DETECTION DATASETS: FORGED CHARACTERS FOR PASSPORT AND DRIVING LICENCE

Teerath Kumar[1], Muhammad Turab[2], Shahnawaz Talpur[2], Rob Brennan[1] and Malika Bendechache[1]

[1]CRT AI and ADAPT, School of Computing, Dublin City University, Ireland
[2]Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan

ABSTRACT

Forged characters detection from personal documents including a passport or a driving licence is an extremely important and challenging task in digital image forensics, as forged information on personal documents can be used for fraud purposes including theft, robbery etc. For any detection task i.e. forged character detection, deep learning models are data hungry and getting the forged characters dataset for personal documents is very difficult due to various reasons, including information privacy, unlabeled data or existing work is evaluated on private datasets with limited access and getting data labelled is another big challenge. To address these issues, we propose a new algorithm that generates two new datasets named forged characters detection on passport (FCD-P) and forged characters detection on driving licence (FCD-D). To the best of our knowledge, we are the first to release these datasets. The proposed algorithm first reads the plain image, then performs forging tasks i.e. randomly changes the position of the random character or randomly adds little noise. At the same time, the algorithm also records the bounding boxes of the forged characters. To meet real world situations, we perform multiple data augmentation on cards very carefully. Overall, each dataset consists of 15000 images, each image with size of 950 x 550. Our algorithm code, FCD-P and FCD-D are publicly available.

KEYWORDS

Character detection dataset, Deep learning forgery, Forged character detection.

1. INTRODUCTION

Personal documents including passport or driving licence, contain key information of a person and these documents are used for various purposes including critical office work, bank account access, any type of insurance and others. But these documents can easily be modified with deep learning algorithms and used for many fraud purpose including theft, robbery, terrorism, etc. [1]. Recent deep learning algorithms demonstrated that character can easily be forged using convolutional neural networks [2, 3, 4, 5, 6] in sequence to sequence manner. These deep learning algorithms can forge the documents from text, colour or background perspectives, but they are computationally very expensive. To detect the forged characters on the documents is a benchmark challenge. As DL algorithms are data-hungry [7], and to detect the forged characters, these algorithms require the high computational resources and labelled training data. Finding dataset(s) for documents is restricted due to many reasons, information privacy, unlabeled data

David C. Wyld et al. (Eds): AISCA, NET, DNLP - 2022
pp. 45-54, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120204

and many other reasons. To reduce this gap, first we propose an algorithm that generates a dataset using the plain background documents of five different countries, and we are the first to release the synthetic two datasets one for passport and other for driving licence using the algorithm.

Previously several methods have been used for forged character detection [8, 9] from document plain text. Algorithm [8] automatically detects tampered characters from document images by measuring distance between feature vectors of Hu moments. Algorithm calculates possible conception errors by considering principal inertia axis, horizontal axis and character size; further character is classified as real or fake based on the score system. Algorithm [9] detects whether the character is real or fake with the help of geometric parameter distortion mutation, for a single character algorithm estimated distortion parameters based on translation and rotation distortion. This algorithm [10] detects characters from an ID card using a traditional image processing method consisting of four stages: pre-processing, text-area extraction, segmentation, and recognition. In the above mentioned works, either dataset is restricted or algorithms are evaluated on private datasets, to bridge the gap, we propose a new algorithm that generates the forged characters dataset. Furthermore we release two datasets i.e. FCD-P and FCD-D using the proposed algorithm.

