(52) International Patent Classification: G06K 9/20, 9/22
(21) International Application Number: PCT/EP2005/052467
(22) International Filing Date: 31 May 2005 (31.05.2005)
(25) Filing Language: English
(26) Publication Language: English
(30) Priority Data: 0413065.4 11 June 2004 (11.06.2004) GB
(72) Inventors; and
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIP (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
Published: with international search report
For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: CAPTURING DATA AND ESTABLISHING DATA CAPTURE AREAS

(57) Abstract: A computer-implemented method of capturing data from a form (200) that has data written on a data entry area (210) of the form (200), the method comprising calculating the area over which extends a data capture area (212) related to said data entry area (210), and capturing data relating to the written data in said data capture area data (212) to establish data capture area data.
CAPTURING DATA AND ESTABLISHING DATA CAPTURE AREAS

Field of the Invention
This invention relates to capturing data and to establishing data capture areas especially but not exclusively to apparatus, methods, and software for capturing data, and for establishing a data capture area associated with a data entry area of a substrate, such as a sheet of paper, the data capture area being the area from which entered data is read by a data capturer.

Background of the Invention
The invention arose from a consideration of a user writing data into data areas of forms using a digital pen of the Anoto™ kind, and the user-written data being allocated to fields associated with the data entry areas for subsequent processing. It will be appreciated that the invention has wider applicability: for example to other digital pen systems where the position of a pen is known to a computer, and to non-pen system such as scanning data capture systems which acquire data by scanning a form. It also applies to non-form related areas.

Data is entered onto forms by a user using a pen (digital or otherwise). (An Anoto-type pen need not be used, for example a form could be completed using a normal pen and then scanned into a computer). The data is entered into specific predetermined data entry areas or fields. A digital version of the manually entered data is known to a computer (for example either via a digital pen or via scanning the form), and application software specific to the field processes the data (e.g. adds it to a database, or evaluates a parameter derived from the data entered in the field). The above scenario overlooks a factor: not all users of a form complete the form perfectly with the lines of characters (e.g. letters or numbers) inside the data entry areas where they are supposed to be: it is not uncommon for handwritten lines defining characters to cross lines defining data entry areas and go outside of the data entry field where they should be.
It is known to have a data capture area associated with a data entry area, with the data capture area extending outside of the area of the data entry area. Thus it is known to have software allocate data acquired from a data capture area to the digital/electronic data entry field with which the data capture area is associated, with some parts of a stroke making up the manually entered characters being outside of the data entry area, but within the data capture area.

In the prior art, when a form is designed by a form designer (a person) they firstly design the layout of the form, its text and associated data entry areas (e.g. boxes), and then they decide for each data entry area how big each associated data capture area is to be and where on the form its boundaries are to be. This may be achieved by, for example, moving a cursor on a computer display screen to delineate data capture areas associated with data entry areas of the form.

The above is described in more detail in relation to Figures 4 to 6. The process is very time consuming since it is required to be done accurately paying attention to avoid having possible overlaps among nearby data capture areas but covering the required area. The process needs to be repeated whenever a change is made to the layout of the data entry areas on the form.

Summary of the Invention

An aspect of the invention provides a computer-implemented method of processing data from a form that has a data entry area, the method comprising calculating automatically a data capture area for said data entry area.
This avoids human error in determining the data capture area, and increases the reliability of OCR/ICR data acquired from the data entry areas. It also speeds us the design process of a form.

The term “data entry area” is to be understood to be, in some embodiments, a graphic representation marked out on a form to indicate the place where a user should write data. The data entry area may take the appearance of a box, or of a line on the form, but there are many other ways of indicating the data entry area.

The term “data capture area” is to be understood to be an area on the form associated with a data entry area such that data written or otherwise entered in the data capture area will be associated with data derived from the data entry area.

The automatic generation of a data capture area for a data entry area removes the need for a human operator to generate the data capture areas for the form. Not only does this speed up the process it also removes any human error in the generation of the data capture areas. Also, the automatic creation of data capture areas results in an improved accuracy in determining captured data (in OCR/ICR conversion of script-applied data/markings).

The method may comprise capturing data relating to writing applied to said data capture area to establish capture area data. This capturing of the written data may comprise scanning, photography or otherwise imaging the form, or establishing pen or data entry device movements over the form whilst the form was being completed by the user (e.g. by a tablet device, or an Anoto pen-type device).

The data capture area data can be assigned to a digital data entry field in a computer, there being a digital data entry field corresponding to each data
capture area and therefore corresponding to each data entry area. The
digital data entry field is an electronic entity such as, for example, an area
of a computer memory that is identified by a particular memory address.

In an embodiment of the invention the boundaries of the data capture areas
are determined using information relating to the data entry area, (for
example using the size and/or position of the data entry area). In one
example, the boundaries of said data capture area may be calculated using a
dimension of the data entry area.

The boundary of the data capture area may be calculated so that it overlaps,
or surrounds, the data entry area. There may be a margin between the data
entry area and the data capture area. For example at least one of, or both of,
the top and bottom margins may be half of the height of the data entry area.

Whereas at least one of, or both of, the left and right margins may have a
width that is equivalent to two thirds of the height of the data entry area. A
data capture area beyond the data input area that projects with a height that
is half of the height of the data input area, and projects sideways by two-
thirds of the height of the data entry area is particularly suitable when the
written data to be captured is in the form of characters from the Roman
alphabet (a, b, c ...z). Other dimensions may be chosen to suit other
alphabets or to suit Japanese or Chinese characters.

It is to be understood that the terms “top”, “bottom”, “left” and “right” take
their normal meaning with respect to a medium (such as a piece of paper, or
stylus-sensitive screen of a Tablet GUI) when written on by a user.

The data capture area may be positioned asymmetrically around the data
entry area. That is the top margin may be smaller or larger than the bottom
margin and/or the left margin may be smaller or larger than the right
margin. For example the right hand margin may be larger than the left hand
margin to account for the fact the when written data is entered into a data
entry box and is written from left to right then the user is more likely to overrun the right hand boundary of the data entry area than the left hand boundary.

The margin at a place in the boundary of the data entry area may be zero: i.e. the boundary of the data capture area may be coincident with a boundary of the data input area. In some circumstances it may be advantageous to have a data capture area boundary inside a data entry boundary.

