AN END-TO-END DEEP NEURAL ARCHITECTURE FOR OPTICAL CHARACTER VERIFICATION AND RECOGNITION IN RETAIL FOOD PACKAGING

Fabio De Sousa Ribeiro*,[1], Liyun Gong*,[2], Francesco Calivá[1], Mark Swainson[3],
Kjartan Gudmundsson[3], Miao Yu[1], Georgios Leontidis[1], Xujiong Ye[2], Stefanos Kollias[1]

[1] Machine Learning Group (MLearn), School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, United Kingdom
[2] Laboratory of Vision Engineering (LoVE), School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, United Kingdom
[3] National Centre for Food Manufacturing, University of Lincoln, Holbeach Technology Park, PE12 7PT, Holbeach, United Kingdom

ABSTRACT

There exist various types of information in retail food packages, including food product name, ingredients list and *use by* date. The correct recognition and coding of *use by* dates is especially critical in ensuring proper distribution of the product to the market and eliminating potential health risks caused by erroneous mislabelling. The latter can have a major negative effect on the health of consumers and consequently raise legal issues for suppliers. In this work, an end-to-end architecture, composed of a dual deep neural network based system is proposed for automatic recognition of *use by* dates in food package photos. The system includes: a *Global* level convolutional neural network (CNN) for high-level food package image quality evaluation (blurry/clear/missing *use by* date statistics); a *Local* level fully convolutional network (FCN) for *use by* date ROI localisation. Post ROI extraction, the date characters are then segmented and recognised. The proposed framework is the first to employ deep neural networks for end-to-end automatic *use by* date recognition in retail packaging photos. It is capable of achieving very good levels of performance on all the aforementioned tasks, despite the varied textual/pictorial content complexity found in food packaging design.

***Index Terms*—** deep learning, optical character verification, transfer learning, adaptation, maximally stable extremal regions

1. INTRODUCTION

Whilst food availability is a primary concern in developing nations and food quality/value a focal point in more affluent societies, food safety is a requirement that is common across all food supply chains. Pre-packaged food products, which are incorrectly labelled (e.g. bearing an incorrect or illegible *use by* date) result in product recalls, as the fault could lead to a food safety incident such as food poisoning. These recalls are usually at very high financial cost to food manufacturers and compromise their reputation. Recurring root causes for mistakes in food package labelling include but are not limited to, human error and varying types of equipment faults. Manual methods of package inspection create mundane and repetitive tasks, therefore placing the human operator in an error-prone working environment. Moreover, these checks also struggle to provide statistically significant correctness data on the packages as the checks are often performed only once every 5 minutes. Therefore, a robust and automatic system capable of recognising *use by* dates for verification is highly desirable to industrial manufacturers.

2. RELATED WORK

The introduction of supervisory Optical Character Verification (OCV) and Recognition (OCR) systems for *use by* dates, is both financially and safety-wise advantageous for suppliers. However, currently existing OCV systems are exclusively off-the-shelf and black-box commercial ones. Moreover, they heavily rely on consistency of date-code format, packaging design and viewing angle, when ascertaining the print quality of a known text by comparing it against a reference image. This consistency is almost unattainable due to high variability of food packaging designs in the food industry. Several traditional image processing methodologies have been applied to images for detection/recognition of text regions with good levels of success, including techniques such as Stroke Width Transform (SWT) [1] and Maximally Stable Extremal Regions (MSER) [2]). Deep learning approaches by way of CNNs/FCNs have shown to be especially effective in these tasks [3, 4, 5, 5, 6, 7, 8, 9, 10, 11]. Given the small volume of available food packaging data, the risk of overfitting is increased when training state-of-the-art architectures. Effective techniques such as Dropout ([12, 13]), data augmentation [14], and transfer learning (TL) [15] can help mitigate this effect. TL involves the adaptation of low- to high-level feature representations learnt from a different distribution to solve new tasks [16]. In this work, TL is performed for adapting a light-weighted fully convolutional network architecture in [3]. Inspired by the recent research in the field of text detection/recognition, state-of-the-art image processing and FCN

*Both authors have contributed equally to this work.

