Machine Learning and Image Processing : A Vision Approach for Validity Date Recognition from Philippine Government-issued ID

Kharisa Mae G. Macaraig and Danilo J. Mercado

***Abstract*— In this paper, the author implemented a more flexible method, other than the study conducted by Akhter et al. [1] on their national ID cards in Bangladesh, as this system accommodated the classification and data extraction from a variety of Government-issued IDs. To test the application, 70 ID cards were used as test data. The system yielded an accuracy rate of 87.14%, and an F1-Score of 91.59% suggesting that the system is highly reliable, effective, and sufficiently competitive.**

***Index Terms*— document analysis and recognition, digital image processing, optical character recognition**

I. INTRODUCTION

Identification cards have always been the main reference for verification and validation of an individual's identity. More often than not, government and business sectors, companies, and offices require a copy of at least one valid ID in order to register, transact, or even just apply for a new ID. Keeping up with the advancement of technology today, offices have started to observe a paperless environment by allowing an online attachment of files on the registration page, or through email, instead of the usual collection of the hardcopy of required documents. Mobile smartphones and digital cameras made it more possible and convenient to obtain a soft copy of ID cards nowadays, as such devices can now detect and recognize objects that go from the classification of coins to identification of natural images, like plate numbers and street signs, to extraction of texts from documents.

However, some companies prefer scanned documents over captured images for readability purposes, since scanning skips the image enhancement step and goes directly to the automatic transformation of a multi-colored image into a binarized image for easier text recognition. Moreover, scanned images are taken under a controlled surface which produces less noise and of high quality. On the contrary, digital images as inputs for object recognition may be challenging due to factors like skewed and distorted images, poor quality, and uneven lighting. In relation to this, most of the ID cards are often printed over a glossy paper which when taken a photo of will produce a reflection on the shiny surface. Retrieval and extraction of information from the card will then be difficult.

Akhter et al. [1] worked on the extraction of information from the national ID cards in their country, Bangladesh, by separating text lines in both the X and Y-axis and saving each as a TIFF file. TIFF files are known for being flexible for handling image and data in the same file. Also, this type of file is widely-supported by scanning and optical character recognition.

Artificial Neural Networks for Document Image Analysis, as worked on by Gori et al. [2] is another approach which used a machine learning technique, instead of using optical character recognition to consider handwritten characters and variations in font styles, colors, and sizes of texts.

The researcher presented a possible way of how information contained in personal documents, like IDs, are stored in databases. Also, in this paper, the researcher developed a system to pre-fill an information field, and validate it as well using the extracted date from the attached image document

In this paper, ID card information extraction went through a combination of machine learning and image processing techniques which resulted to a sufficiently high accuracy rate in information retrieval.

A. Scopes and Limitations

This paper considered digitally taken images of government issued ID cards and passports. Input images are limited to one ID card only, as they are tested one at a time. These images were taken as inputs and were run against a training set in TensorFlow which classified the ID cards.

Input images are assumed to have good lighting and illumination, and good quality already since these cards are assumed to be submitted as a requirement in several offices. The validity date was the primary information extracted from the IDs. However, the name was also extracted to validate and counter check the user using the attached copy of the ID. Output data was sent automatically into the database under a persons credentials.

B. Time and Place

This special problem was conducted for two consecutive semesters in the Academic Year 2017-2018 at the Institute of Computer Science, University of the Philippines Los Banos (UPLB) , College, Laguna.

II. OBJECTIVES

The general objective of this study is to implement an offline computer-based application that will:

- Classify the type of ID a particular query image belongs to,

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

- Validate and cross check the user using the ID, and
- Extract needed information from an ID, specifically the validity date and ID holders name.

III. REVIEW OF RELATED LITERATURE

A wide variety of studies have been conducted on different applications of image processing methods which aims to visually improve images for human interpretation, and to process images for machine recognition. The studies range from classification of coins to identification of natural images, like plate numbers and street signs to extraction of texts from documents. [3]

Extraction of information vary from one government issued id card to the other. This includes the size of the card, the layout of the texts, and the fonts used. These cards, as input images, must undergo multiple steps to get the accurate information needed.