According to another aspect of the invention there is provided a computer-implemented method of capturing data from a form that has had data written on a data entry area of the form, the method comprising calculating the area over which extends a data capture area related to said data entry area and capturing data relating to the written data in said data capture area to establish data capture area data.

According to an aspect of the invention there is provided a computer-implemented method of capturing data for analysis comprising calculating the boundaries of a plurality of data capture areas associated with a respective plurality of data entry areas of a form using the local environment of each data entry area relative to adjacent data entry areas to calculate the position of said data capture areas.

For example the data capture areas may be sized so that they do not overlap.

According to another aspect of the invention there is provided a computer-implemented method of acquiring written data from a form for subsequent processing, the form having first and second adjacent data entry areas, the method comprising using computer logic to determine a first data capture area associated with said first data entry area, and a second data capture
6 area associated with said second data entry area, said first and second data capture areas having an overlap area where they overlap, and using computer logic to allocate data acquired from said overlap area to either a first or a second captured data data set, corresponding to data acquired from said first or said second data capture areas respectively for character recognition processing of said captured data data sets, said captured data data set to which data from said overlap area is allocated being determined by said computer logic using information derived from the data from said overlap area.

Thus, data (e.g. a pen stroke, or part of a pen stroke) is allocated to whichever one captured data set is determined to be appropriate. The determination of to which data set captured data from an overlap area is to be allocated is performed by software.

The method may, in a first example, comprise calculating the centre of gravity of a stroke written on the form and assigning the stroke to a data capture area data set in accordance with the position of the centre of gravity of the stroke on the form. As a second example, the method may comprise calculating the position where a stroke written on the form crosses with a stroke previously written on the form and assigning the stroke to the data capture area data set containing the centre of gravity of the previously written stroke. As a third example, the method may comprise calculating the position at which a later stroke written on the form crosses a stroke previously written on the form, and assigning the later stroke to a data capture area data set in accordance with the position of the crossing of the strokes.

According to another aspect of the invention there is provided software which when run on a control processor of a data processor adapted to process data acquired from a data input form, is adapted to allocate input data to a determined one of a plurality of data input fields, said software
being adapted to determine boundaries of data capture areas adjacent data input areas of said form, said boundaries being calculated or determined automatically by said software using the positions of said data input areas, data derived from said data capture areas being allocated by said software to associated said data input fields automatically.

According to another aspect of the invention there is provided a computer-implemented method of processing data from a form that has had data written on a data entry area of the form, the method comprising using computer logic to calculate a data capture area related to said data entry area and capturing data relating to writing in said data capture area and processing captured data from said data capture area.

According to another aspect of the invention there is provided a computer-implemented method of processing data from a form that has had data written on a data entry area of the form, the method comprising using computer logic to calculate a data capture area related to said data entry area, and capturing data relating to writing in said data capture area data, and processing captured data from said data capture area, wherein said calculating comprises analysing the data written on the form.

Preferably, alpha-numeric characters are identified by analysing the data written on the form.

It will be appreciated that the form, of any aspect of the invention, may be a physical sheet, such as a sheet of paper or plastic. The physical sheet may permanently carry the form (e.g. a sheet of paper), or it may not: it may temporarily carry, or display, the form. An example of this is the display/graphical/user interface of a tablet computer, or Personal Digital Organiser: when the screen of such a device displays a form with data input areas this constitutes a form. When a user writes on the screen with a stylus, they are creating written data on the form.
In addition to pen-type/stylus-type inputs, other hand manipulated input devices could include a mouse-type device. It may even be possible to consider a user's finger as a "pen" if they write with their finger as a position (e.g. touch) sensitive input screen.

"Writing on a form" is intended to cover all of the above, and more.

A stamp, such as a rubber stamp may be used to apply data to a form, or a finger print may be applied to a form. Both processes apply data to a form (i.e. fingerprint, possibly for fingerprint analysis, may comprise use-applied markings to a form, which could comprise a checking screen). Therefore, the mark made by the stamp or a finger are also to be considered to fall within the scope of the term "written data" and the process of applying a stamp or fingerprint are to be considered to fall within the term "writing on a form".

"A form" may have a series of spaces for answers to questions/areas to input data. On the other hand, "a form" could be a single data entry area – for example a signing in box to identify the user of a computer system (possibly to provide access to functionality, or a system, once the identity of the user has been established. Filling in, e.g. by hand, the user ID area on a GUI is filling in a form.

According to another aspect of the invention there is provided a system comprising a processor adapted to process data acquired from a data input form, a data input device for inputting data from a form into the processor, and software which when run on a control processor of the data processor is adapted to allocate input data to a determined one of a plurality of data input fields, said software being adapted to determine boundaries of data capture areas adjacent data input areas of said form, said boundaries being determined automatically by said software using the positions of said data
input areas, data derived from said data capture areas being allocated by said software to associated said data input fields automatically.

The data input device may be chosen from the group: (i) a digital pen; (ii) a scanner; (iii) a PDA and stylus system; (iv) a tablet PC and stylus system; and (v) a touch sensitive screen.

The processor could be part of the data input device such as part of a PDA or tablet PC or the processor could be part of a hand-held device such as a digital pen. If the processor is part of a hand-held device such as a digital pen then this processor could perform OCR/ICR processing and then send this processed data to another off-pen processor for further processing as required.

15 **Brief Description of the Drawings**

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, of which;

20 **Figure 1** schematically illustrates a sheet of prior art Anoto digital paper;

**Figure 2** schematically illustrates an existing Anoto-type digital pen;

**Figure 3** schematically illustrates a prior art form having data entry areas;

25 **Figure 4** schematically illustrates the prior art form of Figure 3 and a data capture area for each of the data entry areas;

**Figure 5** schematically illustrates the prior art form of Figure 3 that has been completed by a user;
Figure 6 is a flow diagram illustrating a prior art method of creating data capture areas;

Figure 7 schematically illustrates a prior art tablet-pen system and a tablet-mouse system;

Figures 8a to 8d schematically illustrates various prior art data entry areas;

Figure 9 schematically illustrates a feature of an embodiment of the invention having a horizontally orientated data entry area with a corresponding data capture area that has been automatically generated according to a dimension of the data entry area;

Figure 10 schematically illustrates a feature of an embodiment of the invention having a data entry area and a corresponding data capture area that is positioned asymmetrically around the data entry area;