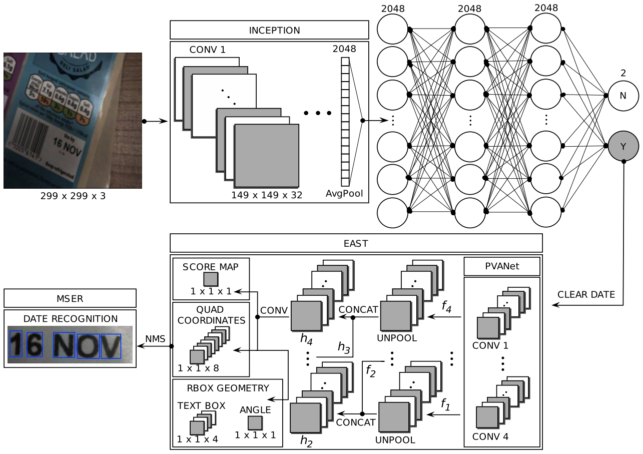


Fig. 1. Illustration of the proposed unified framework for *use by* date OCV and OCR. The top row depicts the first network, responsible for image quality evaluation and filtering. On the bottom, the FCN architecture responsible for date ROI localisation followed by characters recognition.

techniques were combined to propose an end-to-end framework for robust text detection/recognition in retail food packaging.

3. THE PROPOSED APPROACH

To address the aforementioned food package challenges and provide a robust machine vision solution for *use by* date recognition, a unified deep learning framework was devised. Concretely, both global and local approaches were studied and combined to leverage the feature extraction capabilities inherent to Deep Neural Networks (DNNs). Furthermore, by utilising adaptation strategies such as TL, it was possible to mitigate the negative effect of the small dataset available, as well as provide meaningful correctness statistics and filtering of inadequate images from the OCV pipeline. The structure of the proposed end-to-end system can be observed in Fig. 1, which is comprised of two networks. The first network is responsible for the pre-processing and selection of candidate images for date recognition. During this procedure, the first network acts as a filter of blurry/unreadable images or those with missing day/month dates, which could trigger False Positives in the subsequent detection/recognition procedures. By discarding inadequate images from the pipeline, it was possible to reduce the computational cost of the second network. As a byproduct, the first network also produces useful statistical information regarding the image quality and potentially missing *use by* dates. In practice, this is a very desirable property for food manufacturers. For example, a statistically high volume of images with low quality/missing dates could indicate possible equipment faults (printer/camera) and could be reported for maintenance in a more timely manner. The second part of the system consists of a Fully Convolutional Neural Network (FCN), which is responsible for local processing of images. Multiple levels of features from image patches are extracted and exploited to localise the *use by* date ROIs. Date characters could then be segmented and recognised within the respective ROIs through classical image processing and machine learning techniques.

3.1. CNN Transfer Learning

As depicted in the first row of Fig. 1, the first DNN is based on conventional CNN architectures ([17, 18]). CNNs are comprised of filtering layers, in which a number of affine transformations and subsequent non-linearities are applied to an input vector. It is common that CNNs use pooling layers to summarise the activations of multiple adjacent filters within a single response, and also add robustness to the model against input translations. CNN architectures take as input three channelled images and through a series of volume-wise convolutions and feature routing, are capable of selecting the optimal features/filters for classification of particular objects. This in turn eliminates the need for hands-on feature engineering approaches, as customary in classical Machine Learning and Computer Vision. For the problem adressed in this work, it was of particular interest to conduct transfer learning and assess the adaptability of pre-trained CNN weights to food package image datasets. In addition to the traditional architecture of Inception-V3, new fully-connected layers and a final softmax layer were added. In order to optimise the TL performance of the new network, a series of architectural decisions were made empirically. The best performances were achieved with a fully-connected network consisting of two 2048 unit hidden layers with Rectifier Linear Unit ($ReLU :\rightarrow f(x) = max(0, x)$) activation function. As previously alluded to, the risk of overfitting rises as the number of parameters increases w.r.t the number of training examples. Given the scarcity of training data available, it is unfeasible to train very deep models from scratch. Therefore, it was paramount to introduce an effective regulariser in the new network as well as to adapt previously learned low-level features by way of TL. One of the most effective regularisation techniques is Dropout [13]. In practice, to preserve more information in the input layer of the network and thus help learning, the neuron Dropout probability was set to 0.8, whereas for each hidden fully-connected layer of the network it was set to 0.5. Considering the class unbalance of *use by* date information, it was advantageous to use weighted categorical cross entropy as a cost function (1). In (1), ω_j is a weight computed for the j^{th} class of J total number of classes, as a function of the proportion of instances N_j compared to the most populated class (2). This discourages significant reduction of the loss for densely populated classes. Adam optimisation was used as to include adaptive learning rate, momentum, RMSprop and bias correction in weight updates, which helps to obtain faster convergence rate than normal Stochastic Gradient Descent with momentum [19].