Akhter et al. [1] considered manually separating region of interest from the input- text block from a scanned national ID card in their country, Bangladesh, which has the same position, design and size in every card. The framework followed a top-down structural layout analysis by first separating the block of text from the card, dividing the block into text lines, and finally extracting each word from the line. Each word detected is saved as an individual TIFF image which is a file widely-supported by scanning and optical character recognition.

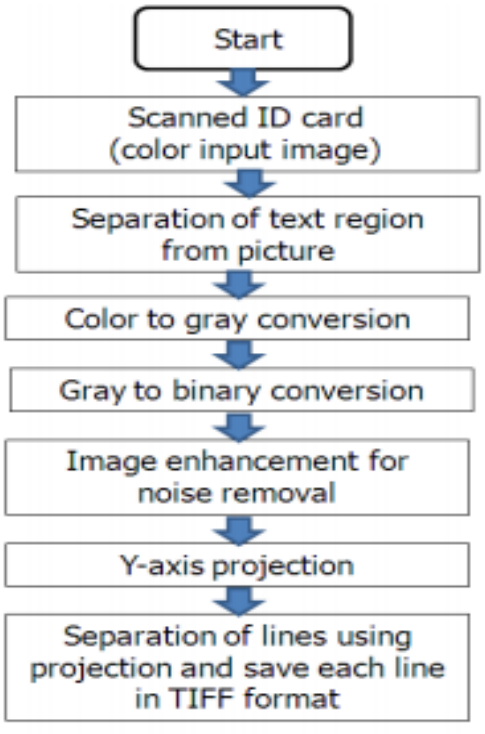

Fig. 1. *Akhter et al. [1] flow diagram of structural layout analysis for line separation.*

In the case where regions of interest are undefined, feature extraction must be recognized first to segregate parts until the main information is found. This is similar to the layout analysis done by Akhter et. al. [4]

Before text extraction, images undergo pre-processing as regions of interest might be skewed or are difficult to be identified. Expiration dates are the subject to be extracted. They come in different fonts and formats. Sometimes, a single identification card contains more than one date. Zaafouri et al. [5] proposed in his paper an approach to recognize expiration dates in any type of products using Stretched Gabor Features.

Optical character recognition provides recognition for optically-processed characters. It converts paper documents to texts. Jajulwar et al. [6] described OCR as a means to digitalized printed documents, digital images containing texts, and even scanned data for extraction, searching, editing, and data and text mining. However, the system, until now is not completely error-free as to it is highly dependent on the input's orientation and quality.

Artificial Neural Networks for Document Image Analysis, as worked on by Gori et al. [2] is another approach which used a machine learning technique, instead of using optical character recognition to consider handwritten characters and variations in font styles, colors and sizes of texts.

Analysis and Recognition can also be done in real-time and using mobiles. Kise et al. [7] compiled several papers, including some of his own works, which aimed to explain and explore Camera-Based Document Analysis and Recognition. He mentioned in his paper that ANNs are commonly used as recognition engines to consider the OCR base system for having predefined classes and the limited lexicon. The use of neural networks in the process addressed the difficulty of classifying handwritten characters, that could possibly be varying, overlapping, or simply touching other characters. Machine learning techniques incorporated in their papers include classification algorithms used in several processing levels, and collection of document images as training classifiers. In a separate study, Document Analysis and Recognition, mostly related to office documents, are suitable for machine learning as to what Marinai worked on in his paper. He classified two factors in applying machine learning techniques: first, the use of classification algorithms starting from pre-processing to character recognition, itself; second is the annotation of images that can be used to train classifiers. [8]

In a different case where a huge number of unique images are to be considered as inputs, image processing techniques alone could no longer be enough. Ezoray et al. [9] discussed the most recent trends, impacts and potentials of the integration of machine learning algorithms to the traditional image processing techniques. Cited in the paper are works on pattern retrieval based on local features, and algorithms for image classification and detection. Wang et al. [10] used the prediction and labeling of images to classify images from a pile of other images. Humans classify images easily if they know what the subject of the image is all about. Image classification is difficult for computers unless they are pre-trained of what an image contains. Annotating images is another problem where labeling each part of the whole image is made aside from classifying an image under a specific group. Other existing works used image features to address

the problem. In their paper, they addressed both problems by using estimation and prediction algorithms.