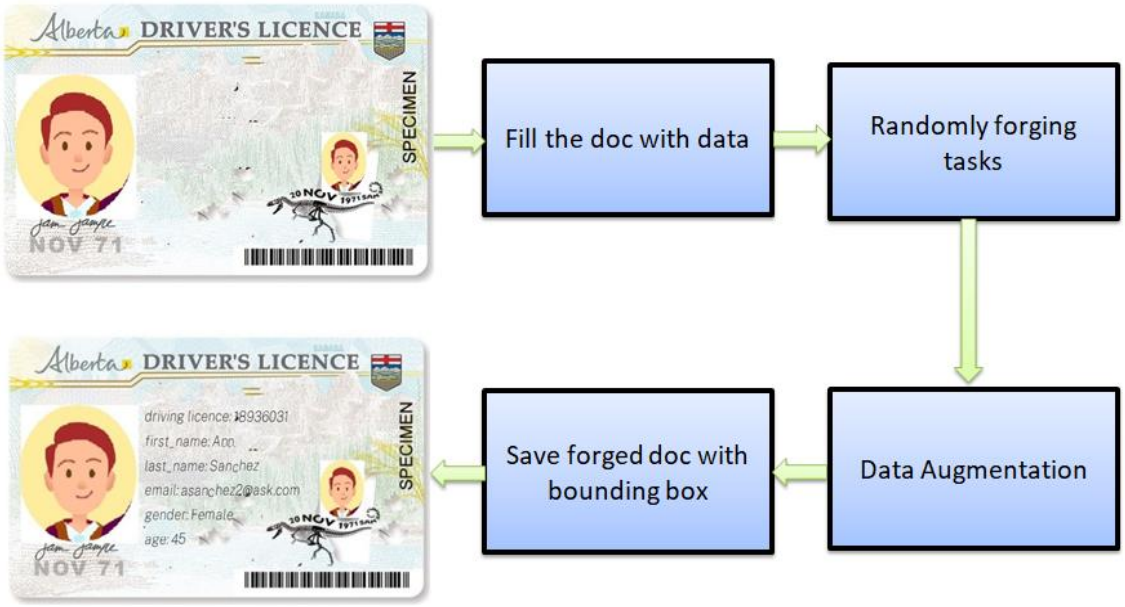

Figure 1. Proposed Algorithm Workflow

2. DESCRIPTION

Each released dataset consists of 15000 RGB images of dimension 900 x 550 each. First we get a plain background image of either passport or driving licence, taken online of five different countries including Australia, Canada, Ireland, Pakistan and USA, as driving licence and passport images are shown in figure 2 and figure 3, respectively, and their sources of the image acquisition are described in table 1. First we remove unwanted text and images on the passport or driving licence using an online website to make those plain, then we apply the proposed algorithm. The proposed algorithm consists of 4 steps, as shown in Figure 1 and described in algorithm 1. 4 steps are described as following:

Algorithm 1: Document Tampering and Augmentation Algorithm

```
fill_doc_with_data(text_font, c_font, data, info):
    /* Fill the doc with given data and return positions */
    margintop ← 10
    marginbottom ← 30
    color← (0, 0, 0)
    size← font.getsize(titlestr)
    x, y ← (info[0] − sz[0])/2, margintop
    Dy ← (info[1] − margintop − marginbottom − size[1] − 30)/4
    y ← margintop + sz[1] + 110
    for key,val in person.items() do
        x ← 10
        size_key ← font.getsize(key)x ← (info[0] − size[0])/2 − 180
        draw text
        x ← x + size_key[0] + 10
    return positions

save_docs(font_styles, backgrounds, csv_file):
    /* Read input data from a csv file then print on document with given
       font styles and save it */

data_augmentation(technique, image):
    /* Apply given augmentation technique on the image (rotation and
       shearing) */
save_tampered_docs(path):
    /* Read position from json file from the given path and apply random
       tampering and augmentation on the image then save it */
```

Algorithm 1. Proposed algorithm

2.1. Fill the document with data

First we read a plain background image of either passport or driving licence using the PIL library in python, then read a csv data file, data [11] was taken from kaggle platform which consist of five attributes including first name, last name, email, gender and age. We get a single record from a data file, and adjust it on a plain background image.

Figure 2. Plain Driving Licences

Table 1. Source of passport and licence sample images

Passport	Australia	Canada	Ireland	Pakistan	USA
Driving licence	Australia	Canada	Ireland	Pakistan	USA

Figure 3. Plain Passports

2.2. Randomly forging tasks

When documents are forged, there are two possibilities, either the forged character is not aligned with other characters or the forged character has a little bit of noise in the background. To keep these possibilities in mind, we randomly pick any character and change character location either up or down as shown in **figure 4 (A)** where in email 's' of the census is moved down, or add a little bit of uniform noise in the character background as shown in **figure 4 (B)** where driving licence, first name, last name and email have a noise in one character of each.

A B

Figure. 4. Forging Tasks

2.3. Data augmentation

In the real world, documents are not placed as straight as shown in the input column of figure 5 and figure 6, documents can be placed at any angle or stretch. To meet the real world scenario, we perform two augmentations namely rotation and shearing for the driving licences and passports image as shown in columns rotation and shearing of each figure 5 and figure 6. The bounding boxes of tampered characters are rotated using the below formula.

$$\begin{bmatrix} alpha & beta & (1-alpha)*x-beta*y \\ -beta & alpha & beta*x+(1-alpha)*y \end{bmatrix}$$

where
*alpha = scale * cos*
*beta = scale * sin*
and theta is the rotation angel
For this case, we use scale=1

Figure 5. Driving licence with applied data augmentation Right to left, input, rotation augmentation and shearing augmentation.