Figure 11 schematically illustrates a feature of an embodiment of the invention having a vertically orientated data entry area with a corresponding data capture that has been generated according to a dimension of the data entry area;

Figure 12 schematically illustrates a feature of an embodiment of the invention having a form having data entry areas in a configuration that does not cause the corresponding data capture areas to overlap;

Figure 13 is a flow diagram illustrating various methods, in accordance with embodiments of the invention, of generating data capture areas;

Figure 14 schematically illustrates a feature of an embodiment of the invention having a form having data entry areas in a configuration that causes the corresponding data capture areas to overlap;
Figure 15 schematically illustrates a portion of the form illustrated in Figure 14 but with the data entry areas repositioned so that the corresponding data capture areas do not overlap;

Figure 16 schematically illustrates a portion of the form illustrated in Figure 14 but with the boundaries of two adjacent data capture areas repositioned by the same amount so that the data capture areas do not overlap;

Figure 17 schematically illustrates a portion of the form illustrated in Figure 14 but with the boundaries of two adjacent data capture areas repositioned by the different amounts so that the data capture areas do not overlap;

Figure 18 schematically illustrates a feature of an embodiment of the invention having data capture areas that are generated to match written data entered onto a form;

Figure 19 schematically illustrates a feature of an embodiment of the invention having written data entered into a data entry area and a data capture area shaped in accordance with the written data;

Figure 20 is a flow diagram illustrating a method in accordance with an embodiment of the invention relating to associating strokes to particular data capture areas;

Figure 21 illustrates a feature of an embodiment of the invention in which a stroke having a centre of gravity falls inside two data capture areas;

Figure 22 illustrates a feature of an embodiment of the invention having written data where there is a first stroke assigned to a first data capture area
and a second stroke having a centre of gravity inside a second data capture area but intersecting the first stroke inside the first data capture area;

**Figure 23** illustrates a prior art scanner connected to a computer;

**Figure 24** illustrates a prior art blank form for use with the scanner of Figure 23;

**Figure 25** illustrates a form of Figure 24 showing, in accordance with an embodiment of the invention, automatically generated data capture areas;

**Figure 26** illustrates a first prior art method of associating written strokes to a data entry field;

**Figure 27** illustrates a second prior art method of associating written strokes to a data entry field;

**Figure 28** schematically illustrates a feature of an embodiment of the invention having a written word and boundary lines calculated for that word;

**Figure 29** schematically illustrates a feature of an embodiment of the invention having two lines of text and boundary lines calculated for the lines of text;

**Figure 30** schematically illustrates a feature of an embodiment of the invention having characters written within data entry areas and data set capture areas calculated for the characters; and

**Figure 31** schematically illustrates a feature of an embodiment of the invention having a written word, part of the word falling outside a data capture area but processed with written data falling inside the data capture area.
Detailed Description

It is convenient to discuss the invention by referring to the prior art Anoto digital pen and paper system, but it will be appreciated that the invention is not restricted to use with any proprietary system.

The prior art Anoto system is described on their website www.anotofunctionality.com. However, since the content of websites can change with time it is to be made clear that the prior art admitted is that which was published on their website no later than the day before the priority date of this patent application. It is also appropriate to include in this application itself a brief review of the Anoto system.

Figure 1 shows schematically an A4 sheet 10 of Anoto digital paper. This comprises a part of a very large non-repeating pattern 12 of dots 14. The overall pattern is large enough to cover 60,000,000 square kilometres. The pattern 12 is made from the dots which are printed using infra-red absorbing black ink. The dots 14 are spaced by a nominal spacing of 300μm, but are offset from their nominal position a little way (about 50μm), for example north, south, east or west, from the nominal position.

In WO 01/126032, a 4x4 array of dots is described, and also a 6x6 array of dots, to define a cell. Each cell has its dots at a unique combination of positions in the pattern space so as to locate the cell in the pattern space. The dot pattern of an area of the dot pattern space codes for the position of that area in the overall dot pattern space. The contents of WO 01/126032 are hereby incorporated by reference, with special reference on the dot pattern and the pen.

The sheet 12 has a pale grey appearance due to the dots 14.
Figure 2 schematically shows a digital pen 20 adapted to write human readable ink in non-machine-readable IR transparent ink and to read a position dot pattern in infra-red. The pen 20 has a housing 22, a processor 24 with access to a memory 26, a removable and replaceable ink nib and cartridge unit 28, a pressure sensor 29 adapted to be able to identify when the nib is pressed against a document, an infra-red LED emitter 30 adapted to emit infra-red light, an infra-red sensitive camera 32 (e.g. a CCD or CMOS sensor), a wireless telecommunications transceiver 34, and a removable and replaceable battery 36. The pen 20 also has a visible wavelength warning light 38 (e.g. a red light) positioned so that a user of the pen can see it when they are using the pen, and a vibration unit 40 adapted to vibrate and to cause a user to be able to feel vibrations through the pen. The pen 20 includes a clock 24' adapted to associate a time value with position data acquired by the pen.

Such a pen exists today and is available from Anoto as the Logitech IOTM pen.

The pen, when in use writing on a page/marking a page, sees a 6x6 array of dots 14 and its processor 24 establishes its position in the dot pattern from that image. In use the LED 30 emits infra-red light which is reflected by the page 12 and detected by the camera 32. The dots 14 absorb the infra-red and so are detectable against the generally reflective background. Of course, the ink of the dots might be especially reflective in order to distinguish them (and the paper less reflective), or they may fluoresce at a different wavelength from the radiation that excites them, the fluorescent wavelength being detected. The dots 14 are detectable against the background page.

The processor 24 processes data acquired by the camera 32 and the transceiver 34 communicates processed information from the processor 24 to a remote complementary transceiver (e.g. to a receiver linked to a PC).
Typically that information will include information related to where in the dot pattern the pen is, or has been, and its pattern of movement, and the time at which the tip of the pen was at any particular position: position values are time-stamped.

There are times when the processor 24 cannot determine its position in pattern space (the overall virtual space defined by the very large dot pattern). For example, if the pen is moved too fast over the pattern the processor cannot process the images fast enough. Also the pen may not be able to see where it is in the dot pattern. This can happen if the page 14 is marked or defaced by colorants, or the pattern covered up with something, or the field of view of the pattern is obscured. The user putting their finger in the way is a common reason why the processor fails to recognise the position of the pen. In order to alert the user to the fact that the pen is not able to determine its position properly the processor 24 is adapted to illuminate the light 38 and cause the vibrator 40 to vibrate. The user gets visual and tactile feedback that the camera is not seeing the dot pattern properly/that the pen is unable to determine its position properly.

