

KYC METHOD AND KYC PLATFORM FOR CORRECTION OF RAW TEXT WITHOUT USE OF TEXT REGIONS

5

TECHNICAL CONTRIBUTION

The present disclosure relates to know your customer review of validation requests. More particularly, the present disclosure relates to a know your customer computer-implemented method or know your customer platform for correction of raw text generated by OCR from
10 images of customer identification documents having printed text.

BACKGROUND

Know your customer or know your client (KYC) is typically a computerized validation process for customer biographical text and identification documents collected from online
15 websites. Customers or staff of a business enter biographical text into the data entry fields of a webpage and upload an image of a form of identification (hereafter, "the ID image") to create a validation request.

The biographical text typically includes names, identification numbers, dates of birth,
20 addresses, effective dates, and/or addresses. The biographical text can be structured using field restrictions in the data entry fields of the data entry webpage, for instance by use of pop-up calendars for the input of dates. For computerized analysis of the ID images, the printed text of the ID images must be captured conducting OCR upon the ID image. The ID images must undergo OCR during the validation process because the biographical data printed on the
25 ID image must be compared against the biographical text submitted with the validation request. The biographical data printed on the ID image may also be compared against other proprietary or public information collected separately from the incoming validation request.

Processing the images of customer identification documents poses various challenges. While
30 each document type may include the same basic biographical data (such as name and date of birth), the template for the placement and presentation of this information varies between each document type. Different document types, for instance, may list the basic biographical data in a different order with different font sizes, and different security features. The

readability of the images may also vary due to image resolution, lighting conditions, font size, security features, and damage to the documents themselves.

5 The formatting and content of identification documents varies between government agencies and may also vary depending on the year of issuance of the identification documents. Philippines identification document types, for instance, can include a driver's license, Government Service Insurance System (GSID) card, a National Bureau of Investigation (NBI) card, a passport, a Philippine Health Insurance Corporation (PhilHealth) card, a postal identity card, a Professional Regulation Commission (PRC) card, a Social Security System
10 (SSS) card, a Tax Identification Number (TIN) card, a Unified Multi-Purpose ID (UMID) card, and a voter ID card. KYC review is adapted to various types of identification or documentation based on the jurisdiction (e.g., country or state) and/or the business needs (e.g., financial, government, etc).

15 An overview of how such a KYC method or KYC platform can be constructed is provided in the article entitled "How we built a modern, state of the art OCR pipeline", AB Saravanan, <https://blog.signzy.com/how-we-built-a-modern-state-of-the-art-ocr-pipeline-preciousdory-dc3a4ae0e87>, 06 October 2018. In this Saravanan article, the author discusses the partitioning of the total ID image area into a series of cropped images, one cropped image for each of a
20 series of text regions. Each text region is identified as a specific area on the ID image set aside for the printing of biographical data for a specific data field. The specific areas can be set aside, for instance, for the printing of the biographical data into a name data field, an identification number data field, an address data field, or a date of birth data field. Each text region is separately processed by the OCR engine and error correction module. The
25 supervised training data for the error correction model includes paired sets of cropped images of specific text regions (from a larger ID image) and the corresponding corrected text for that cropped image. A similar text region approach is also taken in the published patent application numbered WO2020141890 entitled "Method and Apparatus for Document Management" dated 09 July 2020.

30 The driving business demands for a KYC method or KYC platform are its accuracy, speed, throughput, and cost. If the validation requests are not performed accurately, a fraudulent account could be established. If the validation request are not performed quickly or the queue

of pending validation requests backs up, potential customers submitting validation requests could be lost to competitors. If the costs are high, particularly the cost per validation request, running cost may balloon during customer growth periods. Though many software vendors provide turnkey software products or services to perform KYC, the fees for such software or service typically scale with the number of validation requests processed. Especially for financial industries with a large potential customer base, the number of validation requests can be in the hundreds of thousands or more each week, making cost a key concern.

While text regions can be used to further refine and specialize the OCR and error correction process used on each specific text region, use of text regions adds an additional complexity and marries the KYC method or KYC platform to a limited selection of OCR engine choices. What is needed is a KYC architecture that decouples the error correcting step from the OCR conversion step such that a low cost and robust OCR engine can be selected without impacting the error correction step.

SUMMARY

A general embodiment of the invention is a KYC method or KYC platform for receiving and processing a plurality of validation requests which include ID images. Validation requests with fraudulent ID images are red flagged using facial matching. Document types for each ID image are identified without the use of text regions to produce raw text. Raw text from the ID images is post-edited to create corrected text. Post-editing includes use of tailored error correction models associated with the document type of the ID image.

