Fusion of Global-Local Features for Image Quality Inspection of Shipping Label

Sungho Suh†‡, Paul Lukowicz†‡ and Yong Oh Lee∗

∗Smart Convergence Group, Korea Institute of Science and Technology Europe
†Department of Computer Science, TU Kaiserslautern, 67663 Kaiserslautern, Germany
‡German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
Email: s.suh@kist-europe.de, paul.lukowicz@dfki.de, yongoh.lee@kist-europe.de

Abstract—The demands of automated shipping address recognition and verification have increased to handle a large number of packages and to save costs associated with misdelivery. A previous study proposed a deep learning system where the shipping address is recognized and verified based on a camera image capturing the shipping address and barcode area. Because the system performance depends on the input image quality, inspection of input image quality is necessary for image preprocessing. In this paper, we propose an input image quality verification method combining global and local features. Object detection and scale-invariant feature transform in different feature spaces are developed to extract global and local features from several independent convolutional neural networks. The conditions of shipping label images are classified by fully connected fusion layers with concatenated global and local features. The experimental results regarding real captured and generated images show that the proposed method achieves better performance than other methods. These results are expected to improve the shipping address recognition and verification system by applying different image preprocessing steps based on the classified conditions.

I. INTRODUCTION

With the rapid rise of the logistics delivery market, the annual cost associated with failed delivery has also increased. In 2015, this annual cost was estimated to be approximately 1.5 billion dollars for the postal service and 20 billion dollars for the mailing industry in United States of America [1]. One of the main causes of misdelivery is incorrect shipping labels, which may contain falsely given addresses or may be damaged during the logistics process. Packages with problematic labels lead to not only cost loss during unnecessary delivery to undeliverables and return but also damage to the manufacturers and sellers reputations. A process for screening undeliverables and error-containing barcodes in the shipping label is used for identifying packages and their destinations is required.

To verify the information in the shipping label, the commonly used approaches are optical character verification (OCV) and optical character recognition (OCR). These approaches capture an image using a vision system or smartphone, detect regions of interest (ROIs), and recognize optical characters in the ROIs. Several traditional image processing methodologies have been applied for OCV and OCR, but deep-learning-based approaches have been shown to be effective for these tasks [2], [3], [4], [5], [6]. Previous studies adopting a deep neural network framework provided high accuracy of verification and recognition of expiration dates in a food package or address in a shipping label [7], [8], [9].

However, the performance of OCR engines and text detection engines is sensitive to image quality and defects on target objects. The quality of the captured image can be affected by many degradations caused during image acquisition. Further, during the packaging and delivery processes, the shipping label can be damaged. Thus, proper assessment and classification of the captured image and object are required to improve the text detection and recognition processes. Four poor conditions of captured images are defined, and example images of generated and collected datasets are shown in Fig. 1 and Fig. 2.

1) Unreadable: Blurring is the most common issue in real-world applications, occurring because of defocusing or camera motion [10], and an image acquired at the wrong positioning leads to missing the ROI. By detecting the blur degradation and incorrect positioning, the user reacquires the image and the OCR accuracy can be improved.

2) Contaminated: Shipping labels are often contaminated in the delivery process, or some part of the address area in the shipping label can be occluded by a pen mark. By processing with methods such as document image enhancement and binarization [11], [12], the address in the contaminated or occluded label can be recognized better.

3) Handwritten: The shipping label sometimes contains handwritten addresses. Because recognition engines for machine-printed characters and handwritten characters are different, the two types of characters must be classified differently.

4) Damaged: The shipping label can be torn or occluded by another shipping label in the processes of packaging and delivery. By detecting damaged shipping labels, misdelivery can be reduced and the performance of the text recognition process can be improved.

In this paper, we propose an input image quality verification method using convolutional neural networks (CNNs) for shipping label inspection. The proposed method classifies the mentioned previously five image types: normal, contaminated
address, unreadable image, handwritten address, and damaged shipping label.

