## CLAIMS

1. A method for correcting distortion in an image of a page, the method comprising:

identifying a set of undistorted words comprising at least one undistorted word in an undistorted region of one or more images of one or more pages having content related to the content of the page;

identifying at least one distorted word in the image of the page, each distorted word of said at least one distorted word corresponding to an undistorted word from the set of undistorted words;

generating a global transformation function for application to the image of the page so as to substantially transform a distorted word of said at least one distorted word to its corresponding undistorted word; and

applying the global transformation function to pixels of the image of the page.

2. A method as claimed in claim 1, wherein said at least one distorted word comprises at least two distorted words, and wherein generating a global transformation function for application to the image of the page comprises:

generating at least two local transformation functions, each local transformation function of said at least two local transformation functions applicable to a corresponding distorted word of said at least two distorted words so as to substantially transform the corresponding distorted word to its corresponding a undistorted word; and

fitting a global function to said at least two local transformation functions.

3. A method as claimed in claim 2, wherein said at least two local transformation functions comprise polynomial functions.

4. A method as claimed in claim 1, wherein the global transformation function comprises a polynomial function.

5. A method as claimed in claim 1, wherein generating the global transformation function comprises application of an optical flow technique.

6. A method as claimed in claim 1, comprising segmenting the image of the page or said one or more images of one or more pages into words.

7. A method as claimed in claim 1, comprising applying a binarization technique to the image of the page or to said one or more images of one or more pages.

8. A data processing system for correcting distortion in an image of a page, the system comprising:

means for identifying a set of undistorted words comprising at least one undistorted word in an undistorted region of one or more images of one or more pages having content related to the content of the page;

means for identifying at least one distorted word in the image of the page, each distorted word of said at least one distorted word corresponding to an undistorted word from the set of undistorted words;

means for generating a global transformation function for application to the image of the page so as to substantially transform a distorted word of said at least one distorted word to its corresponding undistorted word; and

means for applying the global transformation function to pixels of the image of the page.

9. A data processing system as claimed in claim 8, wherein said at least one distorted word comprises at least two distorted words, and wherein the instructions to generate a global transformation function for application to the image of the page comprises:

means for generating at least two local transformation functions, each local transformation function of said at least two local transformation functions applicable to a corresponding distorted word of said at least two distorted words so as to substantially transform the corresponding distorted word to its corresponding undistorted word; and

means for fitting a global function to said at least two local transformation functions.

10. A data processing system as claimed in claim 9, wherein said at least two local transformation functions comprise polynomial functions.

11. A data processing system as claimed in claim 8, wherein the global transformation function comprises a polynomial function.

12. A data processing system as claimed in claim 8, wherein the instructions for generating the global transformation function comprise instructions to apply an optical flow technique.

13. A data processing system as claimed in claim 8, comprising instructions to segment the image of the page or said one or more images of one or more pages into words.

14. A data processing system as claimed in claim 8, comprising instructions to apply a binarization technique to the image of the page or to said one or more images of one or more pages.

15. A computer program comprising computer program code to, when loaded into a computer system and executed, perform all the steps of the method according to any one of claims 1 to 7.

## CLAIMS

1. A method for correcting distortion in an image of a page, the method comprising:
identifying a set of undistorted words comprising at least one undistorted word in an undistorted region of one or more images of one or more pages having content related to the content of the page wherein identifying at least one undistorted word comprises applying an optical character recognition algorithm for identifying at least one undistorted word having a high confidence value and for each of the at least one undistorted words identified as having a high confidence value generating the set of undistorted words;
identifying at least one distorted word in the image of the page, each distorted word corresponding to an undistorted word from the set of undistorted words, wherein identifying at least one distorted word comprises applying a correction function for comparing the at least one distorted word to an undistorted word in the set of undistorted words and in response to the step of comparing generating a list of candidate words of which the at least one distorted word may be a distortion of;
generating a global transformation function for application to the image of the page so as to transform the at least one distorted word to its corresponding undistorted word by matching the distorted word to each of the undistorted words in the determined list of candidate words to identify the undistorted word; and
applying the global transformation function to pixels of the image of the page, wherein generating a global transformation function for application to the image of the page comprises:
generating a local transformation function, the local transformation function comprising at least two local transformation functions applicable to a corresponding distorted word of said at least two distorted words so as to transform the corresponding distorted word to its corresponding a undistorted word; and
fitting a global function to said at least two local transformation functions.

2. A method as claimed in claim 2, wherein said at least two local transformation functions comprise polynomial functions.

3. A method as claimed in claim 1, wherein the global transformation function comprises a polynomial function.

4. A method as claimed in claim 1, wherein generating the global transformation function comprises application of an optical flow technique.

5. A method as claimed in claim 1, comprising segmenting the image of the page or said one or more images of one or more pages into words.

6. A method as claimed in claim 1, comprising applying a binarization technique to the image of the page or to said one or more images of one or more pages.

7. A data processing system for correcting distortion in an image of a page, the system comprising:
means for identifying a set of undistorted words comprising at least one undistorted word in an undistorted region of one or more images of one or more pages having content related to the content of the page wherein identifying at least one undistorted word comprises applying an optical character recognition algorithm for identifying at least one undistorted word having a high confidence value and for each of the at least one undistorted words identified as having a high confidence value generating the set of undistorted words;
means for identifying at least one distorted word in the image of the page, each distorted word corresponding to an undistorted word from the set of undistorted words, wherein identifying at least one distorted word comprises applying a correction function for comparing the at least one distorted word to an undistorted word in the set of undistorted words and in response to the step of comparing generating a list of candidate words of which the at least one distorted word may be a distortion of;
means for generating a global transformation function for application to the image of the page so as to transform the at least one distorted word to its corresponding undistorted word by matching the distorted word to each of the undistorted words in the determined list of candidate words to identify the undistorted word; and
means for applying the global transformation function to pixels of the image of the page, wherein the means to generate a global transformation function for application to the image of the page comprises:
means for generating at least two local transformation functions, each local transformation function of said at least two local transformation functions applicable to a corresponding distorted word of said at least two distorted words for transforming the corresponding distorted word to its corresponding undistorted word; and
means for fitting a global function to said at least two local transformation functions.

8. A data processing system as claimed in claim 7, wherein said at least two local transformation functions comprise polynomial functions.

9. A data processing system as claimed in claim 8, wherein the global transformation function comprises a polynomial function.

10. A data processing system as claimed in claim 8, wherein the means for generating the global transformation function comprise instructions to apply an optical flow technique.

11. A data processing system as claimed in claim 8, comprising means to segment the image of the page or said one or more images of one or more pages into words.

12. A data processing system as claimed in claim 8, comprising means to apply a binarization technique to the image of the page or to said one or more images of one or more pages.

13. A computer program comprising computer program code to, when loaded into a computer system and executed, perform all the steps of the method according to any one of claims 1 to 7.