Segmentation And Recognition Of Information Printed On Identification Cards

Madhana Gopalapuram Ramesh Rathish Aadiraja

Warsaw University Of Technology, Warsaw, Poland  
01139057@pw.edu.pl

**Abstract.** In the current world of online services, some of them require a remote automatic verification system to verify and validate their users. In these cases the users are asked to provide the photo of their ID card as a proof of verification along with their current picture and their personal information for cross verification. This paper is focused on localizing the ID card from the photo sent by the user using convolutional neural network architecture U-net followed by recognizing the information printed on the localized ID card by using open source Tesseract OCR Engine with the help of open source dataset MIDV-2020 consisting various pictures of artificially generated ID cards which resembles so close to the original ID cards.

**Keywords:** CNN, U-net, Semantic Segmentation, OCR, LSTM

1. Introduction

Due to the increase in demand for online services, most offline services are attempting to transition into the online environment so users may utilize them with ease and complete activities from wherever they want. Among those, some of them such as digital banks, e-commerce websites, cryptocurrency investing applications, car sharing applications validates the user before allowing them to access any of their features. Since these services are available online, companies must authenticate their user’s identity remotely and certify that they are legitimate and eligible to utilize them. The user is verified by just providing a photo of their ID card and a self-portrait, the verification is accomplished in a matter of minutes without the requirement for the user to be present in person at any location. These online services are also required to keep user’s sensitive data secured. The main steps involving this verification process are cropping out the ID card from the whole image, verifying the integrity and textual information followed by a face detection by comparing the current photo and the photo printed on the ID card.

Normally, the photos of the ID cards are taken by the users using their mobile phones with variable backgrounds, occlusions and with different perspective, orientation, lightings and sometimes with low resolutions. To verify the information printed on ID cards it is important to have a clear picture of that ID card in the first place so before obtaining the photo from the user these applications let them know in real time that they must adjust the ID card in the prefect frame with better lightings and resolutions to make the verification more robust. These small steps prevent difficulties in retrieving the information printed on the ID cards. This paper focuses on localizing the ID card from the whole image, preprocessing the ID card image, identification of specific ID type, segmentation of ID fields, recognition of textual information printed and verifying the retrieved information on the ID card. These tasks altogether plays a crucial part in the remote automatic verification systems.

1. Dataset

Unlike the dataset of images of cats, dogs or cars which is easily accessible, real photos and are available in thousands and millions, the dataset of images of ID cards are scarce. Also, it is difficult to collect images of actual ID cards of people since these images contain their private information such as their addresses. Even if the dataset is possibly obtained it is highly important to keep these information as securely as possible. There are few open source datasets of ID cards available and one of them is MIDV 2020 [1] which is used in this paper. It consists of photos, scans and videos of 1000 different ID cards of 10 different countries which has variable text fields and unique artificially generated faces. These photos are taken using mobile phone with different background, perspective, lightings and orientations. The information presented on these ID cards are not actual information of people and it is so close to real ID cards which makes them more applicable for achieving perfect verification without having to deal with original ID cards. This work mainly focuses on Albanian and Spanish ID card types.



**Fig. 1.** Example template images from MIDV-2020 dataset

1. Different Methods for Localization of ID Card From The Input Image

A method based on detection of quadrilateral of document border in the image[2]. It is a combination of contour and region based approach. It is a modified contour approach in which the contours detected are ranked according to contrast between areas inside and outside the border. The quadrilateral with the highest score is selected and it is compared with the ground truth quadrilateral based on Jaccard Index (intersection over union). If the Jaccard index value is greater than the threshold coefficient then the quadrilateral is considered as a valid one.

A method which simultaneously locates the document and recognizes its class then, in next steps the document nature, country, version and the visible side is determined [3]. For each type of ID card, one reference model is created and the keypoints are extracted and classified by SURF method. Each reference model is indexed with random KD- trees. From the query image the keypoints are extracted and matched against all reference models at the same time with a matching score. Then the reference models are reverse matched against the query image and compared with the symmetric mapping of couple of keypoints of direct and reverse matching, histogram of orientation difference between these points and geometric transformation using RANSAC and given a score for each then compared with the previous score. The one with the highest score is considered to be a valid reference model. Then the quadrilateral of the document is detected and verified with set of validations.

A method using a sliding window to detect every region of the image if an object of interest is located[4]. For each window the occurrences of Gradient Orientation(HOG) in the certain portion of the image is calculated. Then each window is classified by SVM either if it contains the document or the background.