However, classification methods based on CNNs are not suitable for direct use for input image quality verification because of two problems. One problem is the varying aspect ratios and sizes of shipping label images. The varying aspect ratios make it difficult to design a CNN with a fixed image size as input. Another problem is that it is difficult to distinguish between contaminants in the address area and contaminants in other areas. In general, there are two methods for handling input images of different sizes: (1) resizing the input image while maintaining the aspect ratio in the padding space and (2) using patches as inputs to the CNN to extract discriminative features. However, resizing the input images leads to increasing the ambiguity between normal and blurred images because of the loss of detailed features, making it difficult to detect defects in ROIs. Conversely, dividing an input image into patches can lead to the loss of global features, degrading the classification performance.

We integrate local and global features from different CNNs. To localize the local features, we detect the barcode and address areas using the deep-learning object detection algorithm You Only Look Once (YOLO) [13], [14]. Then, we localize and crop regions with strong features using one of the famous feature extraction methods, features from accelerated segment test (FAST) [15]. The global features are extracted from the global CNNs with a resized input image, and the local features are extracted from independent local CNNs with the detected address and barcode area and the localized regions using FAST. We concatenate the global and local features together into fully connected fusion layers to classify the shipping label image conditions. The contribution of this paper can be summarized into two points. (1) We detect and localize the ROIs using YOLO and FAST and extract the local features using independent local CNNs. (2) By constructing the variant global and local images to combine the global and local features, we propose a novel CNN model-based stacked generalization ensemble method [16] for image quality verification.

The rest of the paper is organized as follows. Section II summarizes related works. Section III describes the details of the proposed algorithm, and Section IV presents the experimental results. Finally, Section V concludes the paper with discussion of the results.

II. RELATED WORK

Many inspection methods for printed labels have been studied [17], [18], [19], [20]. A previous study [8] proposed a deep-learning system where the shipping address is recognized and verified based on a camera image capturing the shipping address and barcode area. As mentioned in the previous section, proper assessment and classification of the captured image and object are required to improve the text detection and recognition processes. Several traditional image-processing methodologies, including stroke width transform [21], scale-invariant feature transform [22], and maximally stable extremal region [23], have been applied to detect text regions and inspect the printing quality. However, deep-learning-based approaches have been shown to be especially effective in these tasks [7], [8], [9]. Ribeiro et al. [7] proposed automatic recognition of expiration dates in food packages using EAST [5]. This method evaluated food package image quality using CNNs and the localized date region by fully connected networks and EAST. However, the ambiguity between normal and blurred images is aggravated, as is the classification of defects.
in ROI and defects outside of ROI because the image quality evaluation using CNNs used only the resized entire image.

Several papers have proposed CNN-based classification methods that integrate global and local features. Tian et al. [24] proposed a vehicle model recognition model using an iterative discrimination CNN based on selective multi-convolutional region feature extraction. The SMCR features consisted of global and local features, and the classification model combined global and local features. Lu et al. [25] presented a script identification method integrating local and global CNNs in natural scene images. They trained the global CNN with segmented images and the local CNN with patches. He et al. [26] proposed a traffic sign recognition method by combining global and local features, which were extracted by histograms of oriented gradient (HOG), color histograms (CH), and edge feature (EF). These methods showed that the CNN-based classification combining global and local features provided better performance than the CNN-based classification with only an entire image.

III. PROPOSED METHOD

The shipping label inspection system [8] involves six steps: input image quality verification, calibration, salient region detection, image enhancement, text recognition, and address validation. The flowchart of the shipping label inspection system is demonstrated in Fig. 3.

In the first step, the proposed input image quality verification process classifies the five image types: normal, contaminated address, unreadable image, handwritten address, and damaged shipping label. The unreadable image and the damaged label are reported to the user to acquire the image properly or check the condition of the shipping label. The contaminated and handwritten addresses provide information on the address area for image enhancement and/or text recognition processes.

The goal of the input image quality verification method is to classify the five image types. We propose an input image quality verification method using convolutional networks with global and local features. The overall network architecture of the proposed method is presented in Fig. 4.

