

CLAIMS

1. Computer-implemented method for improved table processing, comprising the steps of:
 - 5 - providing a fixed-layout electronic document comprising a grid-less table;
 - training a deep learning model (DLM) on training data relating to a plurality of training-related electronic documents, each comprising a grid-less training table and a label associated thereto indicative of a bounding box of the training table;
 - 10 - determining a bounding box for the grid-less table in said fixed-layout electronic document by means of the trained DLM;
 - extracting an image of the grid-less table from said fixed-layout electronic document based on the determined bounding box; and
 - processing the extracted image by at least performing a dilation step;
 - 15 **characterized in that** the method further comprises the steps of:
 - training a conditional generative adversarial network (cGAN) comprising a generator neural network (GNN) and a discriminator neural network (DNN) on training data comprising a set of real pairs, each comprising a training-related dilated grid-less table image and corresponding table grid image, wherein training of the cGAN comprises a plurality of training steps, each comprising one of steps a) or b), and step c):
 - 20 a) providing, to the DNN, a real pair obtained from the set of real pairs;
 - b) providing, to the DNN, a generated pair comprising a candidate fake grid image and a corresponding training-related dilated grid-less table image obtained from the set of real pairs, wherein the candidate fake grid image is obtained by modifying, by the GNN, the corresponding training-related dilated grid-less table image using a random data vector;
 - 25 c) determining, by the DNN, an indication of whether the to the DNN provided real or generated pair is a by the GNN generated pair;
 - wherein training of the cGAN comprises a plurality of GNN and DNN iterative training cycles, each comprising at least one of said plurality of training steps, wherein during each training cycle a respective GNN or
 - 30 DNN associated loss function is minimized until an associated stability cycle predicate is met and wherein the cGAN is trained by minimizing a
 - 35

- combined loss of the GNN and DNN associated loss functions until a combined loss stability predicate is met;
- determining a table grid image for the processed extracted image by means of the trained GNN comprised by the trained cGAN; and
 - 5 - combining the determined table grid image and the grid-less table from the provided fixed-layout electronic document in a gridded image of the grid-less table.
2. Method according to preceding claim 1, wherein processing the extracted image further comprises, prior to said dilation step, the steps of:
- 10 o converting the extracted image into a greyscale image; and
 - o applying thresholding to the greyscale image by means of an adaptive Gaussian technique;
- wherein said dilation step is performed on said thresholded greyscale image.
3. Method according to any of preceding claims 1 or 2, wherein processing the extracted image further comprises, after said dilation step, the steps of:
- 15 o obtaining outlines of dilated objects in the by said dilation step dilated image;
 - 20 o dilating the outline comprising image; and
 - o applying thresholding to the dilated outline comprising image by means of an adaptive Gaussian technique.
4. Method according to preceding claims 2 and 3.
5. Method according to any of preceding claims 1 to 4, wherein processing the extracted image comprises, after said dilation step, an erosion step.
6. Method according to any of preceding claims 1 to 5, wherein the method further comprises the steps of:
- 30 - training an artificial neural network (ANN) on training data comprising training pairs, each comprising a training-related table grid image and corresponding corrupted table grid image, preferably wherein the ANN is an auto-encoder; and

- determining a restored table grid image from the corrupted table grid image by means of the trained ANN comprised trained GNN determined table grid image;
wherein the determined restored table grid image and the grid-less table from the provided fixed-layout electronic document are combined in the gridded image of the grid-less table.
- 5
7. Method according to any of preceding claims 1 to 6, wherein the set of real pairs, each comprising a training-related dilated grid-less table image and corresponding table grid image, are obtained by:
- 10
- o providing a plurality of PDF documents, each comprising a grid-less table;
 - o converting the plurality of PDF documents to HTML documents;
 - o identifying, in each of the HTML documents, a table tag associated to the grid-less table of the associated PDF document;
 - 15 o determining, for each of the grid-less tables, a table grid image based at least in part on the corresponding identified table tag;
 - o obtaining, from each of said PDF documents, an image of the grid-less table; and
 - 20 o dilating each of the obtained images of the grid-less tables.
8. Method according to preceding claim 7, wherein the image of the grid-less table is obtained, from each of said provided PDF documents, by:
- 25
- o determining a bounding box for the grid-less table in each PDF document by means of the trained DLM; and
 - o extracting, from each of said PDF documents, an image of the grid-less table based on the determined corresponding bounding box.
9. Method according to any of preceding claims 1 to 8, wherein the plurality of training-related electronic documents, each comprising a grid-less training table and a label associated thereto indicative of a bounding box of the training table, are obtained by:
- 30
- o providing a plurality of PDF documents, each comprising a grid-less table;
 - 35 o converting the plurality of PDF documents to HTML documents;

- identifying, in each of the HTML documents, a table tag associated to the grid-less table of the associated PDF document; and
 - associating a label, to each of the plurality of PDF documents, for the bounding box of the grid-less table, wherein said label is based at least in part on the corresponding identified table tag.
- 5
10. Method according to any of preceding claims 1 to 9, wherein the DLM is one or more of an artificial neural network (ANN), a recurrent neural network (RNN) or a convolutional neural network (CNN).
- 10
11. Method according to any of preceding claims 1 to 10, wherein the DLM is a CNN.
- 15
12. Computer system for improved table processing, the computer system configured for performing the computer-implemented method according to any of preceding claims 1 to 11.
- 20
13. Computer program product for improved table processing, the computer program product comprising instructions which, when the computer program product is executed by a computer, cause the computer to carry out the computer-implemented method according to any of preceding claims 1 to 11.
- 25
14. Use of the computer-implemented method according to any of preceding claims 1 to 11, the computer system according to preceding claim 12, or the computer program product according to preceding claim 13, for generating a parsed document from the fixed-layout electronic document.
- 30
15. Use according to preceding claim 14, wherein generating a parsed document from the fixed-layout electronic document comprises the step of performing optical character recognition (OCR) of text associated with tabular data from the gridded image of the grid-less table.