A first embodiment of the invention is a computer-implemented KYC method, the method comprising: (a) receiving at a request database in data communication with a network a plurality of validation requests through a network; (b) storing each validation request in one of a plurality of request records of the request database; (c) facial matching to compare each photograph to each of a plurality of library images to determine a duplication score for the validation request associated with the photograph; (d) identifying a document type of each ID image; (e) conducting OCR to convert the one or more ID images of each validation request to raw text; (f) post-editing the raw text to execute sequence-to-sequence error correction upon the raw text of each ID image of each validation request to create corrected text for each ID image of the validation request; and (g) evaluating each validation request to compare the

biographical text delineated into the first plurality of data fields of the validation request against the corrected text delineated into the second plurality of data fields of the validation request to determine a deviation score for the validation request. Each validation request includes biographical text and one or more ID images purportedly identifying a customer associated with the validation request. The biographical text for each validation request is delineated into a first plurality of data fields within the request record for the validation request. At least one of the one or more ID images of each validation request includes a photograph of the customer. The duplication score of each validation request is evaluated according to a duplication threshold. The facial matching step includes red flagging each validation request exceeding the duplication threshold. The conducting OCR step does not identify a plurality of text regions in each ID image. The raw text created in the conducting OCR step does not include fonts or formatting. The post-editing step includes accessing a plurality of tailored error correction models, each tailored error correction model created by supervised learning specifically for one of the document types without the use of text regions. For each ID image of the validation request, the post-editing step includes selecting the tailored error correction model associated with the document type of the ID image, as identified in the identifying step, to: (1) execute the sequence-to-sequence error correction upon the raw text of the ID image of the validation request; and (2) delineate the corrected text of the ID image into a second plurality of data fields for the validation request. The deviation score of each validation request is evaluated according to a deviation threshold. Each validation request exceeding the deviation threshold is red flagged.

A second embodiment of the invention is a KYC platform for receiving and processing a plurality of validation requests, the platform comprising: (a) a request database in data communication with a network; (b) a facial matching module configured to compare each photograph to each of a plurality of library images to determine a duplication score for the validation request associated with the photograph; (c) a document type identifier configured to determine a document type of each ID image; (d) an OCR engine configured to convert the one or more ID images of each validation request to raw text; (e) a post-editor configured to execute sequence-to-sequence error correction upon the raw text of each ID image of each validation request to create corrected text for each ID image of the validation request; and (f) an evaluation module configured, for each validation request, to compare the biographical text delineated into the first plurality of data fields of the validation request against the

- corrected text delineated into the second plurality of data fields of the validation request to determine a deviation score for the validation request. The request database is configured to receive the validation requests through the network and store each validation request in one of a plurality of request records of the request database. Each validation request includes
- 5 biographical text and one or more ID images purportedly identifying a customer associated with the validation request. The biographical text for each validation request is delineated into a first plurality of data fields within the request record for the validation request. At least one of the one or more ID images of each validation request includes a photograph of the customer. The library images are stored in the KYC platform or are accessible to the KYC
- 10 platform through the network. The duplication score of each validation request is evaluated according to a duplication threshold. The facial matching module red flags each validation request exceeding the duplication threshold. The OCR engine is not configured to identify a plurality of text regions in each ID image. The raw text created by the OCR engine does not include fonts or formatting. The post-editor includes a plurality of tailored error correction
- 15 models, each tailored error correction model created by supervised learning specifically for one of the document types without the use of text regions. For each ID image of the validation request, the post-editor selects the tailored error correction model associated with the document type of the ID image, as identified by the document type identifier, to: (1) execute the sequence-to-sequence error correction upon the raw text of the ID image of the
- 20 validation request; and (2) delineate the corrected text of the ID image into a second plurality of data fields for the validation request. The deviation score of each validation request is evaluated according to a deviation threshold. Each validation request exceeding the deviation threshold is red flagged.
- 25 A third embodiment of the invention is a computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out any one of the methods of the first embodiment and the methods of the alternatives of the first embodiment.

A fourth embodiment of the invention is a computer-readable storage medium comprising instructions which, when executed by a computer, cause the computer to carry out any one of the methods of the first embodiment and the methods of the alternatives of the first embodiment.

5

In contrast to the background art, the invention does not employ text regions during the OCR conversion step by the OCR engine or the post-editor. The OCR engine of the invention performs OCR on each ID image as whole (rather than separating the ID image into a series of text regions). The post-editor of the invention similarly performs error correction upon the raw text of each ID image as a whole (rather than separating the ID image into a series of text regions). The tailored error correction model used by the post-editor of the invention are optimized to each document type through supervised learning. The document type for each ID image is determined prior to the post-editing step and this document type information is fed into the post-editor so that it may select the appropriate tailored error correction model.

15

Without the need for OCR engines with text region capability, the OCR engine can be selected for use with ID images of various photographic quality at a low cost and with high image-processing throughput. Hence, optimization of the OCR engine's creation of raw text and the post-editor's creation of corrected text can be independently optimized.

20

The architecture of the invention also enables easy swapping of OCR engines, as the required output of the OCR engine is limited to raw text. In the event an OCR engine with improved performance on ID images of various photographic quality or an OCR engine provided at a lower cost becomes available, the invention's architecture enables the previous OCR engine to be swapped out and replaced with the new OCR engine.