A. Feature Localization

Classification with a resized entire input image leads to ambiguity between normal and blurred images because of the loss of detailed features, making it difficult to detect contaminated defects and damaged defects in ROIs. For these reasons, we detect barcode and address areas and crop regions with strong features to extract local features. In this study, we train global and local CNNs to extract the global and local features and combine the global and local features, as shown in Fig. 4.

Algorithm 1 Feature localization using the FAST algorithm.

We use the default threshold value of the FAST algorithm, \( t = 50 \), patch size \( D_p = (256, 256) \), and number of patches \( n_p = 3 \).

**Input:** input image of shipping label

**Output:** localized regions using FAST

1. Convert input image to grayscale image
2. Detect feature points \( p \) using FAST with \( t \)
3. \( D_i = \text{size of input image} \)
4. if \( (D_i > D_p) \) then
5. \( \text{patchNum}_{ik} = D_i/D_p \)
6. \( L_p = \text{number of } p \)
7. for each \( j \in [1, \text{patchNum}_{ik}] \) do
8. \( \text{iCountPoint} = 0 \)
9. \( R_j = \text{region of } j\text{-th } \)
10. for each \( k \in [1, L_p] \) do
11. if \( (p(k) \in R_j) \) then
12. \( \text{iCountPoint} = \text{iCountPoint} + 1 \)
13. end if
14. end for
15. \( \text{iCountFeature}(j) = \text{iCountPoint} \)
16. end for
17. Extract the \( n_p \) largest patches and crop the region \( I_p \)
18. else
19. Copy the image to \( I_p \) in the padding space of size \( D_p \)
20. end if
21. return \( I_p \)

To detect and localize the ROIs, we use YOLO to detect barcode and address areas and FAST to localize regions with strong features. YOLO is a fast general-purpose object detector using deep neural networks that can achieve real-time detection [14], which is much faster than R-CNN- [27] and Faster R-CNN [28] methods. Unlike conventional region-based object detection algorithms, the YOLO detector considers object detection as an end-to-end regression problem and uses a single convolutional network to predict the bounding boxes and the corresponding class probabilities. We fine-tune the YOLO detector to locate the barcode and address areas in the shipping label images. We exploit the YOLO network implemented on the Darknet19 [29] framework with a resized input image of size \( 416 \times 416 \). After detecting the barcode and address regions, we crop and resize the detected regions.

![Fig. 3. Flowchart of the shipping label inspection system](image-url)
to $256 \times 256$ while maintaining the aspect ratio in the padding space to set the input in the local CNNs.

The global features are extracted from the global CNNs with a resized input image, and the local features are extracted from independent local CNNs with the detected address and barcode area, and the localized regions using FAST. FAST is a famous corner-detection algorithm that can be used to extract feature points; it demonstrates rapid operation and a low number of computations compared with other feature-detection methods. We calculate the number of feature points from the FAST algorithm in a patch of size $256 \times 256$, the same as the input image size in the local CNNs. By choosing patches with a large number of feature points, regions with strong features, such as barcodes or characters, can be selected as input images of local CNNs. Additionally, because the patches are not resized, but rather maintain the original scale of the input image, they can provide helpful information to classify unreadable input images using the local CNNs.

**B. Global and Local Feature Fusion**

Global features describe the overall outline of the input image, whereas local features focus on the local shape, the degree of contamination in the localized regions, lighting conditions, and other imaging factors. To extract the global and local features, a resized complete input image is processed in the global CNNs, and the localized images are processed in the independent local CNNs, as shown in Fig. 4. In this study, we deploy ResNet-50 [30] pre-trained on ImageNet [31] for global and local feature extractions. In the proposed architecture, four independent CNNs are trained for feature extraction: a global CNN, two local CNNs with address and barcode areas by YOLO, and a local CNN with regions cropped by FAST. The local CNNs with regions cropped by FAST share the network parameters in the training procedure.