While machine learning made progress over almost any field, training the computer requires a lot of time and effort. Goma tested TensorFlow alongside other deep learning techniques to understand the similarities and differences in performance, and speedup the training process for a large-scale and distributed dataset [11]. TensorFlow, introduced by Google, is an open-source software library for machine learning and deep neural network research. The system can be used in various fields and domains.

The preceding review presents relevant methods of data extraction from a wide-ranging natural images. As similar studies on information retrieval have not been done yet on a variety of Philippine government-issued identification cards, the use of Tensorflow and image processing in this study suggests a possible way to address the problem.

IV. METHODOLOGY

This chapter provides the list of applications that should be installed, as well as the minimum specifications of the computer to run the system. Also, this chapter serves as an outline of the research methodology that was used to achieve the objectives of this study.

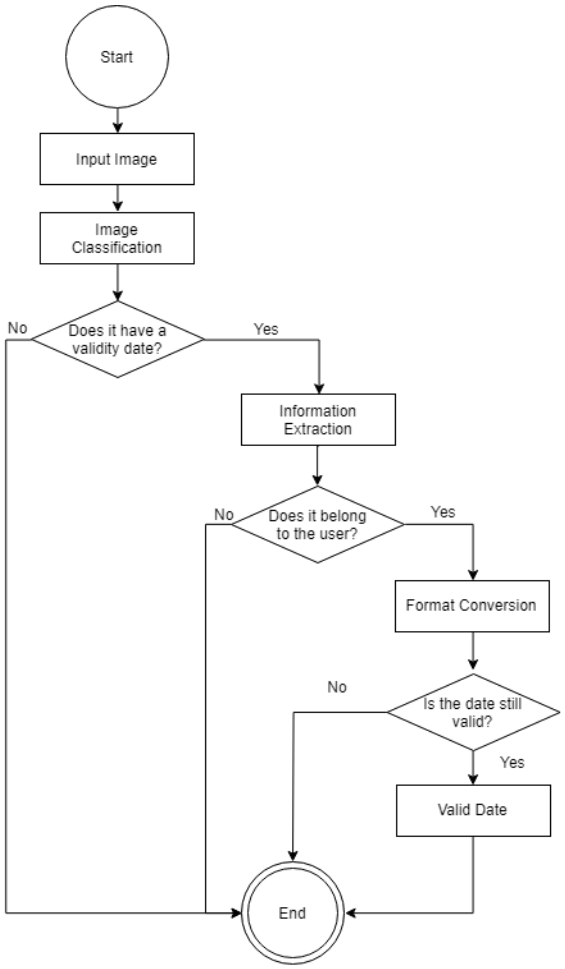

Fig. 2. *Proposed Data Extraction Method for ID Cards.*

A. Development Tools

The system is an offline application developed on a personal computer running on Windows 10, Intel Core i3-6006U CPU @ 2.00GHz with 4GB of RAM, and a storage of 917 GB HDD. The application was developed using TensorFlow for Poets for image classification, Python version 2.7.12 OpenCV for image processing, and OCR Tesseract for data extraction. MySQL was used as the database for storage.

B. Process Model

The flow of data is divided into three main parts: a.) Image Classification, b.) Image Processing, and c.) Data Extraction and Manipulation. Query images undergone all three to achieve the goal. A scanned or a digital image of a card was considered as a query image. It is restricted to a single card per image only. Identification cards considered were government issued IDs only, which may or may not have expiration dates, and may or may not be valid on the present date. Sample IDs include passports, drivers license, professional regulation commission ID, SSS and GSIS ID.

1) Image Classification: Input images were classified using a pre-trained open-source library, TensorFlow, which was developed and maintained by Google.