Referring to Figure 3 a form 200 is shown having several data entry areas 210. Referring to Figure 4, each data entry area 210 has associated with it a data capture area 212 that surrounds the data entry area 212. Referring to Figure 5, written data 214 that falls within the boundary of a data capture area 214 is associated with the digital data entry field associated with that data capture area 214.

Various types of data entry areas are shown in Figures 8a to 8d. Figure 8(a) illustrates a series of one character boxes 150, such a series of boxes 150 is useful for entering data such as name, address, postcode/zip code, date of birth, postcode etc., where the number of characters is either known (e.g. date of birth) or falls within a narrow range (e.g. the lines of an address. Figure 8(b) illustrates a "free form" box 152 designed to allow a user to
write several words in the box 152, such a box 152 may be useful for a user to express comments in his/hers own words. Figure 8(c) illustrates a “comb” style data entry area 154 which is generally used for similar types of data as that for which the arrangement of Figure 8(a) is used. Figure 8(d) illustrates a “baseline” type data entry area 156, this type of data entry area often being used when the written data is to take the form of a signature. Figure 6 shows a prior art method for the design and processing of the form 200 illustrated in Figures 3 to 5.

At step 240 a form designer creates data entry areas 212 for a form, this may be done, for example by manipulating graphical representations of the data entry area 212 on a computer display unit with, say, a mouse. The data entry areas 210 are shown as boxes but the form designer may have chosen other graphical representations according to the type of data to be entered on to the data entry areas 210. Adjacent to each of the data entry areas 210 there is usually text 208 that is pre-printed on the form 200 that indicates to the user the type of information that should be entered into the respective data entry area 210.

At step 242 the form designer creates a data capture area 212 for each data entry area 210. Figure 4 shows the data capture areas 212 associated with the data entry areas 210 for the form illustrated in Figure 3. The data capture areas 212 will not normally be printed on the form but may be displayed on the VDU of a computer so that the form designer can manipulate the data capture areas 212. The form designer will normally define each data capture area 212 so that it surrounds its respective data entry area 210 and does not overlap any other data capture area. That is, a data capture area 212 is larger than its respective data entry area 210 to allow for writing that falls outside the data entry area 210 but within the data capture area 212 to be captured.

This is the end of the known design process, designing a form.
Figure 6 goes on to show the process of how a form is used by a user. At step 244 the form 200 is printed on to a substrate, for example, on to one or more sheets of paper. It is not necessary that the form 200 is printed after the data capture areas 210 have been set, instead the form 200 could be printed after the data entry areas 210 have been set but before the data capture areas are set. Of course, in other embodiments, the form is not printed at all – it may be displayed on a display screen, for example of a PDA, or a Tablet PC.

At step 250 a user fills in the printed (or displayed) form 200 as is illustrated in Figure 5 shows. The user has hand-written data 214 in the data entry areas 210 using a pen, or stylus, or other hand manipulated scribing device. The user may also enter the data on to the form 200 using a typewriter or by feeding the printed form into a printer attached to a computer (data can still be potentially misaligned). For the purposes of this specification, the data entered on to the form by a user will be termed “written data” irrespective of the way in which the user has entered the data, so long as it is capable of being misaligned with the data input area. The term “written data” is also to be understood to include non-textual graphics. Such graphics may include, for example, crosses, ticks, and loops that encircle or partial encircle pre-printed data on the form. In the example illustrated in Figure 5, some of the written data 214 is not within the data entry areas 214 but has fallen within the corresponding data capture areas 214.

At step 252 the completed form is scanned with an optical scanner to obtain a digital version of the completed form and the digital version of the form is then sent to a computer. Software on the computer can determine what handwritten data, or potentially misaligned data applied to the form in some other way, is within what data capture area 212.
At step 254 optical character recognition (OCR) or intelligent character recognition (ICR) software is used to process the captured data.

At step 256 the data resulting from the OCR/ICR processing is associated with the digital data entry field that is assigned to the data capture field.

At step 258 the digital data entry can be processed. For example the data in the digital data entry fields assigned to the “age” data entry area for different completed forms may be added together and divided by the number of completed forms to determine the average age of the users who have filled in the form. The data in the electronic field may be sent to a database, or other memory system. This may be considered “processing”. It may be used to generate metadata, or used to create an invoice, or to order products or services, or to retrieve a record from a computer database, for example. Of course there is a very large number of ways that the data held in digital data fields can be manipulated or used once the characters have been recognised.

Following the process illustrated in Figure 6, there is a question of what happens to data applied to the form that falls outside the data capture areas 212. In one known example, data from a data capture area outside of the data entry area is allocated to a captured data set to be operated upon by OCR/ICR software, using a rule that says any stroke that intersects a data entry area 210 is allocated, in full, to the data set associated with that data entry area, even if it goes outside of the data entry area. Such a system is operated by Anoto, but it does not have an automatically generated data capture area extending beyond the data entry area. Referring to Figure 26 the word “IT” has been written by writing the letters “T” and “T” in first 212x and second 212y respective data entry areas. The cross-bar stroke of the letter “T” falls within both the first 212x and second 212y data entry areas hence the stroke will be treated as if it had been written in both of these data entry areas. Therefore the second data entry area 212y will be
assigned the written data of the letter "T" and the first data area 212x will
be assigned the written data "I" together with the crossbar of the letter "T".
OCR/ICR software may have no problem interpreting the letter "T"
however it may encounter problems interpreting the letter "I" because of
the extraneous written data that has been captured with it.

In another known example, data falling outside the data entry areas 212
will not be processed by the OCR/ICR software (or any other software
designed for operating on the captured data). A data set, for processing by
OCR/ICR software, is created and in one known system data outside of the
data entry area is ignored when constructing the data set associated with the
data entry area. Therefore the OCR/ICR software uses the "clipped data"
to determine what character has been written. This is not as reliable as
using more complete data. For example, the written data word "Harry"
may be interpreted as "Harru" by the OCR/ICR software if the tail of the
written letter "y" is outside the data capture area 212. Of course whole
letters, words or even groups of words may fall outside the data capture
areas 212 and therefore fail to be assigned to a digital data entry field.