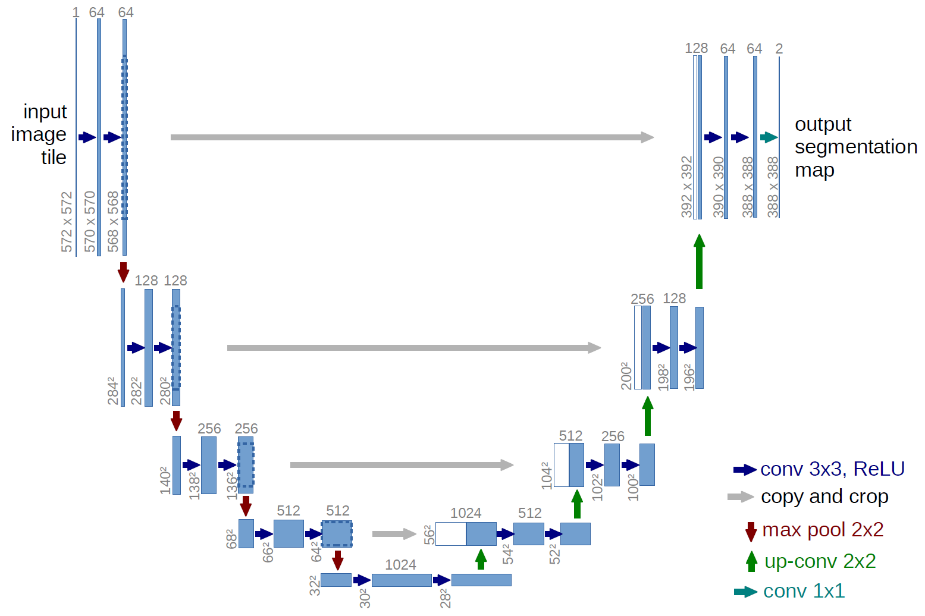
A method based on DenseNet10[4]. The architecture includes Bach normalization, ReLU, Convolutional layer and dropout. It is implemented by using 3 blocks and 10 layers making it more lightweight for mobile device applications. This is done by having one up-sampling path which has one transition up(1 TU) and one down-sampling path with one transition down(1 TD).

A method which uses the information of color difference of the ID card and the background for localization of the ID card[5]. It works by adjusting the vertices of the documents iteratively using data of pixels sampled in outer region of the ID card in the image. It has a priori assumption that the document is approximately in the center of the input image. Another set of vertices start at 30% of the edges of the document which occupies 70% of the whole images and adjusts iteratively until it finds the perfect matching vertices that contains the ID card in the entire image. After the localization it is classified using CNN.

A method that spots the ID card and accurately localizes it from the original image using specific ID document features[6]. It requires a classification a priori along with the list of predefined models. Classification is performed by local key descriptor matching process (SURF) algorithm and RANSAC is used after this step to estimate transformation matrix. A multi-hypothesis approach which runs different crop solutions in parallel and selects the one with the highest score as the cropped document. The main features that are used in the feature extraction and cropping methods are keypoints in the image, vanishing points, MRZ, document border, document corners, photo of a person’s face in the document, landmarks, head pose of the person in the document, logo in the document, variable and invariable fields of the document. Among these features a best set of hypothesis of features are selected to produce the best crop of ID document in any different situation.

1. Localization Using UNET Architecture

U-net is a convolutional neural network (CNN) architecture that was primarily developed for biomedical image segmentation by Olaf Ronneberger, Philipp Fischer, and Thomas Brox[7]. It can successfully perform semantic segmentation with just scant amount of images. By performing semantic segmentation, each and every pixel in an image can be labelled as an object of interest, which is the ID card or the background. This method enables a highly precise prediction of location of ID card in the image. This architecture can be viewed as two main parts, the encoder and decoder path. The encoder path is the contracting path which captures the context, performs a classification and the decoder path is the expanding path which predicts precise localization.



**Fig. 2.** U-net architecture

The above architecture is from the U-net paper that was published originally, at first the input image is taken with the dimensions of 572 x 572 x 1 which is the length, width and number of channels of the image. Where image with single channel represents a binary image. Then a 3 x 3 convolution operation is applied to this image with 64 filters, no padding and with stride of 1. This is followed by a non-linear activation function, rectified linear unit (ReLU). Again for this feature map the same 3 x 3 convolution operation is applied with a ReLU function. After these two convolution operations a 2 x 2 maxpooling operation with stride of 2 is performed, which reduces the length and width of the feature map by half. Now these three steps are repeated for three times which is series of convolutions and maxpooling operations with a small change where, the number of filters doubles in each downsampling step. After these steps the resulting feature map is of dimension 28 x 28 x 1024, which is a dense layer with 1024 feature channels. Now, the expanding path starts with a 2 x 2 up convolution operation, which halves the number of feature channels in every step so the dimensions are changed to 54 x 54 x 512 . This resultant feature map is concatenated with the result that was obtained in each step of downsampling, resulting in a feature map of dimension 54 x 54 x 1024. This concatenation helps in producing a high resolution mask in the segmented result. Again two 3 x 3 convolution operations are performed with ReLU followed by a 2 x 2 up convolution and concatenated with the corresponding result from the downsampling step. Finally from the feature map of dimension 388 x 388 x 64 a 1 x 1 convolution operation is performed to get the final desired number of classes with the dimension of 388 x 388 x 2. The output image has two channels in which one channel is for the foreground class in this case the ID card and the other for the background class. Due to the unpadded convolution operations the final output image is smaller than the input image. This architecture consists a total of 23 convolution layers it takes a raw image as an input and outputs a segmentation mask.