$$\mathscr{L}(x, \hat{x}) = -(\omega_j x \log(\hat{x}) + (1-x)\log(1-\hat{x})) \quad (1)$$

$$\omega_j = \frac{\max(\{N_i\}_{i=[1:J]})}{N_j} \quad (2)$$

3.2. Fully Convolutional Neural Network

To effectively identify and recognise the *use by* date ROI in food packaging, various types of textual/pictorial content must be disregarded by the automated system. A DNN approach was devised to overcome this challenge. Specifically, a FCN architecture originally developed for detecting text, as described in greater detail in [3], was fine-tuned on the food package datasets, for detecting *use by* date ROIs. The full FCN architecture is shown in the lower part of Fig. 1, which is mainly composed of three parts: feature extractor stem, feature-merging branch and output layer. The stem part is a PVANet[20], with interleaving convolution and pooling layers. Four levels of feature maps, denoted as f_i are extracted from the original input image, whose sizes are $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$ of the original input image. Features from different scale levels meet the requirements of detecting text regions with different sizes. In the feature-merging branch, features are merged with the following strategy:

$$\begin{aligned} g_i &= \begin{cases} unpool(h_i) & if \;\; i \leq 3 \\ conv_{3\times3}(h_i) & if \;\; i = 4 \end{cases} \\ h_i &= \begin{cases} f_i & if \;\; i = 1 \\ conv_{3\times3}(conv_{1\times1}([g_{i-1}; f_i])) & if \;\; i = 4 \end{cases} \end{aligned} \quad (3)$$

where g_i is the merge as found in [3] and h_i is the merged feature map. The operator $[;]$, represents concatenation along the channel axis. In each merging stage, the feature map from the last stage is first fed to an unpooling layer to double its size, then concatenated with the current feature map. A $conv_{1\times1}$ bottleneck cuts down the number of channels to reduce computation, followed by a $conv_{3\times3}$ that fuses the information to finally produce the output of this merging stage. Subsequently, a $conv_{3\times3}$ layer produces the final feature map of the merging branch and feeds it to the output layer. The final output layer contains several $conv_{1\times1}$ operations to project 32 channels of feature maps into a 1 channel score map F_s. This map provides the likelihood that a pixel belongs to the *use by* date region, as well as a multi-channel geometry map F_g, which could either be a rotated box (RBOX) or quadrangle (QUAD) representing different geometries. The RBOX geometry map contains a 4-channel map representing 4 distances from every pixel location to the top, right, bottom, left boundaries of a rectangle enclosing the candidate *use by* date region, as well as a 1-channel map representing the angle of the related rectangle. QUAD geometry map is an 8-channel map, which contains the coordinate shift from four corner vertices of a quadrangle (representing candidate *use by* date region) to every pixel position. The loss function to be optimised can be defined as

$$L = L_s + \lambda L_g \quad (4)$$

where L_s and L_g represent losses for score and geometry maps respectively, while λ is a balancing parameter of the two. The term L_s is defined as:

$$L_s = -\beta Y^* \log \hat{Y} - (1-\beta)(1-Y^*)\log(1-\hat{Y}) \quad (5)$$

where $\hat{Y}$ and Y^* represent the predicted and groundtruth score maps respectively. β is a balancing parameter. While the L_g is defined as scale-invariant Intersection over Union (IoU) loss, for the RBOX geometry map and scale-normalised smoothed-$L1$ for the QUAD. The network was then fine tuned end-to-end by optimising the defined loss function using the Adam optimiser, until performance stopped improving. To determine the final *use by* date region from outputs of the fine tuned network, a threshold is firstly set to filter out obtained output geometries with corresponding small scores. Remaining geometries will then be merged by the locality-aware Non-Maximum Suppression (NMS) methodology. Characters in the detected expiry date region can then be segmented, with related features being extracted and classified to proper categories. The segmentation, feature extraction and classification procedures can be implemented by well-built toolboxes, such as Tesseract OCR [21].