Referring to Figure 27 the word "STOP" has been written. The cross-bar of
the letter "T" in this example spans the first three data entry areas 212. The
full extent of the cross-bar stroke can be assigned to each of the first three
data entry areas 212 in a similar manner to the example of Figure 26
(indicated by the continuous arrows in Figure 27). Alternatively, the cross-
bar stroke may be clipped so that only the part of the stroke that actually
falls within a particular data capture area 212 is assigned to that data
capture area 212. This processing is indicated by dashed arrows in Figure
27. Such processing may help in the OCR/ICR processing since for a stroke
that was not intended for a particular data capture area then there would be
less of that stroke in that data capture area. However such processing may
cause problems, for example the letter "P" in the last data entry area
illustrated in Figure 27 would be clipped in a manner that may cause the OCR/ICR processing to interpret the letter as the letter “F”.

Figure 13 illustrates a method for the design of a form according to an embodiment of the invention, and also a method of processing data obtained using such a form.

At step 338 a form designer creates data entry areas 210 for a form on a computer in a similar way to that done in the prior art method (step 240).

At step 340 software operating on the computer automatically generates a data capture area 212 for each of the data entry areas 210. The data capture areas 212 are generated according to the size and/or distribution of the data entry areas 210 according to various rules. A first example of such a rule uses a dimension of the data entry area 210 to determine the dimensions of the data capture area. The data capture area 212 may be defined so that it is larger than the data entry area 210 by a proportion of a dimension of the data entry area 210.

Referring to Figure 9, the software may assign a data capture area 212 to the data entry area 210 so that the data capture area 212 surrounds the data entry area 210 with margins that are determined from a dimension of the data entry area 210. In the example illustrated the top 202 and the bottom 204 margin each has a size equivalent to half of the height, h, of the data entry area 210 whereas the left 208 and right 206 margins each has a size equivalent to two-thirds of the height of the data entry area 210. These proportions are particularly suitable to capture written data that consists of Roman characters.

Since words consisting of Roman characters are written from left to right on the page a user will normally start writing to the left of the left hand
boundary of a data entry area 210, i.e. within the data entry area 210. However, the user may run over the data entry area 210 so that the written data 214 extends past the right hand boundary of the data entry area 214, i.e. outside the data entry area 214 (as illustrated in Figure 10). Therefore, instead of the data entry area 210 being centred within the data capture area 212 the data capture area 212 can be offset with respect to the data entry area 21. Referring to Figure 10 the right margin 206 may be set to be greater than the left margin 208 – (or the data capture area 212 may even be generated so that there is no left margin 208).

Japanese and Chinese characters are written from top to bottom on a page therefore the data entry areas 210 and the data capture areas 212 may be orientated in the vertical direction. The data capture area 212 may be placed symmetrically around the data entry area 21, as shown in Figure 11, or since the user is more likely to write outside the bottom than outside the top of the data entry area 210 the data capture area 212 may be created so that it provides a larger margin at the bottom of the data entry area 212. Similarly, the rules for generating the data capture areas 212 can be modified to suit the written data of various different alphabets or to suit written data that consists of numeric data (e.g., when the data entry area 210 is for a telephone number).

The rules used for calculating the dimensions of a data capture area 212 may be calculated according to the type of data entry area 210. If the data capture area 212 is a baseline 156 then, for example, the data capture area 212 can be defined as a box that has length that is equal to the length of the baseline 156 plus a further length that is determined as a proportion of the length of the baseline, alternatively the further length may be a set length, for example the data capture area 212 may have a length such that it surrounds the baseline 156 with a 0.5 cm margin to the left of the baseline and a 2 cm margin to the right of the baseline 156. The height of the data capture area 212 for a baseline, for example, will generally be set so that
the majority of the data capture area 212 is above the baseline 156 whilst allowing a margin beneath the line to capture letters such as “f”, “g”, “j”, “p”, “q” and “y” that may be written so that part of the tail of the letter will be below the baseline 156. If the data entry area 210 is a free form box 152 proportioned for the entry of several lines of written data 214 then it may, or may not, be appropriate to generate a data capture area 212 that has margins that are a percentage of a dimension of the data entry area 210. For example, if a data capture area 212 was generated to have a margin that was \( \frac{1}{2} \) the height of the free form box 152 then the resulting data capture area 212 may be excessively large. In this case it may be more appropriate for the margins to be a set dimension, e.g. 1 cm (or different sizes at different edges), or for one of the margins to be a set dimension and the other margin to determined to be proportion of the set dimension according to the ratio of the sides of the free form box 152. For example if the ratio of the height to the width of the free form box 152 is 3:2 then the ratio of the top/bottom margins to the side margins can also be set to be 3:2 (same ratio for margins as size of box).

Referring to Figure 13, the generation of data capture areas 212 can be achieved according to processes that follow particular rules. At step 340 the data capture areas 212 are generated according to a dimension of the respective data entry areas 210 as has been described above. At step 342 it is determined whether any of the data capture areas 212 overlap.

Figure 12 illustrates a number of data entry areas 210 on a form and the associated data capture areas 212, the data capture areas 212 having been generated according to a rule that is based on a dimension of the corresponding data entry area 210. None of the data capture areas 212 overlap so there is no ambiguity as to what written data 214 should be assigned to which data capture area 212. Therefore the processing proceeds to step 348 and the data capture areas 212 are assigned to their respective data entry areas 210.
Figure 14 illustrates a form in which relative positions of the data entry areas 210 would cause the data capture areas 212 to overlap following the processing of step 340 – i.e. generating the data capture areas 212 based solely on a dimension of the respective data entry areas 210. The overlapping data capture areas 210 could cause ambiguity as to which written data 214 should be assigned to which digital data entry field unless further processing is performed. Following steps 344 and 346, one solution is to move the data entry areas 210 so that they are far enough apart so that the data capture areas 212 no longer overlap. Figure 15 illustrates part of form of Figure 14 that has undergone this further processing. The form designer may not want the layout/appearance of the form to be altered significantly by the movement of the data entry areas 210 or there may not be enough space on the hardcopy of the form to allow the required movement, therefore the rule that causes the data entry areas 210 to move may include a limit as to the extent that the data entry areas 210 may be moved. However, in many cases only a small shift of the data entry areas 210 may be necessary in order to remove overlaps between the data capture areas 212.