To combine the global and local features, we adopt a stacked generalization ensemble. Stacked generalization is an ensemble method where a new model learns how to combine the predictions from multiple models [16]. To control the importance of the feature combination, we add a fully connected layer with different output feature dimensions to the feature extraction architecture. In the proposed model, we set the dimension of the global feature to 2048 and the dimension of each local feature to 512. To verify the proposed model, we compare the proposed ensemble models with majority voting and weighted majority voting algorithms [32].

**IV. EXPERIMENTAL RESULTS**

In this section, we verify the proposed method on our generated and collected real dataset and compare the results with other image classification methods: ResNet-50 [30] and VGGNet-19 [33]. The experiments were all implemented using Python scripts in the Pytorch framework [34] and tested on a Linux system. Training procedures were performed on NVIDIA Tesla V100 GPUs.

**A. Dataset**

To the best of our knowledge, there is no open database for shipping labels because of privacy issues. Because of the scarcity of shipping label images, we generated and collected 5306 and 1092 images of different types and from various countries using smartphones. Note that we collected real data with the agreement of the owner, and we plan to release the dataset onto our public website. The dataset includes images of the shipping labels, the position information of the barcode and address, and five annotated labels: normal, contaminated, unreadable, handwritten, and damaged. Five people manually annotated the dataset with specific sorting criteria. In the case of the “unreadable” class, the receiver’s address is blurred or distorted, or the resolution is too low to recognize the
address. For the “contaminated” and handwritten class, the receiver’s address is contaminated, occluded by pen mark, or handwritten. Finally, a damaged shipping label indicates that the shipping label is torn or occluded by another shipping label. Detailed information about the dataset is presented in Table I.

To further evaluate the performance of the proposed method, we obtained the shipping label formats from 15 different carrier companies, generated barcodes randomly, and entered random addresses from seven different countries. The label images were generated in various sizes using NiceLabel Designer software. To evaluate the proposed method, we generated contaminated, unreadable, and damaged images using image-processing methods, and then we generated handwritten labels using Google handwriting fonts. Finally, there were 5306 generated images for input image quality verification. We selected 100 images per class as the test set and conducted a 10-fold cross-validation procedure for the dataset.

The collected dataset contains only 1092 images and is under a data-imbalanced condition, while the generated dataset has 5306 images and the number of data for each class is similar. To improve the performance of the classification on the collected real dataset, a data augmentation method, such as rotation (90, 180, 270 degrees) and flip, is adopted to extend the data of minority classes. To evaluate the proposed method using deep neural networks, we conducted a 10-fold cross-validation procedure for the shipping label dataset.

### B. Experimental Results on Generated Images

We evaluate the proposed method on the generated dataset in this subsection. Table II shows the performance of the feature localization by YOLO. The results are the mean average precision and show mean and standard deviation. The feature localization on the generated dataset provided a high performance, above 0.98, and the example images of the detection results are shown in Fig 5 (a).

For the evaluation, VGG-19 and ResNet-50 were used for classification, and we compared the proposed method by combining the global and local features of the classification methods with only global features and other ensemble learning methods, such as majority voting and weighted majority voting. Table III shows the classification accuracy comparison with respect to the generated dataset. The proposed method with ResNet-50 provided the highest classification accuracy of 98.46%, which exceeded ResNet-50 with only global features and other ensemble learning methods by 3.46% and 2.04%, respectively.

### C. Experimental Results on Collected Real Images

We evaluate the proposed method on collected real datasets compared with other classification models with only global features and other ensemble learning methods. Compared with the generated dataset, the collected real dataset has more image types, and it is more difficult to extract local features and classify the input image quality. Table IV presents the classification accuracy comparison on the collected real images. It can be observed that ResNet-50 has better performance than VGGNet-19, and the classification with ensemble learning methods also provided higher accuracy than the classification with only global features. This result means that the feature localization method that we proposed improves the classification of the input image quality.