4. EXPERIMENTAL STUDY

Two datasets comprised of food package label images collected by a leading food company were provided for research purposes. The two datasets included 1404 and 6739 captured images with different colours/contexts respectively. Fig. 2 is exemplary of the datasets utilised. The images were first manually annotated to form separate categories, namely: complete dates, missing day, missing month, no date and unreadable. In the case of the unreadable category, upon inspection, a date was not discernible from the background, potentially due to heavy distortion, non-homogeneous illumination or blur. Otherwise, images in which the day/month or both were missing, were considered as incomplete. Moreover, it was observed that some of these images included packaging which had been folded at crucial points, digits fading over time, or had human-made date occlusions. Annotated images are divided into two groups, the first part is applied to train the first network to classify the high-level image quality information (i.e. readable/unreadable, complete/partial or no date). The performance of the trained network is evaluated on a separate second group. As can be observed in Table 1, despite the small training sets and high data variability, the system was able to obtain high classification accuracy for all

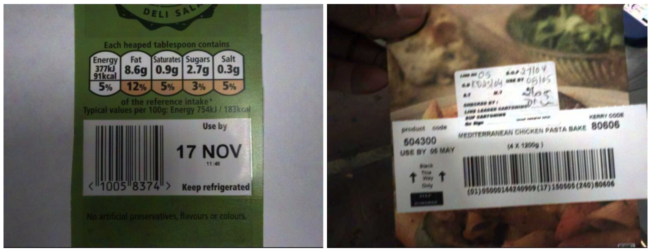


Fig. 2. Demonstration of image variability. On the left, it is relatively simple to localise/recognise the *use by* date. On the right side it becomes significantly more complex.

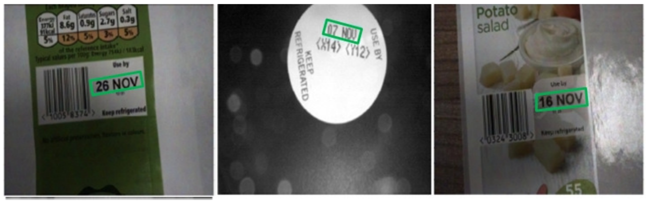


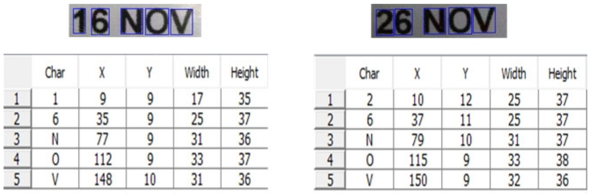
Fig. 3. Date ROI localisation results from the network 2 procedure.

Table 1. Global based experiment results with network 1.

Complete vs.	Dataset	Images	Accuracy %
Unreadable	1	645 vs 645	90.1
	2	2847 vs 2847	96.8
Partial/No Date	1	645 vs 444	89.3
	2	2954 vs 2954	95.9

Table 2. Local based experiment results with network 2.

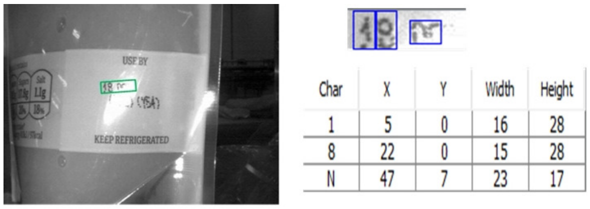
	Tested Clear Images	Accuracy %
Dataset 1	240	98
Dataset 2	482	97.10



	Char	X	Y	Width	Height
1	1	9	9	17	35
2	6	35	9	25	37
3	N	77	9	31	36
4	O	112	9	33	37
5	V	148	10	31	36

	Char	X	Y	Width	Height
1	2	10	12	25	37
2	6	37	11	25	37
3	N	79	10	31	37
4	O	115	9	33	38
5	V	150	9	32	36

Fig. 4. Examples of MSER based date character recognition results.