It is not required to use the exact dimensions and number of filters as proposed in the U-net paper but it is necessary that the length and width of the input image should be equal and an even number so the 2 x 2 maxpooling operations produces better results and even number of filters which is doubled every time in the encoder path and halved every time in the decoder path. This proposed approach is used in this paper for sematic segmentation of ID card from the original image. To train this model, the architecture requires sets of images and its ground truth masks where the ID card is present in respective images. In MIDV-2020 dataset the annotations were given in a file with exact pixel values of vertices of the corners of ID cards in the image which can be masked for the area under the quadrilateral formed by these vertices or can be annotated manually. Even though the dataset consists of less number of images, this architecture can be trained from scratch with the availability of these images resulting precise segmentation.

After obtaining the segmented output from the architecture the accuracy of this localization is verified by finding the value of a Generalized Intersection over Union[8] which is Area of overlap divided by Area of union of the ground truth mask and the predicted mask by the U-net architecture. The value ranges from 0 to 1 where the value close to 1 being more accurate prediction. To crop out the exact ID card from the image a simple corner detection is performed to obtain the pixel values of corners of ID card. If any corner of the ID card is occluded in the image a linear regression of contours[9] is performed on the edges of the ID card to determine the corners, where these contours meet. From the values of these corners a four point perspective transformation is performed to align the perspective of photos of ID cards taken by the users which results in an image containing only the ID card without the background and with the perspective of a scanned document. After these series of steps the resultant image of segmented ID card can be considered for recognition of the information printed on them.

1. Text Recognition Using Tesseract OCR

Tesseract is an open source optical character recognition (OCR) engine which was initially developed at Hewlett-Packard Laboratories and is currently maintained by Google[10]. The engine recognizes the patterns of each character of textual information in the image and matches with the corresponding alphabets and symbols. It also supports line recognition which is achieved by LSTM neural networks architecture. It supports over 100 languages and also has the flexibility to be trained for newer languages. To achieve the best recognized results the resolution of the image and the textual information should be of better quality before passing it into the engine. The preprocessing of the image plays a key role in achieving best results. After obtaining the resultant segmented image from U-net, cropping it and changing the perspective, further more preprocessing techniques are required before passing the image to the OCR machine. Some of them are glare detection, converting the image to grayscale, enhancing the color of textual information on the ID card.

An ID card contains both variable and non-variable text fields. ID card of same type contains all textual information at exact location so the OCR engine is guided to look for those information at a specific location of the image. Since the non-variable text fields are exactly same in specific type of an ID card, the engine can focus more on the variable parts which contains actual information of the card holder. In addition to that, fields like date of birth, age it can be set to look only for numbers and symbols without needing to look for any alphabet and vice versa for the surname and name felids, which makes the engine predict better results. Finally these retrieved information is verified before checking with the actual information provided by the user while registering for the online service using the application. This text retrieval can also be used for automatic data entry which saves more time when compared with manual verification and manual data entry.

1. Conclusion

This work purely focuses on segmenting the ID card from the whole image and retrieving the information from the segmented image and verifying those textual information which is applied in real world remote automatic verification systems for online services. Improvements can be made in the cropping technique where the resolution of cropped picture can be enhanced to make the textual information more readable for OCR engine. It can also be improved by using a different localization technique or different OCR machine which can provide a better accuracy and a better training time. This work can be extended by training additional ID types all around the world by using additional datasets of those ID cards. It can be embedded to a whole process where further works like facial recognition and ID card integrity verification can be added providing a whole end to end remote automatic verification system and for identity fraud prevention.

References

1. K.B. Bulatov, E.V. Emelianova, D.V. Tropin, N.S. Skoryukina, Y.S. Chernyshova, A.V. Sheshkus, S.A. Usilin, Z. Ming, J.-C. Burie, M. M. Luqman, V.V. Arlazarov: “MIDV-2020: A Comprehensive Benchmark Dataset for Identity Document Analysis”, Computer Optics (submitted), 2021.
2. Tropin, Daniil V., et al. "Approach for document detection by contours and contrasts." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.
3. Awal, Ahmad Montaser, et al. "Complex document classification and localization application on identity document images." *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE, 2017.
4. Lara, Rodrigo, et al. "Towards an Efficient Semantic Segmentation Method of ID Cards for Verification Systems." *arXiv preprint arXiv:2111.12764* (2021).
5. Attivissimo, Filippo, et al. "An automatic reader of identity documents." *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019.
6. Chiron, Guillaume, Nabil Ghanmi, and Ahmad Montaser Awal. "ID documents matching and localization with multi-hypothesis constraints." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.
7. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
8. Rezatofighi, Hamid, et al. "Generalized intersection over union: A metric and a loss for bounding box regression." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
9. Castelblanco, Alejandra, et al. "Machine learning techniques for identity document verification in uncontrolled environments: A case study." *Mexican Conference on Pattern Recognition*. Springer, Cham, 2020.
10. Smith, Ray. "An overview of the Tesseract OCR engine." Ninth international conference on document analysis and recognition (ICDAR 2007). Vol. 2. IEEE, 2007.