A different solution, following step 350, is to reduce the size of the data capture areas 212 so that they do not overlap. This may be done by moving the boundaries of capture data areas 212 that lie between the adjacent data entry areas 210 so that the boundaries of capture data areas 212 are equidistant from the boundaries of their respective data entry areas 210. As a modification of this technique, and as is illustrated in Figure 16, the boundaries may be shifted the same distance to prevent the boundaries overlapping – this may or may not result with the boundaries of capture data areas 212 being equidistant from the boundaries of their respective data entry areas 210. Alternatively, as illustrated in Figure 17 and following step 354, the boundary of one of the data capture areas 212 may be moved more than the boundary of the adjacent data capture area 212. In
this case the right hand boundary of the left hand data capture area 212 is moved to the left to an extent that is less than the right hand boundary of the left hand data capture area 212 is moved to the right (the rule may even be set so that the right hand boundary of the right hand data capture area 212 is not be moved at all). Such a rearrangement is advantageous if the user is more likely to write over the right hand boundary of the data entry area 210. Once the data capture areas 212 have been calculated, either following step 352 or step 354, then the data capture areas can be assigned to their respective data entry areas 210 (step 348).

The processing for moving the data entry areas 210 (step 344)/calculating the position of the data entry areas and the processing for calculating the size and position of the data capture areas 212 are not mutually exclusive. For example the data entry areas 210 could be moved to a limited extent and if the data capture areas 212 still overlap then the size of the data capture areas 212 could be reduced until the overlap is removed. Alternatively, the data capture areas could be reduced in size until a minimum size is reached, according to a minimum required margin around the respective data entry areas 210, and if the overlap remains then the data entry areas can be moved until the overlap is removed. Which option is taken can be pre-set by the form designer or the processing software according to one option can be followed and the result can be reviewed by the form designer. If the result does not meet the designer’s approval then the form can be reprocessed according to the other option, or in some other way.

If the written data was handwritten using a digital pen or via a stylus on a PDA or Tablet device, then the digital pen or device can be operated to transmit the written data to a processor together with the position of the writing on the form as the writing is being written.
If the handwritten data 214 is entered using a conventional pen, or a typewriter, or printed on the form, then an optical scanner may be used to read the form so that the "handwritten" data can be obtained in digital form and its position relative to the data capture area 212 can be determined by software running on a computer. Once the data capture areas 212 have been assigned then written, or applied, data that has been written or applied to the form inside the data capture areas 212 can be processed, e.g., by OCR/ICR. The OCR/ICR processed data can then being assigned to the digital data entry fields associated with the respective data capture areas 212.

In an embodiment of the invention, following step 356, the data capture areas 212 are assigned to the data entry areas 210 irrespective of whether the areas 212 overlap. In this case the processing follows on to the method illustrated in Figure 20 and will be described later.

The data capture areas 212 may have their position and size determined (e.g. they may be resized, reshaped, or moved) pursuant to an identification of the user who will complete the form. In this case, once the software has identified the user then the data capture areas 212 are generated so that they are customised to the user. The data capture areas 212 will be optimised according to the style of the user, for example as determined from a history of previously completed forms. For example, a certain user may only rarely (or never have) overrun the left hand margin of a data entry area 210 but often enter written data that is beneath a data entry area 210. In this case the data capture areas 212 can be generated to surround their respective data entry areas 210 so that there is little or no left hand margin but with a bottom margin that is greater than the top margin. As an option, in this example, the data entry areas 210 may be rearranged so that there is less of a gap between horizontally adjacent data entry areas 210 but a larger gap between vertically adjacent data entry areas 210. In this way the space on the form 200 may be optimised whilst providing data capture areas 212 of a
sufficient shape and size to enable the written data to be effectively captured. User-specific, algorithm or heuristic, data capture area determination, bespoke to an identified user, may be helpful, especially in the case of users whose writing is more stylised than usual.

When the identity of the intended user is known then the identity can be entered into the computer, or selected from a list displayed on the computer’s display, so that personalised data capture areas 212 can be assigned before the user fills out the form 200. Alternatively, the identity of the user may be known from the data input device (e.g. an Anoto-type pen, or a PDA or Tablet PC will typically have an owner -- an identified user). Alternatively, the identity of the user may be determined from the written data that has been entered on to the form 200, for example when the user fills out a data entry area 210 associated with a “name” digital data entry field, or by the user writing out a unique identifying code (such as an employee number), or from a signature written on the form. The identification of the user can be made after the form has been completed and all the written data 214 on the completed form has been read into the computer, e.g. by OCR/ICR. The data capture areas 212 can then be generated so that the written data 214 can be assigned to the correct digital data entry field. Thus the calculation of the size and position of the data capture areas for a particular form-filling event may be determined after (or of course before) the form-filling event itself.

If the written data is entered onto the form using a digital pen or other hand-pointed device (stylus, mouse etc.) that can send the written data, and its position on the form, back to the computer then it is possible to identify the user before the form is completed, e.g. as soon as the name data entry area has been filled in by the user (the name data entry area is usually the first, or one of the first, data entry areas to be filled in by a user). Therefore, the data capture areas 212 can also be assigned before the form is completed.
Additionally or alternatively, the software can also adapt the data capture areas 212 according to a history of how a particular form has been previously filled in by users without having any information identifying the next particular user of that form. Heuristic development of data capture areas for a specific form can be helpful.

Alternatively or additionally, the software may determine the size and position of the data capture areas for each form-filling event “on-the-fly”, individually during or after each form-filling event, using knowledge of how the form has actually been filled in. In this method the data capture areas 212 are generated according to an analysis of the form of the handwritten, or applied, data 214.

Referring to Figure 18 the word “Blue” has been written in respective data entry areas 210 by two different users – Sarah and Bill. The written data entered by Sarah has letters that are elongated in the vertical direction whereas the written data entered by Bill has letters that are elongated in the horizontal direction. The software can be configured to analyse the characters entered onto a form and assign the data capture areas 212 with a shape and size that will most efficiently capture the characters. The analysis of a character may involve calculating one or more parameters for the character such as the shape and/or size of the character or the degree to which the character is written outside a particular boundary of a data capture area. The characters can be analysed after the form has been filled in for example when the form is processed, or alternatively after data collection (e.g. via a pen or via a scanner), but before OCR/ICR processed data is processed further, or is subsequently additionally processed. Alternatively, if a digital pen is used to enter the handwritten data 214 (or other hand pointed device is used where the written data, and its position on the form, is known to the computer) then it is possible for the software to calculate a running average of the character parameter(s) as the form is
being filled out, for example the average is updated every time a character
is written on the form. In this way it is possible to assign/refine the data
capture areas 212 as the form is being filled in.