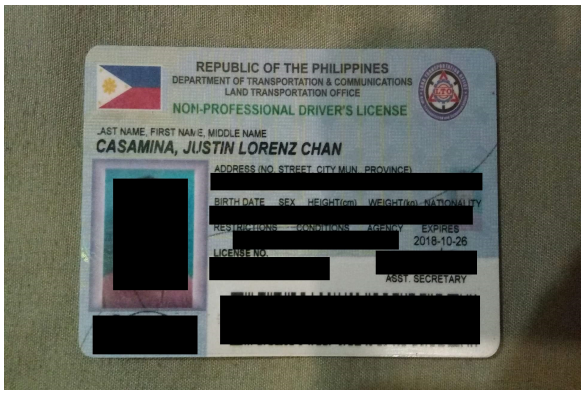

Fig. 3. *Sample Query Image*

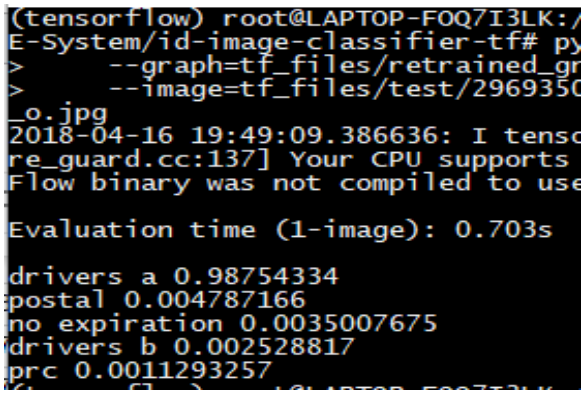

Fig. 4. *Classified type of ID*

Training sets are compiled into one folder containing cards of the same type. TensorFlow compared the input image to

all possible types and returned the classification with the highest probability of likeness which informed the system where it belongs. The ID card's classification, together with the image file name, were returned to the system which served as parameters in the next step - image processing.

2) Image Processing:

- Pre-processing- OpenCV (Open Source Computer Vision) version 3.2 is a library using C functions and C++ classes, commonly used for image manipulation and processing. However, Python version 2.7.12 was used to conform with TensorFlows implementation.
  Preprocessing includes noise removal, edge detection, and extraction of the region of interest. Input images undergone smoothening and noise removal, first, using Gaussian filter, which is commonly and widely used for that function. The background and foreground were separated through binary image processing. Lastly, texts were treated as the foreground which were turned into white.

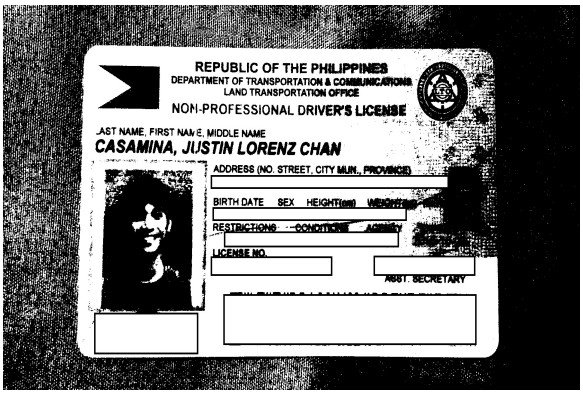

Fig. 5. *Thresholding to locate the cards from the background*

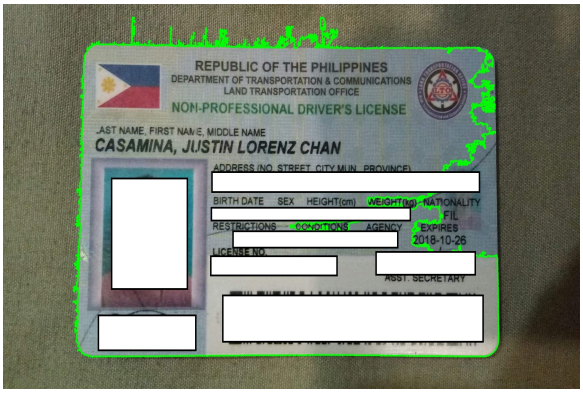

Fig. 6. *Connected Components to detect borders of the card*

- Segmentation and Localization- Images contained a single card only. Canny edge detection was used to locate the card in the image. The image was then resized to maintain a uniform size per each type. With that, position of the needed information remained almost the same in every image. Each information were saved separately as regions of interest.
  Dilation was also applied to expand components of the texts. Lastly, regions of interest were dependent from the IDs classification. A defined set of values were used as coordinates to form a rectangle. The cropped rectangle included the card holders name and the cards validation date, and therefore, were treated as the regions of interest.