25

Another technical benefit of the invention is that it does not strive to output a final decision for every validation request. To the extent that the invention is unable to meet a final decision on the validation request, the invention sends the validation request to a second level review.

30

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the present disclosure are described herein with reference to the drawings in which:

5 **FIG. 1** is a flowchart representation of a computer-implemented KYC method in an embodiment of the invention.

FIG. 2 is a block diagram representing a KYC platform in a KYC system in an embodiment of the invention.

10

FIG. 3 is a textual flowchart representation of a in an embodiment of the invention.

FIG. 4 is a block diagram representing an identifying step for a document type in an embodiment of the invention.

15

FIG. 5 is a chart detailing the accuracy rates of document type identification in an embodiment of the invention.

20 **FIG. 6** is a chart detailing the accuracy rates of post-editing error correction in an embodiment of the invention.

DETAILED DESCRIPTION

In the following detailed description, reference is made to the accompanying drawings, which form a part hereof. The illustrative embodiments described in the detailed description, drawings and claims are not meant to be limiting. Other embodiments can be utilized, and other changes can be made, without departing from the spirit or scope of the subject matter presented herein. Unless specified otherwise, the terms “comprising,” “comprise,” “including” and “include” used herein, and grammatical variants thereof, are intended to represent “open” or “inclusive” language such that they include recited elements but also permit inclusion of additional, un-recited elements. As used herein, a computer or server may include a single stand-alone computer, a server, multiple dedicated servers, and/or a virtual server running on a larger network of servers and/or cloud-based service. As used herein, a data processing system may reside within a single stand-alone computer, a server, multiple

dedicated servers, a cloud-based service, and/or a virtual server running on a larger network of servers. As used herein, a database may store data to and access data from a single stand-alone computer, a data server, multiple dedicated data servers, a cloud-based service, and/or a virtual server running on a larger network of servers.

5

FIG. 1 is a flowchart representation of a computer-implemented KYC method **10** in an embodiment of the invention. Validation requests **11** include biographical text **11A** and at least one ID image **11B**. The ID image **11B** is the input for a facial matching step **1-01**, a document type identifying step **1-02**, and a conducting OCR step **1-03**.

10

In a facial matching step **1-01**, the ID images **11B** are compared against a plurality of library images (not illustrated) to determine a duplication score for the validation request **11**. Validation requests **11** with a duplication score exceeding a duplication threshold are red flagged and either rejected outright or fed into a second level review step **1-07**.

15

Validation requests **11** with a duplication score not exceeding the duplication threshold are further processed in the document type identifying step **1-02** to determine a document type (identified in **FIG. 1** as document types **01 . . . n**).

20 The conducting OCR step **1-03** is performed upon each ID image **11B** to create raw text **12**. Raw text **12** does not include fonts or formatting. Raw text **12** could be represented using standard text codes such as ASCII, CCCII, or Unicode. The conducting OCR step **1-03** is intentionally limited to use of a simple OCR engine **24E**, avoiding the capture of distinct font or formatting information. The conducting OCR step **1-03** does not delineate text fields as in
25 the background art.

The post-editing step **1-04** has two inputs. A first input is the document type determined in the document type identifying step **1-02** (identified in **FIG. 1** as document types **01 . . . n**). The document type is used in the post-editing step **1-04** by the post-editor **24F** to select a
30 tailored error correction model **13** (identified in **FIG. 1** as MODEL_{01 . . . n}) for execution of a sequence-to-sequence error correction upon the raw text **12** to produce corrected text **14**. Corrected text **14**, like raw text **12** does not include fonts or formatting.

The corrected text **14** is compared against the biographical text **11A** submitted by the customer with the validation request **11** in an evaluation step **1-05** that determines a deviation score. The corrected text **14** can also be compared against third party data records **15** (from one of a plurality of third party data servers, as depicted in **FIG. 2** item **21**). Validation requests **11** with a deviation score exceeding a deviation threshold are red flagged and either rejected outright or fed into a second level review step **1-07**. Validation requests **11** with a deviation score not exceeding the duplication threshold are deemed to have entered a passing step **1-06**. Third parties could be governmental agencies, non-governmental agencies, credit agencies, partner financial institutions, and/or any other type entity with potentially relevant data for the KYC process.

The passing step **1-06** typically instigates further processing of the application for which the validation request **11** was submitted by the customer. For example, once a potential banking customer's validation request **11** makes it through the KYC requirements, the banking customer bank account request application moves forward toward activation.

FIG. 2 is a block diagram representing a KYC platform **24** in a KYC system **20** in an embodiment of the invention. **FIG. 2** depicts the entire KYC system **20**. The KYC system **20** includes a KYC platform **24** in data communication with an off-platform **25** infrastructure **25**. The off-platform **25** infrastructure **25**, as illustrated, includes a laptop computer **22** accessing an example banking website for upload of biographical text **11A** and at least one ID image **11B**. The laptop computer **22** and a plurality of third party data servers **21** are connected to a web server **24A** of the KYC platform **24** through the network **23**. Validation requests **11** include the biographical text **11A** and ID images **11B**. The validation requests **11** are delivered from the customer's laptop computer **22** to the KYC platform **24** via the network **23**.

