themselves	[10]	131	488,520	8280	[a]1680
collected form website	[15]	—	2116	—	—
collected form website	[16]	568	568	—	—
collected form website	[17]	2000	2500	>12,000	—
pill database 'drug.com'	[19, 20]	—	540	—	—
DPID	**this study**	**612**	**2,393,585**	**34,975**	[a]**1193**

The authors of [15, 19, 20] are focused on feature extractor on drug imprint, without classification task.

[a]Each image includes 5–8 drugs.

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

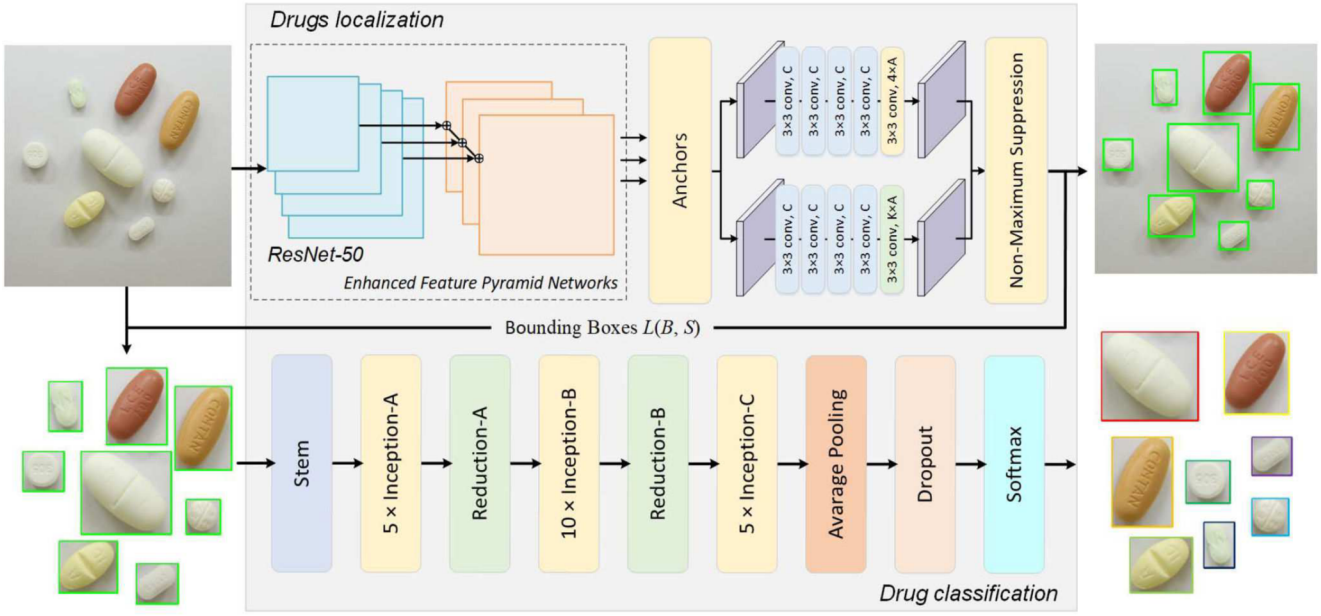

Fig. 1 *System overview of automatic drug pills detection. The proposed system contains a two-stage architecture, localisation, and classification. In the localisation step, the proposed EFPN and ResNet-50 are used for backbone building; two sub-models are applied for bounding box estimation, non-maximum suppression is utilised to merge the prediction results and reduce bounding boxes. Inception-ResNet V2 is used for drug classification*