Referring to Figure 19, a user has filled in a data entry area 210 with
written data 214 that has characters that slope forward. The software can be
configured to recognise the slope and adapt the shape of the data capture
area 212 to match the characters. For example the data capture areas 212
can be defined by forward sloping parallelograms. Similarly for writing
that slopes backwards the data capture areas 212 can be defined by
backward sloping parallelograms.

The written data 214 can also be analysed according to the pen strokes that
constitute the written characters. Figure 20 illustrates a method of
determining how a particular pen stroke can be assigned to a data capture
area 212. At step 380 a uses enters written data on to the form. If the
written data 214 is read from the form using a scanner then a stroke can be
defined as a continuous line of ink, or an algorithm may notionally divide
cursive, joined up, writing into separate characters. If the written data 214
is written using a digital pen or other hand pointed device (stylus, mouse
etc.) that can send the written data, and its position on the form, back to the
computer processor, then a stroke can be defined by the data sent to the
computer processor when the pressure sensor of the pen is activated. For
example a stroke may constitute the movement of a pen across the printed
form without removing pressure from the nib of the pen (Anoto-type pens
have a pressure sensor).

At step 382 the 'centre of gravity' of the stroke is determined. For the
purposes of this specification the centre of gravity of the stroke is the
centroid of the stroke, the centroid being a point in a set of points the
coordinates of the centroid being the mean value of the coordinates of the other
points in the set. Therefore the centroid can be easily calculated by the computer,
for example by pixelating the stroke and finding the average position of the pixels. Of course, alternatively a stroke could be treated as a line, and the center point of the line could be determined.

At step 384 the software determines whether the center of gravity of the stoke falls within a data capture area 212. If this is the case then, at step 386, the software determines if the center of gravity falls within a single data capture area. If the center of gravity does fall within a single capture area then, at step 388, then the stroke is assigned to that data capture area 212. If at step 386 it is determined that the center of gravity falls within more than one data capture area 212 then, at step 392, the stroke is assigned to the data capture area 212 that has a center of gravity that is closest to the center of gravity of the stroke. This situation is illustrated in Figure 21 where the stroke that forms the character "O" has been assigned to the data capture area 212a on the left hand of the character.

If at step 384 it is determined that the center of gravity of the stroke does not fall within a data capture area 212 then, at step 390 the stroke is assigned to the nearest data capture area. The nearest data capture area 212 can be defined either as the data capture area 212 that has a boundary closest to the center of gravity of the stroke or the data capture area 212 that has a center of gravity that is closest to the center of gravity of the stroke.

The processing defined in the flow diagram of Figure 20 can be used to process data on a form that has data capture areas that do or do not overlap. That is the flow diagram of Figure 21 can follow on from either point "A" or point "B" on the flow diagram of Figure 13.

If a stroke begins in a particular data capture area 212 but extends outside of it then a rule may be applied so as to assign the stroke to that data capture area 212.
When the letter "t" is handwrittren a user may write the cross stroke of the letter in a way that causes the majority of the stroke to fall outside the data capture area 212c that contains the rest of the letter (similar concepts apply to other letters or numbers having cross strokes). Referring to Figure 22 the word "tomato" has been written so that most of the word falls within a first data capture area 212c and the word "plug" has been written in a second data capture area 212d. The word "tomato" has been written with a number of distinct strokes, that is strokes that form the first "t", the letters "om", the letters "ato" and the cross stokes for the first and second "t"s. These strokes may be readily assigned to the first data capture area 212c using one or more of the rules already described. The cross stroke 400 of the second "t" may be assigned to the first data capture area 212 by using a rule that recognises that the cross stroke 400 intersects the stroke "ato". The centre of gravity of the cross stroke 400 is outside the correct data capture area 212c and closer to the second data capture area 212d however a rule may be defined so as to use the position of the crossing point of the cross stroke 400 and the "ato" stroke. If the crossing point is within the first data capture area 212c the cross stroke 400 is assigned to the first data capture 212c area. In contrast the tail of the letter "g" of the word "plug", which is written in the second data capture area 212d, intersects the stroke "ato" stroke outside the first data capture area 212c and will therefore not be assigned to this area.

The various rules described may be applied according to a set hierarchy of rules. For example, if the centre of gravity of a stroke falls within a single capture area 212 then it is assigned to that data capture area 212. If it does not fall with a single capture area but the stroke starts in a single capture area 212 then it is assigned to that data capture area 212. If the stroke does not start in a data capture area 212 then stroke is assigned to the data capture area that has a centre of gravity that is closest to the centre of gravity of the stroke. It will be appreciated that the rules can be arranged in many different ways to form different hierarchies. The hierarchy that is
applied to a particular form may be selected to match the identity of the user of the form or if the identity of the user is not known the hierarchy can be adapted to suit the written data actually entered on to the form, or the hierarchy may be chosen by the form designer for a particular form identity or layout (i.e. different forms may have different rules associated with them).

For cursive writing ("joined-up writing") a word may consist of a single stroke, or at least fewer strokes than there are characters. For non-cursive writing each character may correspond to a single stroke, or two strokes, or a few strokes. It is possible to assign time data to transmission of data produced from a pen during a stroke. The software can be configured to determine the chronological sequence of strokes, and additionally it may time the period from when one stroke has finished to when another stroke has started. Generally when a user writes out a group of words into a data entry area 210 one word will be shortly followed by another. Similarly in any particular non-cursively written word a stroke forming one letter will be followed by a stroke forming another letter in the word until the complete word is written. Therefore, if a stroke or a portion of a stroke falls outside a specific data capture area 212 then if the stroke or portion of the stroke is written with a set time period of a previously written stroke that falls with the data capture area 212 then the stroke or a portion of a stroke falls outside a specific data capture area 212 may also be also be assigned to that data capture area 212. The same ideas could apply to different strokes within the same letter.