Char	X	Y	Width	Height
1	5	0	16	28
8	22	0	15	28
N	47	7	23	17

Fig. 5. Example of False Positive date recognition when omitting network 1 verification and image/date quality evaluation.

tasks. Furthermore, clear images including full date information selected by the first DNN on two different datasets were collected. The second (FCN) network was utilised in the *use by* date ROI localisation in these images. 70% of collected images were used for fine-tuning the FCN and 30% testing. The detection accuracies are summarised in Table 2, for the selected datasets. Examples are shown in Fig. 3, illustrating the FCN's high detection accuracy on a variety of food packages present in both datasets. By focusing solely on the detected *use by* date ROI, the date characters can more easily be recognised. The MSER algorithm is applied to segment and recognise the date from the extracted ROI region, selected examples can be seen in Fig. 4. Finally, we illustrate one of the advantages of applying the cascade of two networks. As shown in Fig. 5, a blurry image can lead to False Positive recognition of *use by* dates. In this example, the ground truth *use by* date is $18DEC$; however, the recognised date information, by adopting only the second network, is closer to $18NOV$. With the aid of the first DNN, this image is classified as inadequate, thus it was not fed into the second network for processing, therefore False Positive recognition can be avoided and computational costs reduced.

5. CONCLUSION

In this work, we have proposed an end-to-end deep neural architecture for the automatic verification and recognition of *use by* dates in food package photos. The architecture consists of two networks; the first is responsible for OCV of candidate images for OCR. These images are then fed into the second FCN which is responsible for the identification of *use by* date ROIs. Date characters within the ROIs were then recognised utilising the MSER algorithm. Promising experimental results have been obtained on a myriad of real life food package photos with varying textual/pictorial contexts. As a future step, the current framework will be made more robust and accurate in the *use by* date recognition of lower quality images, ultimately targeting the deployment of a system to be adopted in the food industry.

6. ACKNOWLEDGEMENTS

The research presented in this paper has received funding from EPSRC (Reference number EP/R005524/1) and Innovate UK (Reference number 102908), in collaboration with the Olympus Automation Limited Company, for the project Automated Robotic Food Manufacturing System.

7. REFERENCES

[1] Boris Epshtein, Eyal Ofek, and Yonatan Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2963–2970.

[2] Huizhong Chen, Sam S Tsai, Georg Schroth, David M Chen, Radek Grzeszczuk, and Bernd Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2609–2612.

[3] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "East: an efficient and accurate scene text detector," *arXiv preprint arXiv:1704.03155*, 2017.

[4] Zheng Zhang, Wei Shen, Cong Yao, and Xiang Bai, "Symmetry-based text line detection in natural scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2558–2567.

[5] Francesco Caliva, Fabio De Sousa Ribeiro, Antonios Mylonakis, Christophe Demaziere, Paolo Vinai, Georgios Leontidis, and Stefanos Kollias, "A deep learning approach to anomaly detection in nuclear reactors," in *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[6] Fabio De Sousa Ribeiro, Francesco Caliva, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis, and Stefanos Kollias, "An adaptable deep learning system for optical character verification in retail food packaging," in *Evolving and Adaptive Intelligent Systems, IEEE Conference on*, 2018.

[7] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[9] Baoguang Shi, Xiang Bai, and Cong Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[10] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng, "End-to-end text recognition with convolutional neural networks," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3304–3308.

[11] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176.

[12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] Martin A Tanner and Wing Hung Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[15] Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan, "Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 117–129, 2017.

[16] Yoshua Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] K. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "Pvanet: Deep but lightweight neural networks for realtime object detection," *arXiv preprint arXiv:1608.08021*, 2016.

[21] "Tesseract-ocr," *https://github.com/tesseract-ocr/tesseract.*