The KYC platform **24** includes a web server **24A**, a request database **24B**, a facial matching module **24C**, a document type identifier **24D**, an OCR engine **24E**, a post-editor **24F**, and an evaluation module **24G**. The request database **24B** can be a separate database or a portion of another database or another platform module. E.g., the request database **24B** can be a queue or buffer in the web server **24A**, the facial matching module **24C**, the document type

identifier **24D**, the OCR engine **24E**, the post-editor **24F**, and/or the evaluation module **24G** of the KYC platform **24**.

FIG. 3 is a textual flowchart representation of a in an embodiment of the invention. The flowchart **3-00** illustrates the steps **3-01** to **3-04** (see below) for implementing an embodiment of the invention.

3-01 Customer enters biographical text **11A** into a first plurality of data fields via a banking website using a laptop computer **22** and uploads an ID image **11B** of the customer's passport. The banking website sends the customer's biographical text **11A** and ID image **11B** to the KYC platform **24** as a validation request **11** through the network **23**.

3-02 A facial matching module **24C** compares the customer's passport photograph with images in a library of images. Passport photographs matching an image from the library results in either rejection or a second level review of the validation request **11**.

3-03 The document type of the ID image **11B** is determined first and used to select a tailored error correction model **13** associated with that document type to correct raw text **12** from the OCR engine **24E**.

3-04 Corrected text **14** is compared to the biographical text **11A** entered by the customer and possibly also third party data records **15**. If the deviation score exceeds a deviation threshold, the validation request **11** is red flagged for rejection or a second level review.

FIG. 4 is a block diagram representing an identifying step for a document type in an embodiment of the invention. **FIG. 4** illustrates details of the document type identifying step **1-02** of **FIG. 1**. In the document type identifying step **1-02**, ID images **11B** undergo CNN analysis **40**. The CNN analysis **40** does not involve use of any biographical text **11A** of the validation request **11**. A single CNN model can be used on ID images **11B** of all document types. Alternatively, multiple CNN models can be used by the KYC method **10**, such as

where each individual CNN model is used on just a subset of identification documents types, such as just the identification documents from a specific country or region. The CNN analysis **40** can determine a probability score for each document type. As illustrated in **FIG. 4**, the CNN analysis **40** has resulted in a 95% likelihood of the specific ID image **11B** being a driver's license.

The document type identifying step **1-02** depicted in **FIG. 4** is performed prior to the post-editing step **1-04** because the document type conclusion is used to select the appropriate tailored error correction model **13**. E.g., while there may be only one CNN model to determine the document type, there are multiple tailored error correction models **13**, each tailored error correction model **13** specifically created by supervised learning for just one of the document types.

FIG. 5 is a chart **5-00** detailing the accuracy rates of document type identification in an embodiment of the invention. The accuracy rates are determined across a set of sample ID images **11B** including various document types commonly submitted in the Philippines for KYC validation requests **11**. The overall accuracy in identifying the document type using CNN analysis **40** in an embodiment of the invention was over 93%.

FIG. 6 is a chart **6-00** detailing the accuracy rates of post-editing error correction in an embodiment of the invention. The accuracy rates in the chart **6-00** are detailed for a set of sample ID images **11B** including driver's licenses and UMID cards document types. The post-editing step **1-04** leverages the previous document type identifying step **1-02**. With the knowledge of the document type, the KYC method **10** and KYC platform **24** are enabled to select the appropriate tailored error correction model **13** to correct the raw text **12**. The overall accuracy in error correcting the raw text **12** using an embodiment of the invention was over 94%. Note that the less standardized data from the ID images **11B** (such as the customer's address) has a lower accuracy rating after error correction than the more standardized fields.

A general embodiment of the invention is a KYC method **10** or KYC platform **24** for receiving and processing a plurality of validation requests **11** which include ID images **11B**. Validation requests **11** with fraudulent ID images **11B** are red flagged using facial matching.

Document types for each ID image **11B** are identified. ID images **11B** undergo OCR processing without the use of text regions to produce raw text **12**. Raw text **12** from the ID images **11B** is post-edited to create corrected text **14**. Post-editing includes use of tailored error correction models **13** associated with the document type of the ID image **11B**.