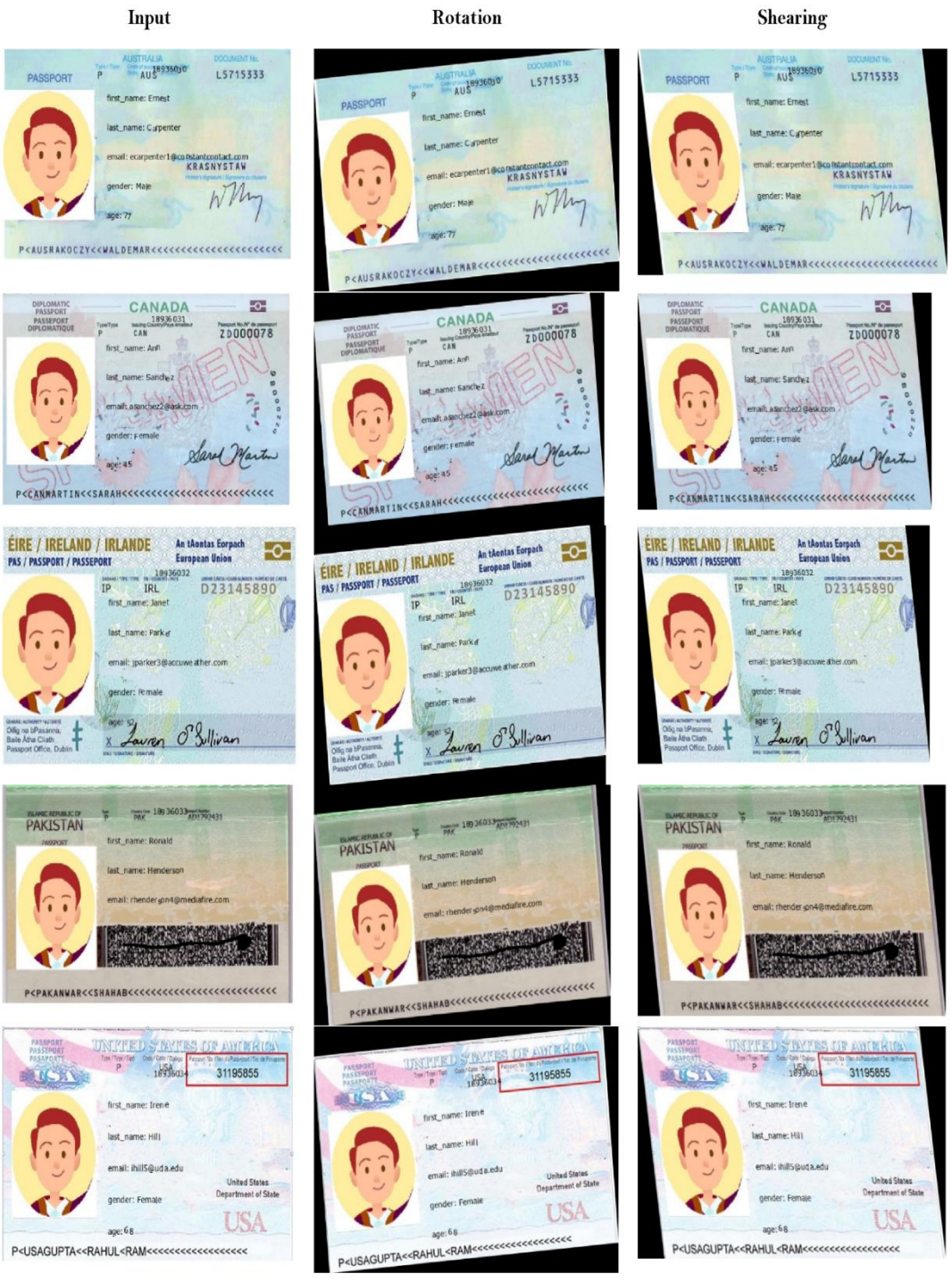

Figure 6. Passport with applied data augmentation Right to left, input, rotation augmentation and shearing augmentation.

2.4. Save forged document with bounding box

In this step we describe the motivation of recording the bounding box of tempered character, as unlabeled data can be found online in large volume, but to get the data labelled is extremely difficult, time consuming, tedious and expensive task [12]. To mitigate that issue, our proposed

algorithm records the bounding boxes of the tampered characters and saves them in the json files so that the research community can use it for training networks for detection. Above four step process is described for one document. To synthesise more data, we repeat this process to generate 3000 documents for each country document, finally we save forged document images with their bounding boxes in a json format file.