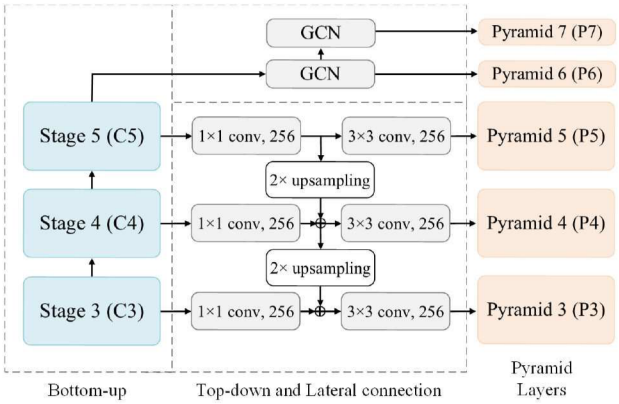

Fig. 2 *Architecture of EFPN for the pyramid layers generation*

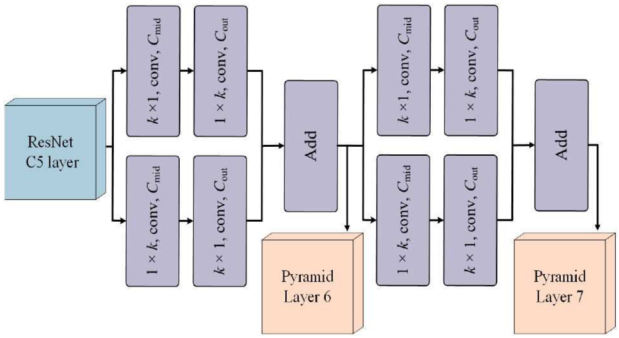

Fig. 3 *Generation of pyramid layers P6 and P7*

of the EFPN architecture are shown in Fig. 2. The bottom-up pathway, lateral connections, top-down pathway, and the GCN have been included in the proposed EFPN architecture. The bottom-up pathway is produced by the forward propagation of the ResNet-50 backbone. The backbone network reduces the size of feature maps through convolution layers with stride and pooling. The lateral connection is composed of 1×1 convolution layer with 256 channels, which are used to reduce the channel number and execute information fusion across the channels. The top-down pathway produces the feature maps with stronger semantics on the lower pyramid level, and the nearest-neighbours interpolation has been utilised to up-sample the top-down pathway. In the EFPN, the GCN has been used to build the pyramid layers P6 and P7 based on the C5 layer. The GCN increases the receptive field of the FPN for the detection of larger objects. The GCN factorises the $k \times k$ convolutions into $k \times 1$ and $1 \times k$ without the Rectified Linear Unit (ReLU) function. The generation of the pyramid layers P6 and P7 is shown in Fig. 3. The pyramid layers P3, P4, P5, P6, and P7 have been constructed by the EFPN with the ResNet-50 as the backbone. The pyramid layers are considered for object localisation with different scales and ratios. In addition, the pyramid layer P2 in the EFPN has been removed because the pyramid layers from P3 to P7 can cover the localisation task for inconsistent drug sizes.

The anchors have been utilised to generate the pre-selected boxes for regression and classification models in each of the pyramid layers. The anchors have been generated to increase the anchor density, through the multiple aspect ratios {1:2, 1:1, 2:1} and sizes $\{2^0, 2^1, 2^2\}$. The purpose of the anchors is to cover various appearances of the bounding boxes in each position of the feature map. In the proposed system, nine anchors have been created with different aspect ratios and sizes at each coordinate point. The Intersection-over-Union (IoU) ratio is used to filter the excess anchors, which compares the similarity between the anchor and the near ground-truth bounding boxes. The IoU formula is shown in the following equation:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

Subsequently, a positive label is assigned to an anchor if the IoU ratio exceeds 0.5 when compared to the ground-truth box and a negative label is given if the IoU ratio is less than 0.4; otherwise the anchor is abandoned. Both the regression and the classification of the bounding box are calculated when the anchor has a positive label. To exclude the irrelevant anchors in the proposed system, the search areas for the IoU calculation are set to 32^2, 64^2, 128^2, 256^2, and 512^2 on the pyramid layers P_3, P_4, P_5, P_6, and P_7, respectively.

To detect the object location on the pyramid levels, a classification model and a regression model have been utilised. The purpose of the classification model is to predict the object category of the anchors. The classification model contains five CNN layers. The first four layers are composed of 3×3 convolution layers with C channels and ReLU activation. The 3×3 convolution layers with $K \times A$ channels and linear activation are used as the output layers. The parameters K, C, and A, are the one-hot vector of the categories, the number of channels, and the number of anchors, respectively. The regression model is used to estimate the bounding

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

box $R_b^a\{b_x^a,b_y^a,b_w^a,b_h^a\}$ according to the anchor R_r^a and the ground-truth object $R_g^a\{g_x^a,g_y^a,g_w^a,g_h^a\}$. The regression model is similar to the classification model, the only difference is that the 3 × 3 convolution layers with linear activation and $4 \times A$ channels are considered as the output layer. With respect to the bounding box, the parameters x, y, w, and h, respectively, represent the x and y coordinates, the width, and the height. The anchor is represented as a, where $a = 1, 2, \ldots, A$. The regression model has been adopted to find a mapping $f(R^a) = R_b^a \simeq R_g^a$ and the parameterisation of bounding box [22] is taken as the reference in our regression model. In each anchor R_r^a, the predicted bounding box $R_b^a\{b_x,b_y,b_w,b_h\}$ is calculated by the relative offset t and the anchor box R_r^a. The predicted bounding box R_r^a is denoted as follows:

$$b_x = r_w t_x + r_x, \tag{2}$$

$$b_y = r_h t_y + r_y, \tag{3}$$

$$b_w = r_w \exp(t_w), \tag{4}$$

$$b_h = r_h \exp(t_h), \tag{5}$$

where the relative offset $t\{t_x,t_y,t_w,t_h\}$ is estimated by the regression targets of the nearby ground-truth bounding box $R_n^a\{n_x,n_y,n_w,n_h\}$ and the anchor box $R_r^a\{r_x,r_y,r_w,r_h\}$. The regression target of the ground-truth bounding box R_n^a is estimated by the anchor R_r and the ground-truth object $R_g^a = \{g_x,g_y,g_w,g_h\}$, and defined as follows:

$$n_x = (g_x - r_x)/r_w, \tag{6}$$

$$n_y = (g_y - r_y)/r_h, \tag{7}$$

$$n_w = \log(g_w/r_w), \tag{8}$$

$$n_h = \log(g_h/r_h). \tag{9}$$

To merge the predicted results of the pyramid layers, the non-maximum suppression (NMS) [35] is used to eliminate the overlap IoU. The prediction results $L(B, S) = \{b_n,s_n\}$ are sorted in descending order S, where n is the index of the anchor, with $n = 1, 2, \ldots, N$ and N is the number of the bounding boxes. NMS processing has two steps. In the first step, the bounding box which has the maximum confidence score M (b_m,s_m) is selected and moved to the localisations list, D. In the second step, the IoU is calculated between the score, M (b_m,s_m) and each bounding box of list $L(B,S)$; and the bounding boxes are excluded when the IoU is greater than the IoU threshold T_{IoU}. Finally, the prediction results $L(B,S)$ are evaluated by multiple iterations of the two steps.

The localisation network is constructed by the regression and classification models, which are used to predict the position $R_b\{b_x,b_y,b_w,b_h\}$ and the object categories $P = \{p_0,p_1,\ldots,p_{K-1}\}$. Both of the models are trained by a multi-task loss function $L_{\text{ML}}(u,t,R_b,P_c,Y_{\text{gt}})$, which is denoted in the following equation:

$$L_{\text{ML}}(u, t, R_b, P, Y_{\text{gt}}) = uL_{\text{loc}}(t, R_b) + L_{\text{cls}}(P, Y_{\text{gt}}) \tag{10}$$

where L_{loc} and L_{cls} are the regression loss and the classification loss, the parameter u reflects the positive (target) or the negative (background) labelling. R_n is the position of the near ground-truth object target. The one-hot vector $\boldsymbol{Y}_{\text{gt}}$ is the representation of the ground-truth class label. The regression model takes the standard $L1$-norm loss for the bounding box regression. The formula is denoted as follows:

$$L_{\text{loc}}(t, R_b) = \sum_{i \in \{x,y,w,h\}} \text{Smooth}_{L1}(t_i - b_i) \tag{11}$$

$$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}. \tag{12}$$

An extreme imbalance in the training phase may result since there are always more background samples than the foreground targets. The focal loss approach [36] is used to solve the imbalance problem through the down-weighting inlier. The focal loss approach consists of an improved cross-entropy CE (p,y) formula as shown in (13)

$$\text{CE}(p, y) = \text{CE}(p_t) = -\log(p_t), \tag{13}$$

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}, \tag{14}$$

where y specifies the ground-truth class, and p is the estimated probability for the calls with label $y = 1$. This approach adopts the deep neural network to focus on the samples that are difficult to predict, and the focal loss is generalised according to the commonly used cross-entropy by reducing the penalties from the well-classified samples. The focal loss approach adds a modulating factor $(1-p_t)^\gamma$ to the cross-entropy with an adjustable parameter γ. The equation for the focal loss approach is shown below:

$$\text{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t). \tag{15}$$

The focal loss samples with small losses can dominate the gradient when the accumulated classification loss exceeds a large number of inliers. The classification loss L_{cls} is defined by the sum of the focal loss samples over K classes, and the formula L_{cls} is shown in the following equation:

$$L_{\text{cls}}(P, Y) = \sum_{i=1}^{K} \text{FL}(p_i, y_i) \tag{16}$$

where i is the index of the category, p_i and y_i are the ground-truth class and the probability, respectively.

3.2 Inception-ResNet v2 for drug classification

The proposed system attempts to merge another CNN model for drug classification. To improve accuracy, the classifier needs to have strong imprint understanding capabilities. GoogLeNet [12, 37, 38] has received significant attention in the classification of a large number of categories by stacking inception modules and increasing the inception modules. Inception V3 [37], Xception [12], Inception V4 [38], Inception-ResNet v1 [38], and Inception-ResNet v2 [38] have all been evaluated with respect to drug classification.

Inception-ResNet v2 has been selected for the drug classification task because of its experimental performance. Stem block is used as the backbone, and three modules, named Inception-A, Inception-B, and Inception-C, have been used for feature map fusion in different scales. The architecture of Inception-ResNet v2 is described in the drug classification phase of Fig. 1. The stem and the three modules, Inception-A, Inception-B, and Inception-C are shown in Figs. 4*a* and *b*.

4 Experiments and discussion

The proposed automatic drug pills detection system is evaluated on the DPID. The drug pills detection has several challenges, including multiple categories, inconsistent sizes, similar appearances, and random rotation. In this section, we describe the database and experimental setting, as well as the reliability evaluation of the proposed system, followed by the time complexity analysis and a discussion of mistaken cases.