5

A first embodiment of the invention is a computer-implemented KYC method **10**, the method comprising: (a) receiving at a request database **24B** in data communication with a network **23** a plurality of validation requests **11** through a network **23**; (b) storing each validation request **11** in one of a plurality of request records of the request database **24B**; (c) facial
10 matching to compare each photograph to each of a plurality of library images to determine a duplication score for the validation request **11** associated with the photograph; (d) identifying a document type of each ID image **11B** (e) conducting OCR to convert the one or more ID images **11B** of each validation request **11** to raw text **12**; (f) post-editing the raw text **12** to execute sequence-to-sequence error correction upon the raw text **12** of each ID image **11B** of
15 each validation request **11** to create corrected text **14** for each ID image **11B** of the validation request **11**; and (g) evaluating each validation request **11** to compare the biographical text **11A** delineated into the first plurality of data fields of the validation request **11** against the corrected text **14** delineated into the second plurality of data fields of the validation request **11** to determine a deviation score for the validation request **11**. Each validation request **11**
20 includes biographical text **11A** and one or more ID images **11B** purportedly identifying a customer associated with the validation request **11**. The biographical text **11A** for each validation request **11** is delineated into a first plurality of data fields within the request record for the validation request **11**. At least one of the one or more ID images **11B** of each validation request **11** includes a photograph of the customer. The duplication score of each
25 validation request **11** is evaluated according to a duplication threshold. The facial matching step includes red flagging each validation request **11** exceeding the duplication threshold. The conducting OCR step does not identify a plurality of text regions in each ID image **11B**. The raw text **12** created in the conducting OCR step does not include fonts or formatting. The post-editing step includes accessing a plurality of tailored error correction models **13**, each
30 tailored error correction model **13** created by supervised learning specifically for one of the document types without the use of text regions. For each ID image **11B** of the validation request **11**, the post-editing step includes selecting the tailored error correction model **13** associated with the document type of the ID image **11B**, as identified in the identifying step,

to: (1) execute the sequence-to-sequence error correction upon the raw text **12** of the ID image **11B** of the validation request **11**; and (2) delineate the corrected text **14** of the ID image **11B** into a second plurality of data fields for the validation request **11**. The deviation score of each validation request **11** is evaluated according to a deviation threshold. Each
5 validation request **11** exceeding the deviation threshold is red flagged.

In an alternative of the first embodiment: (i) each validation request **11** red flagged for exceeding the duplication threshold is given a rejection status; and (ii) each rejection status includes a first log file detailing a rejection rationale.

10

In an alternative of the first embodiment: (a) each validation request **11** red flagged, in the evaluating step, for exceeding the deviation threshold is given a second level review recommendation; and (b) each second level review recommendation includes a second log file detailing a second level review rationale.

15

In an alternative of the first embodiment, the evaluating step further includes: (a) linking, through the network **23**, to one or more third party data servers **21**; (b) submitting a data request to at least one of the third party data servers **21** for each validation request **11**, each data request associated with the customer purportedly identified in the validation request **11**;
20 (c) receiving a third party data record **15** for the customer in response to the data request submission, the third party data record **15** including additional biographical data for the customer; and (d) comparing the additional biographical data for the customer against the second plurality of data fields associated with the validation request **11** of the customer. In this embodiment: (a) to identify the customer in the data request, each data request can
25 include a subset of the bibliographic text included in the validation request **11** associated with the data request; and (b) the subset of the bibliographic text can be at least one of a driver's license number, a passport number, a tax identification number, a formal name, and an address.

30 In an alternative of the first embodiment, the post-editing step includes natural language processing with part-of-speech tagging.

In an alternative of the first embodiment: (a) the step of identifying the document type of each ID image **11B** is performed by a convolutional neural network analysis upon the ID image **11B**; and (b) the identifying step includes determining the document type of each ID image **11B** without the use of either the raw text **12** or the corrected text **14** associated with the ID image **11B**.

A second embodiment of the invention is a KYC platform **24** for receiving and processing a plurality of validation requests **11**, the platform comprising: (a) a request database **24B** in data communication with a network **23**; (b) a facial matching module **24C** configured to compare each photograph to each of a plurality of library images to determine a duplication score for the validation request **11** associated with the photograph; (c) a document type identifier **24D** configured to determine a document type of each ID image **11B**; (d) an OCR engine **24E** configured to convert the one or more ID images **11B** of each validation request **11** to raw text **12**; (e) a post-editor **24F** configured to execute sequence-to-sequence error correction upon the raw text **12** of each ID image **11B** of each validation request **11** to create corrected text **14** for each ID image **11B** of the validation request **11**; and (f) an evaluation module **24G** configured, for each validation request **11**, to compare the biographical text **11A** delineated into the first plurality of data fields of the validation request **11** against the corrected text **14** delineated into the second plurality of data fields of the validation request **11** to determine a deviation score for the validation request **11**. The request database **24B** is configured to receive the validation requests **11** through the network **23** and store each validation request **11** in one of a plurality of request records of the request database **24B**. Each validation request **11** includes biographical text **11A** and one or more ID images **11B** purportedly identifying a customer associated with the validation request **11**. The biographical text **11A** for each validation request **11** is delineated into a first plurality of data fields within the request record for the validation request **11**. At least one of the one or more ID images **11B** of each validation request **11** includes a photograph of the customer. The library images are stored in the KYC platform **24** or are accessible to the KYC platform **24** through the network **23**. The duplication score of each validation request **11** is evaluated according to a duplication threshold. The facial matching module **24C** red flags each validation request **11** exceeding the duplication threshold. The OCR engine **24E** is not configured to identify a plurality of text regions in each ID image **11B**. The raw text **12** created by the OCR engine **24E** does not include fonts or formatting. The post-editor **24F**