Fig. 7. *Located ID Card*

Fig. 8. *Located card holder's name*

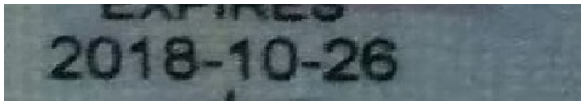

Fig. 9. *Located validity date*

3) Data Extraction and Manipulation: Given the cropped image containing the name and date, OCRs Engine, Tesseract, was used to extract the information left on the image.

Names from ID cards are usually separated. The system concatenated each component of the name as it belong to the same attribute in the database. The system converted the extracted date into a date format as it originally was of the type string. Also, date formats vary from one card to another and from the date format used in the database. Extracted dates were parsed per part and were rewritten in the format YYYY-MM-DD, which was used in MySQL.

Fig. 10. *Connected Components*

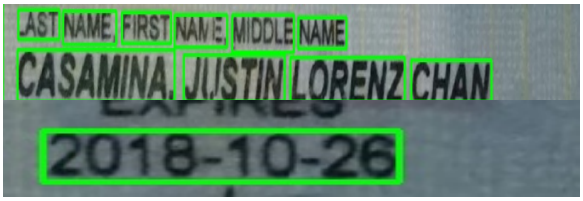

Fig. 11. *Contours*

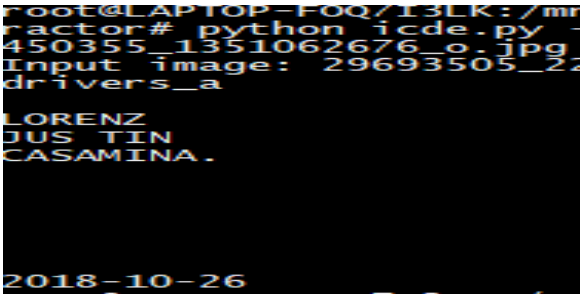

Fig. 12. *Extracted Information*

V. RESULTS AND DISCUSSION

A. Evaluation of Experimental Results

A total of 350 Identification cards of different types were gathered to be used in the system. 280 (80% of 350) of these cards were used as training data for the image classifier, and 70 (20% of 350) for testing.

TABLE I
Table containing the ID types considered in this study including the number of data per type

ID Types	Total number of Data	20% of Data
Drivers_a	65	13
Drivers_b	45	9
No_expiration	75	15
Passport	75	15
PRC	90	18
Total	350	70

TABLE II
Tally of correctly identified and classified IDs

ID Types	Image Classification	Data Extraction
Drivers_a	13/13	11/13
Drivers_b	9/9	8/9
No_expiration	15/15	13/15
Passport	15/15	11/15
PRC	18/18	14/18
Total	70/70	57/70

Each type has a folder containing the same type of IDs that vary on perspective, orientation, and even number of cards per image. The folder containing cards with no expiration has a variety of cards including SSS, GSIS, UMID IDs. 15 cards in the test data have no expiration dates, while 55 of these cards have. The accuracy of the classifier is dependent on the uniformity of cards in a single classification or folder. By removing unnecessary training data, the system avoided a 1.8182% error from previous test, which was obtained due to the insufficient and unequal distribution of training images in the No_expiration folder.

B. Discussions

TABLE III
Cross-Validated Confusion Matrix for ID Card Data Extraction

		Actual: Valid ID (58)	Actual: Invalid ID (12)	
Predicted	Valid ID (49)	49	0	PPV=100%
F1-Score = 84.48%	Invalid ID (21)	9	12*	NPV=57.14%
		TPR=84.48%	SPC=100%	ACC=87.14%

* Four (4) invalid IDs are safely considered as invalid because the system cannot read any word from the image

Different performance measures were computed to test the system's credibility. The result of the observations (Table 3) clearly show that it is safe to say that the system properly identified 84.48% out of all IDs that are valid, and 100% of the time, it is correct. Also, 100% of the time, the system correctly identified invalid IDs. The small number (9 out of 70 cards tested) of cards with false negatives only gave us a 57.14% confidence to say that the IDs, which were identified as invalid, are true.