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

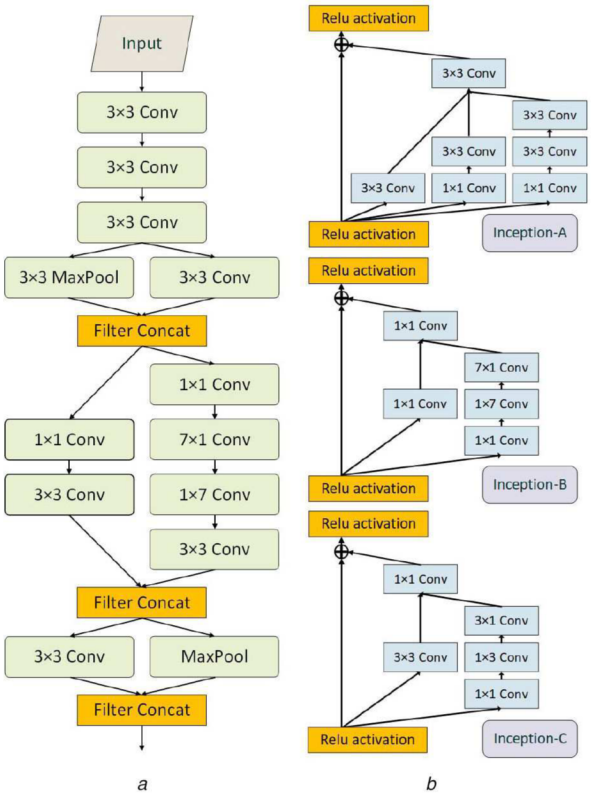

Fig. 4 *Modules of Inception-ResNet v2*
(a) Architecture of stem, ***(b)*** Those models, Inception-A, Inception-B, Inception-C, are represented from top to down

Fig. 5 *Examples of DPID*
(a) Samples with rotation factor on training and validation datasets, ***(b)*** Example with regard to changes in lighting and shot angles, ***(c)*** Samples of multiple drugs on testing dataset

4.1 Drug Pills Image Database

The DPID is collected through the cooperation with the KVGH, which has been presented primarily to support research in computer vision of pharmaceutical applications. To maintain the image quality and capture the imprint in the DPID, a digital single-lens reflex camera, Canon EOS 80D, has been used for data collection. Images with variations in the light angles, rotation, and drug angles have been taken into account in the DPID. 30–40 drug images have been captured from different angles for each category and the drug images are generated by rotating the images from 0° to 359°.

The DPID contains 612 categories of drugs from the KVGH, and three datasets have been independently collocated for the training, validation and testing purposes. The training and validation datasets have been used in the training phase of drug classification, and the testing dataset is used for the verification of our proposed system and the proposed EFPN. The training and validation datasets include over 3800 images in each category, with one single drug contained in each image. On the other hand, the testing dataset covers 1193 images with 8490 annotations, where each image includes 5–8 different drugs randomly selected from the 612 categories. Samples of the DPID are shown in Figs. 5*a–c*.

4.2 Implementation details

The drug pills detection system consists of two stages: the localisation stage and the classification stage. In the training phase of the localisation stage, the labels are either object (belonging to the foreground) or not (part of the background). For an unknown drug image, the drug category is determined by the Inception-ResNet v2 in the classification stage. The classification stage focuses on the classification task with a large number of drug categories. The proposed system includes two CNN models, and the parameters of the models are unrelated. In the case where both the shapes and colours of the drugs are covered in the drug localisation model, the drug localisation model does not need to be retrained when the number of drug categories for drug classification is increased.

All of the architectures are end-to-end trained by two NVIDIA 1080Ti GPU, and the Adam optimiser is used for the gradient descent. Four images have been used in the mini-batch with an initial learning rate of 10^{-4}. The learning rate can be adjusted by a value of 10^{-1} when the accuracy rates have not improved in more than five epochs during the validation phase. The proposed system works under Ubuntu 16.04 LST with Python version 3.6.4. Tensorflow-gpu 1.5.1 and Keras 2.1.6 have been used for the training and testing phases.

4.3 Evaluation and discussion

We have used the training, validation and testing datasets for the evaluation of the proposed system. The training and validation datasets have been used for the evaluation of the classifier models, which are used during the training and validation phases. The testing dataset has been used for the drugs localisation and the complete system evaluation.

In the drugs localisation stage, the design of the EFPN is based on the FPN coupled with the GCN, and ResNets-50 has been chosen as the backbone. To further reduce the computational complexity, the P2 layer has been removed. The three layers, P3, P4, and P5 of the FPN have been retained and two layers, P6, and P7 have been generated by the GCN to improve the objects localisation ability. The comparisons between the EFPN and the FPN, as well as the comparisons between the backbone ResNet-50 and the ResNet-101, are shown in Table 2. The FPN-ResNet-50 and the FPN-ResNet-101 have been implemented for drugs localisation with the correct rates of 48 and 76%, respectively. Figs. 6*a–c* show that the larger objects are difficult to locate by both the FPN-ResNet-50 and the FPN-ResNet-101. However, the GCN can effectively provide the receptive field for the drugs localisation task. The detected results of the proposed EFPN-ResNet-50 and the EFPN-ResNet-101 are 96.3 and 96.5%, respectively. The performances of the EFPN-ResNet-101 and the

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

EFPN-ResNet-50 are similar, but the computational complexity of the ResNet-101 is higher than the ResNet-50. The results show that the EFPN has achieved good performance on the drugs localisation task. Therefore, the ResNet-50 has been chosen as the EFPN backbone.