includes a plurality of tailored error correction models **13**, each tailored error correction model **13** created by supervised learning specifically for one of the document types without the use of text regions. For each ID image **11B** of the validation request **11**, the post-editor **24F** selects the tailored error correction model **13** associated with the document type of the ID image **11B**, as identified by the document type identifier **24D**, to: (1) execute the sequence-to-sequence error correction upon the raw text **12** of the ID image **11B** of the validation request **11**; and (2) delineate the corrected text **14** of the ID image **11B** into a second plurality of data fields for the validation request **11**. The deviation score of each validation request **11** is evaluated according to a deviation threshold. Each validation request **11** exceeding the deviation threshold is red flagged.

In an alternative of the second embodiment: (a) each validation request **11** red flagged for exceeding the duplication threshold is given a rejection status; and (b) each rejection status includes a first log file detailing a rejection rationale.

In an alternative of the second embodiment: (a) each validation request **11** red flagged by the evaluation module **24G** for exceeding the deviation threshold is given a second level review recommendation; and (b) each second level review recommendation includes a second log file detailing a second level review rationale.

In an alternative of the second embodiment, the evaluation module **24G** is further configured to: (a) link, through the network **23**, to one or more third party data servers **21**; (b) submit a data request to at least one of the third party data servers **21** for each validation request **11**, each data request associated with the customer purportedly identified in the validation request **11**; (c) receive a third party data record **15** for the customer in response to the data request submission, the third party data record **15** including additional biographical data for the customer; and (d) compare the additional biographical data for the customer against the second plurality of data fields associated with the validation request **11** of the customer. In this embodiment: (a) to identify the customer in the data request, each data request can include a subset of the bibliographic text included in the validation request **11** associated with the data request; and (b) the subset of the bibliographic text can be at least one of a driver's license number, a passport number, a tax identification number, a formal name, and an address.

In an alternative embodiment of the second embodiment, the post-editor **24F** is configured to execute natural language processing with part-of-speech tagging.

- 5 In an alternative of the second embodiment: (a) the document type identifier **24B** employs a convolutional neural network analysis upon the ID image **11B** to determine the document type; and (b) the document type identifier **24D** is configured to determine the document type of each ID image **11B** without the use of either the raw text **12** or the corrected text **14** associated with the ID image **11B**.

10

A third embodiment of the invention is a computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out any one of the methods of the first embodiment and the methods of the alternatives of the first embodiment.

- 15 A fourth embodiment of the invention is a computer-readable storage medium comprising instructions which, when executed by a computer, cause the computer to carry out any one of the methods of the first embodiment and the methods of the alternatives of the first embodiment.

- 20 The KYC method **10** and KYC platform **24** provide a technical solution directed at a cost-effective but accurate validation request **11** review process for large volumes of incoming validation requests **11** where the identification document images have varying readability, format, and content. Assisted by AI / ML techniques, the KYC method **10** and KYC platform **24** perform a first level automated review of each validation request **11**. Validation requests
25 **11** that cannot be validated by the KYC platform **24** can be either rejected outright or routed to a second level review.

- The second level review can be performed by third party software and/or personnel, making the second level review more costly (per validation request **11**) than the first level review. If
30 personnel are involved in the second level review, this human review can also introduce a significant time delay for completion of the review. The KYC method **10** and KYC platform **24** is used to reduce the total number of validation requests **11** that are passed through to the second level review, thereby reducing the overall processing costs for KYC validation.

There are two initial steps performed by the KYC method **10** and KYC platform **24** which are performed prior to the conducting OCR step **1-03**. A first initial step, employed to detect duplicate or fraudulent accounts, is a facial matching step **1-01** comparing the faces pictured on the incoming ID images **11B** with the faces captured in a plurality of library images. The library images can be photographs of existing customers, people in a proprietary database, and/or people in a public database. Matching facial features between the photographs within the ID images **11B** submitted by a potential new customer and an existing customer can, for instance, indicate a duplicate or fraudulent KYC validation request **11**.

10

A second initial step taken by the KYC method **10** and KYC platform **24** is the document type identifying step **1-02**. This second initial step is performed on the ID image **11B**, without the use of raw text **12** or corrected text **14**. This second initial step is vital as it allows the KYC method **10** and KYC platform **24** to select a tailored error correction model **13** for the post-editing sequence-to-sequence error correction performed in the post-editing step **1-04**. Use of the tailored error correction models **13** enables the post-editor **24F** to consider the “context” of the raw text **12** without the need for the complexity and cost resulting from the use of text regions.

