## CLAIMS

What is claimed is:

1. A method for extracting personal data from a sensitive document, the method comprising:

predicting type of document by keypoint matching of the sensitive document;

cropping the sensitive document in response to the type of document predicted to extract document shape and extract one or more fields, the one or more fields comprising text or pictures and the one or more fields comprising at least one categorical value field;

denoising the one or more fields to prepare them for optical character recognition;

performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields; and

categorizing the characters in the at least one categorical value field by calculating similarity scores between predefined category data corresponding to the at least one categorical value field and the detected characters in the at least one categorical value field.

2. The method in accordance with Claim 1 wherein denoising the one or more images comprises utilizing a deterministic denoising approach combined with a Generative Adversarial Network (GAN) based noise modelling to denoise the one or more fields.

3. The method in accordance with Claim 1 or Claim 2, further comprising correcting character spelling mistakes and/or auto-generating missed or invisible characters using a trained Bidirectional Long Short-Term Memory (Bi-LSTM) language model.

4. A method for extracting personal data from a sensitive document, the method comprising:

predicting type of document by keypoint matching of the sensitive document;

cropping the sensitive document in response to the type of document predicted to extract document shape and extract one or more fields, the one or more fields comprising text or pictures;

denoising the one or more fields to prepare them for optical character recognition; and

performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields, wherein predicting the type of document by keypoint matching of the sensitive document comprises:

extracting transformation invariant features corresponding to keypoints of the sensitive document;

using a KAZE-based approach to associate each of the keypoints with a best-describing vector;

building a Bag of Visual Word model having a plurality of clusters by clustering ones of the keypoints sharing same identified properties;

counting the keypoints in each cluster to determine a number of clusters, the number of clusters corresponding to a size of a feature vector combining the vectors associated with the keypoints in the clusters; and

predicting the type of document in response to the size of the feature vector.

5. The method in accordance with Claim 4 wherein using a KAZE-based approach to associate each of the extracted features with a best-describing vector comprises fine-tuning the KAZE-based approach with hyperparameters to associate each of the extracted features with a best-describing vector.

6. The method in accordance with either Claim 4 or Claim 5 wherein counting the keypoints in each cluster to determine a number of clusters comprises:

using a pre-trained K-means model to determine a number of the keypoints in each cluster; and

normalizing the numbers of the keypoints in the clusters to determine the number of clusters.

7. The method in accordance with Claims 4 to 6 wherein predicting the type of document comprises using a pre-trained neural network to predict the type of document in response to the size of the feature vector.

8. A method for extracting personal data from a sensitive document, the method comprising:

predicting type of document by keypoint matching of the sensitive document;

cropping the sensitive document in response to the type of document predicted to extract document shape and extract one or more fields, the one or more fields comprising text or pictures;

denoising the one or more fields to prepare them for optical character recognition; and

performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields, wherein performing optical character recognition (OCR) on the denoised one or more fields comprises utilizing a Tesseract OCR engine modified by providing a plurality of rescaled predicted field images to the Tesseract OCR engine to derive a most effective set of parameters for each of the one or more fields for control of the Tesseract OCR engine when detecting characters in the one or more fields.

9. The method in accordance with Claims 1 to 8 wherein the method for extracting personal data from a sensitive document comprises a lite mode of operation which increases speed in favor of accuracy.

10. A system for extracting personal data from a sensitive document, the system comprising:

a document prediction module for type of document prediction of the sensitive document using a keypoint matching-based approach;

a cropping module to extract document shape and extract one or more fields comprising text or pictures from the sensitive document, wherein the one or more fields comprise at least one categorical value field;

a denoising module to prepare the one or more fields for optical character recognition (OCR);

an OCR module for performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields; and

a categorization module for categorizing the characters in the at least one categorical value field by calculating similarity scores between predefined category data corresponding to the at least one categorical value field and the detected characters in the at least one categorical value field.

11. The system in accordance with Claim 10 wherein the denoising module utilizes a deterministic denoising approach combined with a Generative Adversarial Network (GAN) based noise modelling to denoise the one or more fields.

12. The system in accordance with Claim 10 or Claim 11, further comprising a correction module for correcting character spelling mistakes and/or auto-generating missed or invisible characters using a trained Bidirectional Long Short-Term Memory (Bi-LSTM) language model.

13. A system for extracting personal data from a sensitive document, the system comprising:

a document prediction module for type of document prediction of the sensitive document using a keypoint matching-based approach;

a cropping module to extract document shape and extract one or more fields comprising text or pictures from the sensitive document;

a denoising module to prepare the one or more fields for optical character recognition (OCR); and

an OCR module for performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields,

wherein the document prediction module comprises:

a feature extraction device for extracting transformation invariant features corresponding to keypoints of the sensitive document;

a keypoint-vector association device which uses a KAZE-based approach to associate each of the keypoints with a best-describing vector;

a Bag of Words module for building a Bag of Visual Word model having a plurality of clusters by clustering ones of the keypoints sharing same identified properties;

a keypoint in cluster counting module for counting the keypoints in each cluster to determine a number of clusters, the number of clusters corresponding to a size of a feature vector combining the vectors associated with the keypoints in the clusters; and

a prediction module for predicting the type of document in response to the size of the feature vector.

14. The system in accordance with Claim 13 wherein the keypoint-vector association device uses a fine-tuned KAZE-based approach with hyperparameters to associate each of the extracted features with a best-describing vector.

15. The system in accordance with either Claim 13 or Claim 14 wherein the keypoint in cluster counting module comprises:

a determiner using a pre-trained K-means model to determine a number of the keypoints in each cluster; and

a normalizer to normalize the numbers of the keypoints in the clusters to determine the number of clusters.

16. The system in accordance with Claims 13 to 15 wherein the document prediction module comprises a pre-trained neural network to predict the type of document in response to the size of the feature vector.

17. A system for extracting personal data from a sensitive document, the system comprising:

a document prediction module for type of document prediction of the sensitive document using a keypoint matching-based approach;

a cropping module to extract document shape and extract one or more fields comprising text or pictures from the sensitive document;

a denoising module to prepare the one or more fields for optical character recognition (OCR); and

an OCR module for performing optical character recognition on the denoised one or more fields to detect characters in the one or more fields, wherein the OCR module comprises a Tesseract OCR engine modified by providing a plurality of rescaled predicted field images to the Tesseract OCR engine to derive a most effective set of parameters for each of the one or more fields for control of the Tesseract OCR engine when detecting characters in the one or more fields.

18. The system in accordance with Claims 10 to 17 wherein the system for extracting personal data from a sensitive document comprises a lite mode of operation which increases speed in favor of accuracy.