Since the ResNet is the backbone of the FPN, we have also evaluated the ResNet-50 and the ResNet-101 classifiers in drug classification. The error rates of the ResNet-50 and the ResNet-101 in 200 epochs are shown in Table 3. To provide the error rate curve in different epochs, the graphically evaluated results of the ResNet-50 and the ResNet-101 are given in Figs. 7*a* and *b*. Both of the graphs have the same behaviour with the error rates converging after 200 epochs. However, the performance of the ResNet series is unstable. The complete system with the ResNet has been evaluated on the validation dataset, and the error rates of Top-1, Top-3, and Top-5 are shown in Table 4. The Top-1 error rates of the ResNet-50 and the ResNet-101 are 43.5 and 39.7%, respectively. Although the EFPN-ResNet-50 models perform well in drugs localisation, the similar appearances of the drugs in multiple categories and the rotation factors still lead to the poor recognition rate in drug classification. To deal with this problem, the proposed system has two stages, the localisation stage, and the classification stage. The first stage focuses on the locations of the drugs within the original image. The second stage tries to recognise the drug category from the original image.

The proposed system has been constructed using the EFPN and the CNN models. The choice of classifiers is an important issue in the proposed system. Several CNN models have been evaluated with 300 epochs in Fig. 7*c* and *d* including the VGG-16/19 [33], the Inception V3 [37], the Inception V4 [38], the Xception [12], and the Inception-ResNet v1/v2 [38]. Table 3 lists the error rates of the CNN models for the validation of single drug classification. The results provided in Fig. 7 and Table 3 show that the VGG and the ResNet both reach convergence after 200 epochs, but the curves are still in the non-steady state. In contrast, the Xception, the Inception V4, and the Inception-ResNet v2 have faster convergence in 40–60 epochs. According to the different convolution kernels used to increase the kernel width of classifier units, the Inception series achieves better performances in the feature extraction of the drug content.

The complete system evaluates the EFPN with different CNN classifiers. Table 4 presents the comparison results with the Top-1, Top-3, and Top-5 error rates. The VGG and the ResNet are unable to serve as the classifiers for the drug classification. The Inception series have good performance in the complete system evaluation. According to the experimental results, the combination of the EFPN and the Inception-ResNet v2 has the best performance in the system evaluation. The Top-1, Top-3, and Top-5 error rates of the EFPN-Inception-ResNet v2 are 17.9, 7.6, and 5.3%, respectively. Based on the performances, the Inception-ResNet v2 has been selected as the classifier for the proposed system. The experimental results from the proposed system are presented in Fig. 8, where the locations of the drugs are marked by the green boxes and the detected results are shown over the drugs.

4.4 Time complexity analysis

The time-cost calculation has been used in the time complexity analysis. Two CNN models, the ResNet and the Inception-ResNet v2, have been used for the drug localisation task and the drug classification task, respectively. Each image contains 0–8 drugs for

Table 2 Comparison of the EFPN and FPN on drugs localisation task

Backbone	Pyramid layers	Annotation (detected/totals)	Correct rate, %
FPN-ResNet-50	{P2, P3, P4, P5}	4093/8490	48.2
FPN-ResNet-101	{P2, P3, P4, P5}	6403/8490	75.5
EFPN-ResNet-50	**{P3, P4, P5, P6, P7}**	**8164/8490**	**96.3**
EFPN-ResNet-101	{P3, P4, P5, P6, P7}	8180/8490	96.5

a *b* *c*

Fig. 6 *Comparison of drugs localisation with difference detection models*
(a) EFPN-ResNet-50, ***(b)*** FPN-ResNet-50, ***(c)*** FPN-ResNet-101

Table 3 Error evaluation of CNN classification models with 200 epochs on single drug classification

Network models\epoch	20	40	60	80	100	200
VGG-16 [30]	77.0%	29.7%	18.3%	11.3%	9.3%	4.4%
VGG-19 [30]	85.1%	57.6%	27.7%	16.6%	13.8	5.1%
ResNet-50 [31]	69.1%	52.6%	27.4%	21.4%	44.9%	13.4%
ResNet-101 [31]	79.0%	48.9%	33.7%	22.0%	23.8%	16.3%
Inception V3 [33]	48.2%	35.6%	26.5%	16.4%	11.9%	2.7%
Inception V4 [35]	22.0%	9.9%	6.6%	3.8%	3.4%	2.0%
Xception [34]	17.1%	5.3%	2.9%	2.4%	1.7%	1.6%
Inception-ResNet v1 [35]	44.8%	29.6%	23.0%	10.4%	7.9%	2.1%
Inception-ResNet v2 [35]	25.8%	13.7%	9.4%	4.4%	3.8%	1.5%