3. LIBRARIES / PACKAGE USED

We used multiple libraries to forge the character in the passport and driving licences. Libraries/packages including PIL [13] library to perform operation of image reading, setting and drawing font with its position, numpy [14] to deal rotation and other mathematical operations, random [15] to perform randomness for position and random noise adding, os [16] library to deal with file listing and path handling, json [17] library to deal with json file as annotation and cv [18] library to deal augmentation.

4. CONCLUSIONS

This paper addresses the gap of forged character detection of documents i.e. passport and driving licence, due to datasets unavailability. To fill this gap, this paper presents a new algorithm of synthesising data considering real world scienarios and releases two new datasets named forged characters detection on passport (FCD-P) and forged characters detection on driving licence (FCD-D), using five different countries' passport and licence. This research work opens new challenges for forged character detection on passport and driving licence. Finally, we release our code and datasets and the research community can use it for their research purpose. Possible future work is to include more countries' passports and driving licences and apply state-of-the-art detection algorithms.

ACKNOWLEDGEMENTS

This publication has emanated from research [conducted with the financial support of/supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6223 and is supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106/_P2), Lero SFI Centre for Software (Grant 13/RC/2094/_P2) and is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] Fake identity brits warned that their lives are in danger, Online Available:https://www.independent.co.uk/news/world/middle-east/fake-identity-brits-warned-that-their-lives-are-in-danger-1905971.html . Accessed on:

[2] Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., & Bai, X. (2019, October). Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1500-1508).

[3] Yang, Q., Huang, J., & Lin, W. (2020). Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14700-14709).

[4] Roy, P., Bhattacharya, S., Ghosh, S., & Pal, U. (2020). STEFANN: scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13228-13237).

[5] Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

[6] Zhao, L., Chen, C., & Huang, J. (2021). Deep Learning-based Forgery Attack on Document Images. *arXiv preprint arXiv:2102.00653*.
[7] Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data*, *8*(1), 1-54.
[8] Bertrand, R., Gomez-Krämer, P., Terrades, O. R., Franco, P., & Ogier, J. M. (2013, August). A system based on intrinsic features for fraudulent document detection. In *2013 12th International conference on document analysis and recognition* (pp. 106-110). IEEE.
[9] Shang, S., Kong, X., & You, X. (2015). Document forgery detection using distortion mutation of geometric parameters in characters. *Journal of Electronic Imaging*, *24*(2), 023008.
[10] Ryan, M., & Hanafiah, N. (2015). An examination of character recognition on ID card using template matching approach. *Procedia Computer Science*, *59*, 520-529.
[11] https://www.kaggle.com/avkash/5feature30kcsv/version/1 (accessed on 1/17/2022)
[12] Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.
[13] https://pillow.readthedocs.io/en/stable/ (accessed on 1/17/2022)
[14] https://numpy.org/ (accessed on 1/17/2022)
[15] https://docs.python.org/3/library/random.html (accessed on 1/17/2022)
[16] https://docs.python.org/3/library/os.html (accessed on 1/17/2022)
[17] https://docs.python.org/3/library/json.html (accessed on 1/17/2022)
[18] https://pypi.org/project/opencv-python/ (accessed on 1/17/2022)

AUTHORS

Teerath kumar received his Bachelor's degree in Computer Science with distinction from National University of Computer and Emerging Science (NUCES), Islamabad, Pakistan, in 2018. Currently, he is pursuing PhD from Dublin City University, Ireland. His research interests include advanced data augmentation, deep learning for medical imaging, generative adversarial networks and semi-supervised learning.

Muhammad Turab is an undergraduate final year student at Computer Systems Engineering MUET, Jamshoro. He has done 60+ projects with java and python, all projects can be found on GitHub. His research interests include deep learning, computer vision and data augmentation for medical imaging.

Shahnawaz Talpur is the chairman of Computer Systems Engineering Department at Muet Jamshoro. He has done his masters from MUET and PhD from Beijing Institute of Technology, China. His research interests include high performance computing, computer architecture and big data.

R. Brennan is an Assistant Professor in the School of Computing, Dublin City University, Chair of the DCU MA in Data Protection and Privacy Law and a Funded investigator in the Science Foundation Ireland ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund, His main research interests are data protection, data value, data quality, data privacy, data/AI governance and semantics.

M. Bendechache is an Assistant Professor in the School of Computing at Dublin City University, Ireland. She obtained her Ph.D. degree from University College Dublin, Ireland in 2018. Malika's research interests span the areas of Big data Analytics, Machine Learning, Data Governance, Cloud Computing, Blockchain, Security, and Privacy. She is an academic member and a Funded Investigator of ADAPT and Lero research centres.