The tailored error correction models **13** can be trained with supervision to be familiar with the vocabulary and format of the typical biographical data found within each document type. The post-editing step **1-04** is approached as a “sequence-to-sequence” problem, where the raw text **12** is the source sequence and the corrected text **14** is the target sequence. The tailored error correction models **13** are each created using supervised learning. Each model is limited to a single document type. The training set for each model can include thousands of sample ID images (each including printed text), each sample ID images true text (e.g., the actual text printed in onto the identification card), each sample ID’s raw text **12** outputted from a given OCR engine **24E**, and each sample ID image’s corrected text **14** from a sample post-editor **24F**. In this manner the post-editor **24F** (with its set of tailored error correction models **13**) has an understanding of how the given OCR engine **24E** misreads the printed text of the incoming ID images **11B** and has learned how to correct these mistakes for each document type.

To address readability variance for the printed text of the incoming ID images **11B**, the KYC method **10** and KYC platform **24** have the flexibility to select a robust (and potentially low-cost) OCR engine **24E**. An effective OCR engine **24E** typically exhibits good performance with varying lighting conditions of the ID images **11B** and also has the capability of reading longer texts. Because the KYC method **10** and KYC platform **24** only require raw text **12** of the entire ID image **11B**, it is very simple to swap out and replace earlier OCR engines **24E** with new revisions of the same OCR engine **24E**, OCR engines **24E** licensed by different third parties, or other home-grown OCR engines **24E**. The invention is also flexible enough to accommodate different OCR engines **24E** for use in different countries, for use with different languages, for use in different business application, and for use with different throughput volumes. In short, the invention is directed at optimizing flexibility, efficiency, accuracy, and costs. As a safety valve, to the extent the invention is unable to process specific validation requests **11**, the invention directs these validation requests **11** to second level review. The second level review can be a third party KYC software product, a third party KYC service, or review by personnel.

While various aspects and embodiments have been disclosed herein, it will be apparent that various other modifications and adaptations of the invention will be apparent to the person skilled in the art after reading the foregoing disclosure without departing from the spirit and scope of the invention and it is intended that all such modifications and adaptations come within the scope of the appended claims. The various aspects and embodiments disclosed herein are for purposes of illustration and are not intended to be limiting, with the true scope and spirit of the invention being indicated by the appended claims.

CLAIMS

1. A computer-implemented KYC method, the method comprising:
 - (a) receiving at a request database in data communication with a network a plurality of
5 validation requests through a network;
 - (b) storing each validation request in one of a plurality of request records of the request database;
 - (i) wherein each validation request includes biographical text and one or more ID
images purportedly identifying a customer associated with the validation request;
 - 10 (ii) wherein the biographical text for each validation request is delineated into a first plurality of data fields within the request record for the validation request; and
 - (iii) wherein at least one of the one or more ID images of each validation request includes a photograph of the customer;
 - (c) facial matching to compare each photograph to each of a plurality of library images to
15 determine a duplication score for the validation request associated with the photograph,
 - (i) wherein the duplication score of each validation request is evaluated according to a duplication threshold; and
 - (ii) wherein the facial matching step includes red flagging each validation request exceeding the duplication threshold;
 - 20 (d) identifying a document type of each ID image;
 - (e) conducting OCR to convert the one or more ID images of each validation request to raw text,
 - (i) wherein the conducting OCR step does not identify a plurality of text regions in
each ID image; and
 - 25 (ii) wherein the raw text created in the conducting OCR step does not include fonts or formatting;

- (f) post-editing the raw text to execute sequence-to-sequence error correction upon the raw text of each ID image of each validation request to create corrected text for each ID image of the validation request,
- (i) wherein the post-editing step includes accessing a plurality of tailored error correction models, each tailored error correction model created by supervised learning specifically for one of the document types without the use of text regions; and
- (ii) wherein, for each ID image of the validation request, the post-editing step includes selecting the tailored error correction model associated with the document type of the ID image, as identified in the identifying step, to:
- (1) execute the sequence-to-sequence error correction upon the raw text of the ID image of the validation request; and
- (2) delineate the corrected text of the ID image into a second plurality of data fields for the validation request; and
- (g) evaluating each validation request to compare the biographical text delineated into the first plurality of data fields of the validation request against the corrected text delineated into the second plurality of data fields of the validation request to determine a deviation score for the validation request,
- (i) wherein the deviation score of each validation request is evaluated according to a deviation threshold; and
- (ii) wherein each validation request exceeding the deviation threshold is red flagged.
2. The method of claim 1,
- (a) wherein each validation request red flagged for exceeding the duplication threshold is given a rejection status; and
- (b) wherein each rejection status includes a first log file detailing a rejection rationale.
3. The method of claim 1,
- (a) wherein each validation request red flagged, in the evaluating step, for exceeding the deviation threshold is given a second level review recommendation; and
- (b) wherein each second level review recommendation includes a second log file detailing a second level review rationale.