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

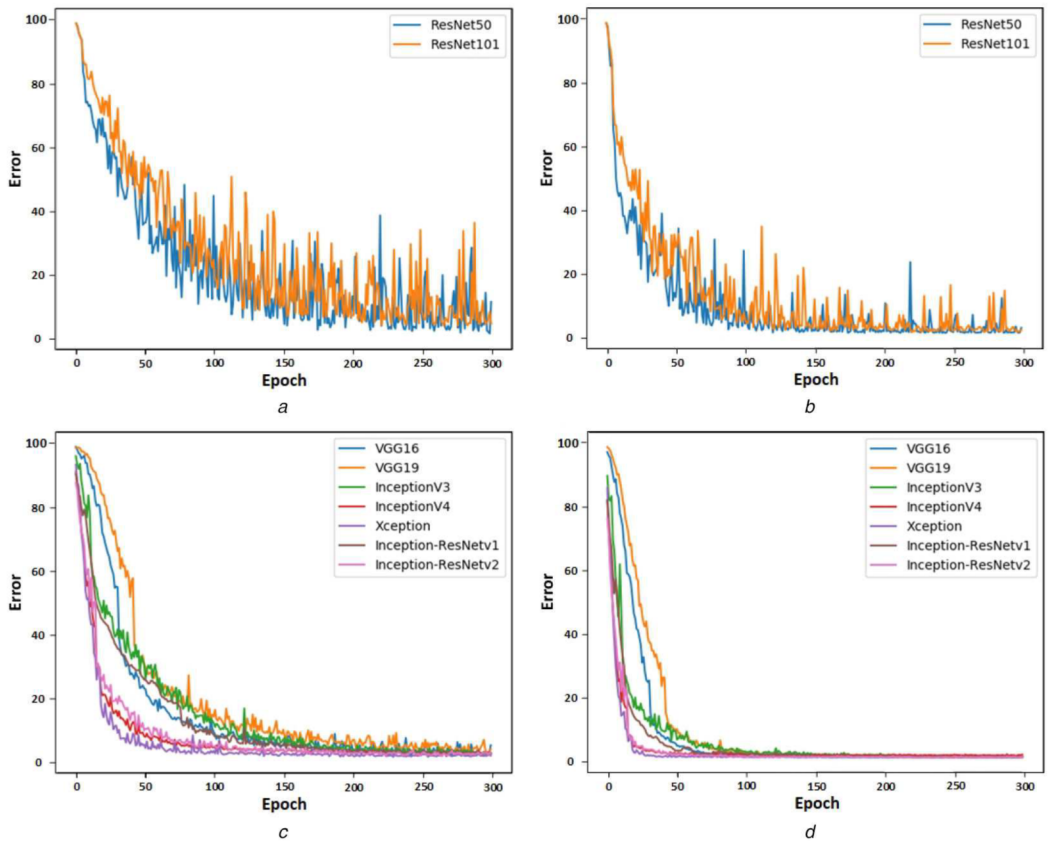

Fig. 7 *Error rate curve of signal drug classification with 300 epochs*
(a), ***(b)*** Top-1 and Top-5 error rate curves of RestNet-50 and ResNet-101, ***(c)***, ***(d)*** Top-1 and Top-5 error rate curves of CNN-based classifiers

Fig. 8 *Experimental results of the proposed automatic drug pills detection, EFPN-Inception-ResNet v2*

Table 4 Comparison with different classifiers of a complete system with top-1, top-3, and top-5 error evaluations

Network models	Top-1 error	Top-3 error	Top-5 error
EFPN-VGG 16	59.5%	54.9%	48.2%
EFPN-VGG 19	55.7%	49.7%	43.5%
EFPN-ResNet-50	43.5%	34.7%	28.2%
EFPN-ResNet-101	39.7%	31.0%	26.5%
EFPN-Inception v3	23.4%	13.1%	9.3%
EFPN-Inception v4	20.9%	11.4%	9.4%
EFPN-Xception	18.7%	9.5%	6.2%
EFPN-Inception-ResNet v1	19.1%	11.8%	8.9%
EFPN-Inception-ResNet v2	**17.9%**	**7.6%**	**5.3%**

the time-cost calculation. The time-cost calculations of the drug localisation task and the drug classification task are discussed separately. In the drug localisation task, the number of anchors to be calculated is the same in each image and the averages locating time is 65 ms. The drug images have been cropped for drug classification through predicted coordinates from the drug localisation task. In the drug classification task, the average time for drug classification is 26.7 ms. The time-cost of the classification task increases proportionally as the number of drugs in each time step. The time complexity analysis is shown in Table 5.

4.5 Mistaken cases

Although the proposed system performs well in the validation step of the training phase, the results obtained from the testing phase are not as good as the training phase. The extremely high similarity or the similar appearance of the drugs is still the reason for mis-identification. Examples of erroneous recognitions are shown in Fig. 9. In Fig. 9*a*, the pills have been produced by the same pharmaceutical factory with the same manufacturer's mark on one side and a similar imprint code on the other side. In Fig. 9*b*, the

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5 Time complexity analysis of the proposed system with images of 0–8 drugs

	0	1	2	3	4	5	6	7	8	Avg.
proposal	65.1	93.6	117.5	148.6	174.2	197.3	224.2	251.3	278.3	172.2
localisation	65.1	65.7	64.9	64.9	64.8	64.9	65.0	64.9	65.1	65.0
classification	0.0	27.8	52.5	83.7	109.4	132.1	159.1	186.4	213.2	26.7[a]

[a]The average time is the classification time of each drug. Unit: millisecond (ms).