We then take note of the accuracy rate of 87.14% which indicates a high reliability and effectiveness of the system. Lastly, an F1-Score of 91.59% suggests that the system is sufficiently competitive.

VI. CONCLUSION

In this paper, an offline desktop application was implemented for the purpose of automatic classification and data extraction of a variety of Government-issued IDs. The system was able to classify the ID type that a query image belongs to, among five classifications of IDs considered in this study. Also, the system was able to counter check if the ID belongs to a particular person or not using the extracted card holder's name. Lastly, the system was able to pre-fill an information, and validate it as well, using the extracted date from the attached image document.

The system was proven to be effective and reliable as it yielded an accuracy rate of 87.14% in eliminating redundant forms by automatically validating and filling up fields with important information extracted from the attached image document. Also, an F1-Score of 91.59% suggests that the system is sufficiently competitive.

VII. RECOMMENDATION

In this study conducted, however, it is only suitable for identification cards that are taken flat, and images of high quality. To further enhance this application, future developers can perform a perspective crop to accommodate IDs that are taken from a lower or higher angle. Also, school IDs can be included in the list of IDs of interest as they are also accepted in offices. Lastly, since the application is only computer-based, future developers may expand and create an Android, iOS or other platform version for mobile access.

REFERENCES

[1] R. Akhter, H. Bhuiyan, and S. Uddin, "Extraction of words from the national id cards for automated recognition," 2010. [Online]. Available: https://www.researchgate.net/publication/258549622_Extraction_of_Words_from_the_National_ID_Cards_for_Automated_Recognition

[2] M. Gori, S. Marinai, and G. Soda, "Artificial neural networks for document analysis and recognition," Tech. Rep., 2003. [Online]. Available: http://www.dsi.unifi.it/~simone/ANNxDAR/TR-DSI-01-03.pdf

[3] K. J. Abriol-Santos, "Image enhancement," 2017.

[4] O. Augereau, J.-P. Domenger, and N. Journet, "Document image recognition and classification," 2014. [Online]. Available: https://www.researchgate.net/publication/265008808

[5] A. Zaafouri and F. Sayadi, M.and Fnaiech, "A vision approach for expiry date recognition using stretched gabor features," *International Arab Journal of Information Technology*, 2015. [Online]. Available: http://ccis2k.org/iajit/PDF/Vol%2012,%20No.%205/6700.pdf

[6] K. Jajulwar, "Raisoni college of engineering session: 2016-2017 semester/branch/section: vii / etc c name of subject: Optical communication tae-2 optical character recognition." 2017.

[7] K. Kise and D. Doermann, "Camera-based document analysis and recognition," vol. 2, 2007. [Online]. Available: http://imlab.jp/cbdar2007/proceedings/proc_full.pdf

[8] S. Marinai, "Introduction to document analysis and recognition," 2008. [Online]. Available: https://www.springer.com/cda/content/document/cda_downloaddocument/9783540762799-c1.pdf?SGWID=0-0-45-480217-p173779118

[9] O. Lezoray, C. Charrier, H. Cardot, and S. Lefevre, "Machine learning in image processing," *EURASIP Journal on Advances in Signal Processing*.

[10] C. Wang, D. Blei, and L. Fei-fei, "Simultaneous image classification and annotation," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, 2009. [Online]. Available: http://www-cs.stanford.edu/groups/vision/documents/WangBleiFei-Fei_CVPR2009.pdf

[11] M. Yages Gom, J. Torres Vials, and R. Tous Liesa, "Image recognition with deep learning and tensorflow," 2016. [Online]. Available: https://upcommons.upc.edu/bitstream/handle/2117/106009/118046.pdf?sequence=1&isAllowed=y