#### Table I

| Defect Type   | # of Images (Generated) | # of Images (Collected) |
|---------------|-------------------------|-------------------------|
| Normal        | 1283                    | 660                     |
| Contaminated  | 1054                    | 139                     |
| Unreadable    | 904                     | 107                     |
| Handwritten   | 988                     | 52                      |
| Damaged       | 1077                    | 134                     |
| **Total**     | **5306**                | **1092**                |

#### Table II

| Region   | mAP (Generated) | mAP (Collected) |
|----------|-----------------|-----------------|
| Address  | 0.9653 ± 0.1637 | 0.8907 ± 0.0213 |
| Barcode  | 0.9976 ± 0.0018 | 0.9778 ± 0.0065 |
| **Total**| **0.9814 ± 0.0085** | **0.9343 ± 0.0105** |

Fig. 5. Examples of detected barcode and address areas (a) on generated images and (b) on the real dataset: red boxes are ground truth, and blue boxes are detected regions.
TABLE III

| Methods                                | VGG-19 Classification Accuracy | Resnet-50 Classification Accuracy |
|----------------------------------------|--------------------------------|-----------------------------------|
| Only global features                   | 95.98 ± 0.74 %                 | 95.80 ± 0.38 %                    |
| Global-local fusion(majority voting)   | 96.40 ± 0.65 %                 | 96.62 ± 0.43 %                    |
| Global-Local fusion (weighted majority voting) | 97.16 ± 0.43 %                 | 97.02 ± 0.77 %                    |
| Global-Local fusion (ours)             | 98.32 ± 0.49 %                 | 99.06 ± 0.66 %                    |

TABLE IV

| Methods                                | VGG-19 Classification Accuracy | Resnet-50 Classification Accuracy |
|----------------------------------------|--------------------------------|-----------------------------------|
| Only global features                   | 85.40 ± 2.43 %                 | 86.00 ± 3.40 %                    |
| Global-local fusion(majority voting)   | 84.67 ± 2.27 %                 | 86.40 ± 2.97 %                    |
| Global-Local fusion (weighted majority voting) | 85.00 ± 2.68 %                 | 87.60 ± 2.70 %                    |
| Global-Local fusion (ours)             | 87.80 ± 2.13 %                 | 89.26 ± 2.70 %                    |

TABLE V

| Class          | Only global features | Global-local fusion (ours) |
|----------------|----------------------|-----------------------------|
| Normal         | 93.33 %              | 95.00 %                     |
| Contaminated    | 78.33 %              | 85.00 %                     |
| Unreadable      | 83.33 %              | 88.33 %                     |
| Handwritten     | 98.33 %              | 98.33 %                     |
| Damaged         | 76.67 %              | 79.31 %                     |
| Total           | 86.00 %              | 89.26 %                     |

In this paper, we have presented an input image quality verification method using CNNs combining global and local features for shipping label inspection. We detected and localized the regions of interest using YOLO and FAST, and we extracted the global and local features using several independent CNNs. To combine the global and local features, we adopted a stacked generalization ensemble. To train and test the proposed method, we generated and collected a shipping label dataset. The experimental results show that the proposed method improved the classification accuracy by 3.46% and 3.86%, respectively, compared with the classification model with only global features, and by 2.04% and 2.86%, respectively, compared with the other ensemble learning method combining global and local features. In addition, we plan to release the generated and collected real dataset onto our public website. As future work, the proposed method will be applied to a packaging machine with an industrial camera and will be tested in the logistics industry. We will also improve the image enhancement and text recognition process by applying the proposed method.

VI. ACKNOWLEDGMENT

This research was supported by the KIST Europe Institutional Program (Project No. 12005).

REFERENCES

[1] U. S. P. S. O. of Inspector General, “Undeliverable as addressed costs more than you think,” [https://www.uspsoig.gov/], 2015.
[2] Z. Zhang, W. Shen, C. Yao, and X. 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.
[3] A. Gupta, A. Vedaldi, and A. 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.
[4] B. Shi, X. Bai, and C. 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.