Fig. 9 *Examples of erroneous recognitions*
(a) Different pills have same manufacturer's mark and similar imprint code, ***(b)*** Different pills have similar colours and without any imprint code, ***(c)*** Similar colours and imprints on different pills, ***(d)*** Similar colours and shapes on different pills

pills have similar colours without any imprint code on either side. In Fig. 9*c*, the pills have a line imprinted in the middle and an imprint code on both sides of the line imprint. In Fig. 9*d*, the pills have an imprint code on one side with similar colours and shapes.

5 Conclusion

In this paper, an automatic drug pills detection system has been proposed for multiple drug detection. The proposed system solves two issues in the drug pills detection for practical applications, namely the multiple pills detection and the randomly placed pills detection. The proposed EFPN is used in multiple drug localisation, and the CNN classifiers, Inception-ResNet v2, and a drug image database are used to detect randomly placed drug pills.

Two stages have been implemented in the proposed system and each stage contains an independently trained CNN model. The proposed EFPN is first used to find the drug locations regardless of the shapes and colours of the drugs; the localisation model focuses on finding the drug's location regardless of the classification task. As a result, the localisation model does not need to be retrained in most cases. The Inception-ResNet v2, which concentrates on multi-category classification tasks, has been chosen for drug classification. Moreover, we have constructed the DPID, which includes 612 categories of drugs for deep learning research. According to the experimental results, the EFPN can correctly detect 96.3% of the drug locations for the validation dataset in the DPID. The accuracy rates of the proposed system, the EFPN-Inception-ResNet v2, are 82.1, 92.4, and 94.7% for rankings of Top-1, Top-3, and Top-5, respectively. There are three main goals for our future work: the continued improvement of the drug detection accuracy, the simplification of the two-CNN model architecture, and the expansion of the DPID to cover more of the prescription drugs.

6 Acknowledgments

This work conducted in collaboration with the KVGH. Specially thank to the Chief of Pharmacy Department, Professor E.L. Lee, and all pharmacists of the KVGH for their support.

7 References

[1] Ushizima, D., Carneiro, A., Souza, M., *et al.*: 'Investigating pill recognition methods for a new national library of medicine image dataset'. Int. Symp. on Visual Computing, Las Vegas, NV, USA, December 2015, pp. 410–419
[2] Bekker, C., Gardarsdottir, H., Egberts, A., *et al.*: 'Unused medicines returned to community pharmacy: an analysis of medication waste and possibilities for redispensing', *Int. J. Clin. Pharm.*, 2017, **39**, (1), pp. 240–240
[3] Lenzer, J.: 'US could recycle 10 million unused prescription drugs a year, report says', *Br. Med. J.*, 2014, **349**, pp. 1–1
[4] National Health Service, U.K.: '*Pharmaceutical waste reduction in the NHS*' (NHSEngland, England, 2015), pp. 1–24
[5] Neto, M.A.V., Souza, J.A.W.M.D., Filho, P.P.R.C., *et al.*: 'Cofordes: an invariant feature extractor for the drug pill identification'. IEEE 31st Int. Symp. on Computer-Based Medical Systems, Karlstad, Sweden, June 2018, pp. 30–35
[6] Zeng, X., Cao, K., Zhang, M.: 'Mobiledeeppill: a small-footprint mobile deep learning system for recognizing unconstrained pill images'. Proc. of the 15th Annual Int. Conf. on Mobile Systems, Applications, and Services, Niagara Falls, NY, USA, June 2018, pp. 56–67
[7] Wang, Y., Ribera, J., Liu, C., *et al.*: 'Pill recognition using minimal labeled data'. IEEE Third Int. Conf. on Multimedia Big Data, Laguna Hills, CA, USA, April 2017, pp. 346–353
[8] Yu, J., Chen, Z., Kamata, S.-I., *et al.*: 'Accurate system for automatic pill recognition using imprint information', *IET Image Process.*, 2015, **9**, (12), pp. 1039–1047
[9] Chen, R.-C., Chan, Y.-K., Chen, Y.-H., *et al.*: 'An automatic drug image identification system based on multiple image features and dynamic weights', *Int. J. Innov. Comput., Inf. Control*, 2012, **8**, (5), pp. 2995–3013
[10] Ou, Y.-Y., Tsai, A.-C., Wang, J.-F., *et al.*: 'Automatic drug pills detection based on convolution neural network'. Int. Conf. on Orange Technologies, Bali, Indonesia, October 2018, pp. 1–4
[11] Lin, T.-Y., Dollár, P., Girshick, R., *et al.*: 'Feature pyramid networks for object detection'. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp. 936–944
[12] Chollet, F.: 'Xception: deep learning with depthwise separable convolutions'. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp. 1800–1807
[13] Peng, C., Zhang, X., Yu, G., *et al.*: 'Large kernel matters – improve semantic segmentation by global convolutional network'. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, June 2017, pp. 1743–1751
[14] Hartl, A.: 'Computer-vision based pharmaceutical pill recognition on mobile phones'. 14th Central European Seminar on Computer Graphics, Budmerice, Slovakia, May 2010, pp. 51–58
[15] Lee, Y.-B., Park, U., Jain, A.K.: 'Pill-ID: matching and retrieval of drug pill imprint images'. 20th Int. Conf. on Pattern Recognition, Istanbul, Turkey, August 2010, pp. 2632–2635
[16] Caban, J.J., Rosebrock, A., Yoo, T.S.: 'Automatic identification of prescription drugs using shape distribution models'. 19th IEEE Int. Conf. on Image Processing, Orlando, FL, USA, October 2012, pp. 1005–1008
[17] Chen, Z., Kamata, S.-I.: 'A new accurate pill recognition system using imprint information'. Sixth Int. Conf. on Machine Vision, London, UK, December 2013, pp. 1–8
[18] Yu, J., Chen, Z., Kamata, S.-I.: 'Pill recognition using imprint information by two-step sampling distance sets'. 22nd Int. Conf. on Pattern Recognition, Stockholm, Sweden, August 2014, pp. 3156–3161
[19] Suntronsuk, S., Ratanotayanon, S.: 'Automatic text imprint analysis from pill images'. 9th Int. Conf. on Knowledge and Smart Technology, Chonburi, Thailand, February 2017, pp. 288–293
[20] Suntronsuk, S., Ratanotayanon, S.: 'Pill image binarization for detecting text imprints'. 13th Int. Joint Conf. on Computer Science and Software Engineering, Khon Kaen, Thailand, July 2016, pp. 1–6
[21] Yaniv, Z., Faruque, J., Howe, S., *et al.*: 'The national library of medicine pill image recognition challenge: an initial report'. IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, October 2016, pp. 1–9
[22] Girshick, R., Donahue, J., Darrell, T., *et al.*: 'Rich feature hierarchies for accurate object detection and semantic segmentation'. IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, June 2014, pp. 580–587
[23] He, K., Zhang, X., Ren, S., *et al.*: 'Spatial pyramid pooling in deep convolutional networks for visual recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **37**, (9), pp. 1904–1916
[24] Girshick, R.: 'Fast R-CNN'. IEEE Int. Conf. on Computer Vision, Washington, DC, USA, December 2015, pp. 1440–1448
[25] Ren, S., He, K., Girshick, R., *et al.*: 'Faster r-cnn: towards real-time object detection with region proposal networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **39**, (6), pp. 1137–1149
[26] He, K., Gkioxari, G., Dollár, P., *et al.*: 'Mask r-cnn'. IEEE Int. Conf. on Computer Vision, Venice, Italy, October 2017, pp. 2980–2988