4. The method of claim 1, wherein the evaluating step further includes:
- (a) linking, through the network, to one or more third party data servers;
 - (b) submitting a data request to at least one of the third party data servers for each validation request, each data request associated with the customer purportedly
5 identified in the validation request;
 - (c) receiving a third party data record for the customer in response to the data request submission, the third party data record including additional biographical data for the customer; and
 - (d) comparing the additional biographical data for the customer against the second plurality
10 of data fields associated with the validation request of the customer.
5. The method of claim 4,
- (a) wherein, to identify the customer in the data request, each data request includes a subset
15 of the bibliographic text included in the validation request associated with the data request; and
 - (b) wherein the subset of the bibliographic text is at least one of a driver's license number, a passport number, a tax identification number, a formal name, and an address.
6. The method of claim 1, wherein the post-editing step includes natural language
20 processing with part-of-speech tagging.
7. The method of claim 1,
- (a) wherein the step of identifying the document type of each ID image is performed by a convolutional neural network analysis upon the ID image; and
 - (b) wherein the identifying step includes determining the document type of each ID image
25 without the use of either the raw text or the corrected text associated with the ID image.
8. A computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out the method of any one of claims 1-7.
30
9. A computer-readable storage medium comprising instructions which, when executed by a computer, cause the computer to carry out the method of any one of claims 1-7.

10. A KYC platform for receiving and processing a plurality of validation requests, the platform comprising:

- (a) a request database in data communication with a network,
 - (i) wherein the request database is configured to receive the validation requests through the network and store each validation request in one of a plurality of request records of the request database;
 - (ii) wherein each validation request includes biographical text and one or more ID images purportedly identifying a customer associated with the validation request;
 - (iii) wherein the biographical text for each validation request is delineated into a first plurality of data fields within the request record for the validation request; and
 - (iv) wherein at least one of the one or more ID images of each validation request includes a photograph of the customer;
- (b) a facial matching module configured to compare each photograph to each of a plurality of library images to determine a duplication score for the validation request associated with the photograph,
 - (i) wherein the library images are stored in the KYC platform or are accessible to the KYC platform through the network;
 - (ii) wherein the duplication score of each validation request is evaluated according to a duplication threshold; and
 - (iii) wherein the facial matching module red flags each validation request exceeding the duplication threshold;
- (c) a document type identifier configured to determine a document type of each ID image;
- (d) an OCR engine configured to convert the one or more ID images of each validation request to raw text,
 - (i) wherein the OCR engine is not configured to identify a plurality of text regions in each ID image; and
 - (ii) wherein the raw text created by the OCR engine does not include fonts or formatting;

- (e) a post-editor configured to execute sequence-to-sequence error correction upon the raw text of each ID image of each validation request to create corrected text for each ID image of the validation request,
- (i) wherein the post-editor includes a plurality of tailored error correction models, each tailored error correction model created by supervised learning specifically for one of the document types without the use of text regions; and
- (ii) wherein, for each ID image of the validation request, the post-editor selects the tailored error correction model associated with the document type of the ID image, as identified by the document type identifier, to:
- (1) execute the sequence-to-sequence error correction upon the raw text of the ID image of the validation request; and
- (2) delineate the corrected text of the ID image into a second plurality of data fields for the validation request; and
- (f) an evaluation module configured, for each validation request, to compare the biographical text delineated into the first plurality of data fields of the validation request against the corrected text delineated into the second plurality of data fields of the validation request to determine a deviation score for the validation request,
- (i) wherein the deviation score of each validation request is evaluated according to a deviation threshold; and
- (ii) wherein each validation request exceeding the deviation threshold is red flagged.
11. The platform of claim 10,
- (a) wherein each validation request red flagged for exceeding the duplication threshold is given a rejection status; and
- (b) wherein each rejection status includes a first log file detailing a rejection rationale.
12. The platform of claim 10,
- (a) wherein each validation request red flagged by the evaluation module for exceeding the deviation threshold is given a second level review recommendation; and
- (b) wherein each second level review recommendation includes a second log file detailing a second level review rationale.

13. The platform of claim 10, wherein the evaluation module is further configured to:
- (a) link, through the network, to one or more third party data servers;
 - (b) submit a data request to at least one of the third party data servers for each validation request, each data request associated with the customer purportedly identified in the validation request;
 - (c) receive a third party data record for the customer in response to the data request submission, the third party data record including additional biographical data for the customer; and
 - (d) compare the additional biographical data for the customer against the second plurality of data fields associated with the validation request of the customer.
14. The platform of claim 13,
- (a) wherein, to identify the customer in the data request, each data request includes a subset of the bibliographic text included in the validation request associated with the data request; and
 - (b) wherein the subset of the bibliographic text is at least one of a driver's license number, a passport number, a tax identification number, a formal name, and an address.
15. The platform of claim 10, wherein the post-editor is configured to execute natural language processing with part-of-speech tagging.
16. The platform of claim 10,
- (a) wherein the document type identifier employs a convolutional neural network analysis upon the ID image to determine the document type; and
 - (b) wherein the document type identifier is configured to determine the document type of each ID image without the use of either the raw text or the corrected text associated with the ID image.