© The Institution of Engineering and Technology 2019

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[27] Zhang, X., Yuan, Y., Wang, Q.: 'ROI-wise reverse reweighting network for road marking detection'. 29th British Machine Vision Conf., Newcastle, UK, 3–6 September 2018, pp. 1–12
[28] Li, K., Cheng, G., Bu, S., *et al.*: 'Rotation-insensitive and context-augmented object detection in remote sensing images', *IEEE Trans. Geosci. Remote Sens.*, 2018, **56**, (4), pp. 2337–2348
[29] Zhang, S., Wen, L., Bian, X., *et al.*: 'Single-shot refinement neural network for object detection'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 4203–4212
[30] Yuan, Y., Xiong, Z., Wang, Q.: 'VSSA-NET: vertical spatial sequence attention network for traffic sign detection'. *IEEE Transactions on Image Processing*, 2019, pp. 3423–3434
[31] Cheng, G., Zhou, P., Han, J.: 'Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (12), pp. 7405–7415
[32] Cheng, G., Han, J., Zhou, P., *et al.*: 'Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection', *IEEE Trans. Image Process.*, 2019, **28**, (1), pp. 265–278
[33] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition'. Int. Conf. on Learning Representations, San Diego, CA, USA, May 2015, pp. 1–14
[34] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 770–778
[35] Neubeck, A., Gool, L.V.: 'Efficient non-maximum suppression'. 18th Int. Conf. on Pattern Recognition, Hong Kong, China, August 2006, pp. 850–855
[36] Lin, T.-Y., Goyal, P., Girshick, R., *et al.*: 'Focal loss for dense object detection'. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017
[37] Szegedy, C., Vanhoucke, V., Ioffe, S., *et al.*: 'Rethinking the inception architecture for computer vision'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, July 2016, pp. 2818–2826
[38] Szegedy, C., Ioffe, S., Vanhoucke, V., *et al.*: 'Inception-v4, inception-resnet and the impact of residual connections on learning'. Proc. of the Thirty-First AAAI Conf. on Artificial Intelligence, San Francisco, CA, USA, February 2017, pp. 4278–4284

17519640, 2020, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-cvi.2019.0171. By Institution of Engineering and Technology, Wiley Online Library on [13/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

© The Institution of Engineering and Technology 2019