The invention may be realised using an Anoto digital pen and paper system (as has been described with reference to Figures 1 and 2) or in many other ways, for example by scanning the form with an optical scanner, or using touch-position sensitive screens (e.g. of Tablet devices or of PDA’s. Other prior art systems may be used that determine the position of writing that is written on the form and the invention is applicable to them. Referring to
Figure 8(a), one such system uses a tablet 400 and a pen 410 that is attached to the tablet 400. The position of the tip of the pen (or stylus) 410 with respect to the tablet 400 can be fed back to the computer. If a form 200 is placed in a known position on to the tablet 400 the position of the pen on the form 200 can also be determined. Instead of a pen a different type of hand pointed device may be used such as a mouse 420, as is illustrated in Figure 7(b). The mouse 420 or stylus/pen 410 may be physically connected by a wire to either a computer or to a processor in the tablet 200 so that relative moment of the pen/mouse can be fed back to the computer/processor. Alternatively, the pen/mouse may communicate with computer/processor using wireless communication for example using Bluetooth™ technology.

In another system the handheld device has a transmitter to transmit to at least two receivers that are positioned on the tablet 200. In this way the position of the handheld device can be calculated using triangulation.

A touch sensitive screen may be used to determine the position of the pen/stylus/user’s finger, or a position (of stylus or pen) sensing screen may be used.

As has been mentioned, for some embodiments of the invention a scanner or camera may be used as a means to input data from a form. Figure 23 shows a scanner 500 that can be used to scan a form 200 into a computer 510. A flatbed scanner 200 has been illustrated but other types of scanner can be used. Figure 24 illustrates a blank form 200 (i.e. an uncompleted form), for use with the scanner 500, that comprises a number of data entry areas 210. Once the blank form 200 has been scanned software on the computer can be configured to recognise the boundaries of the data entry areas 210. The software then automatically generates data capture areas 212 from information relating to the size and position of the data entry areas 212 as has been previously described. Figure 25 illustrates a representation
of the form 201 showing the automatically generated data capture areas 210 along with the respective data entry areas 210. Such a representation 201 may be viewed on a computer screen 512 but generally when the form 200 is printed for use by a user the data capture areas 212 will not be visible.

A completed form may be scanned into the computer 510 so that written data on the form can be assigned to the appropriate data capture areas 212. The written data may be assigned to a particular data capture area 212 according to any of the rules that have been previously described (for example by determining the position of the centre of gravity of a stroke relative to a data capture area 212 or by determining data capture area 212 that a stroke starts in). The captured data can then undergo further processing such as OCR/ICR.

It is possible for the software to be configured so that it is not necessary for the blank form 200 to be scanned into the computer 510. In this case a single scan of the document may be used to generate the data capture areas 212 and to determine what written data is assigned to what data capture area. This technique allows for the data capture areas 212 to be generated or altered in accordance with the actual data that is written on the form 200, for example by determining if the written data overruns the data entry areas 210 in a particular direction and extending the data capture areas 212 accordingly, or using one of the other techniques that have been previously described.

In an embodiment of the invention written data words on a form 200 can be assigned to data entry areas 210 following an analysis of the extent or area that the words occupy space on the form 200. Referring to Figure 28, the word 214 “Guidelines” has been written on a form 200 and is captured by a computer via use of a digital pen, scanner, PC tablet etc. in a manner that has already been discussed for other embodiments of the invention. Software operating on the computer analyses the captured data to determine
upper 600 and lower 601 boundary lines for the written word 214. In the case of the written word 214 "Guidelines" illustrated, the upper boundary line 600 can be determined from the top points of the letters "G" and "L" in the written word 214 and the lower boundary line 601 can be determined from the lowest point of the letter "G" in the written word 214. The position of the upper 600 and/or lower 601 boundary line in relation to data entry areas 210 or data capture areas 212 on the form 200 can then be used to assign the written word 214 to a particular data entry area 210.

Referring to Figure 29, one line of text 604 has been written above another line of text 603. For the sake of clarity the lines of text have been horizontally offset, but imagine that they are one directly above the other. The upper boundary line 600 of the lower line of text 603 is above the lower boundary line 601 of the upper line of text 604. To enable the correct word to be assigned to the correct data entry area 210 each line of text is analysed to determine a baseline 602 for that line of text. The baseline 602 may be defined to be the lowest line that all the letters in a line of text intersect. The position of the baselines 602 in relation to a data entry area 210 or a data capture area 212 can then be used to assign the respective written words to the appropriate digital data entry fields.

According to a further embodiment of the invention data may be assigned to a data set for OCR/ICR processing according to the actual area on the form 200 that the written data 214 occupies. Figure 30 illustrates a collection of single character data entry areas 210 together with a data capture area 212 that surrounds the collection of data entry areas 210. The data capture area 212 has been automatically generated according to the size and/or position of the data entry areas 210. Rather than a single data capture area 212 that surrounds the collection of data entry areas 210 there could, instead, be a data capture area 212 for each single character data entry area 210.
Individual letters have been written in respective single character data entry areas 210 to form written data, however some of the letters intersect more than one data entry area 210. The written data is captured and then analysed so that a data set capture area 211 is generated for each letter in accordance with the actual area on the form that the letter occupies. Only data that falls within a data set capture area 211 is assigned to a data set, the data set then being operated on by OCR/ICR processing (other data, which may be gathered from area outside of the computer-generated area 211, is not passed to the data set that is operated upon by the OCR/ICR software). By generating a data set capture area 211 in this way it is possible to reduce the area of the form that is analysed by the OCR/ICR software. This has the advantage of reducing pressure on the computer’s resources, for example by reducing the workload on the computer’s processor and the usage of the computer’s memory. This can make the processing faster.

The letter “S”, illustrated in Figure 30, extends into a data entry area 210 occupied by another character, i.e. the letter “E”. The data set capture area for the letter “E” is smaller than the data entry area 210 containing that letter. Therefore, a data set capture area 211 can be generated for the letter “S” that covers part of the data entry area for the letter “E”. This enables ICR/OCR processing to be performed on the entire letter “S” without that data set capture area 211 for that letter overlapping with the data set capture area 211 generated for the letter “E”.

The example illustrated in Figures 29 shows a set of individual letters, however this embodiment of the invention is equally applicable to other types of written data e.g., several words, parts of words, non-cursive words, strokes, non-alphanumeric data and the like.

Figure 31 illustrates a data entry area 210, a data capture area 212 that has been automatically calculated for the data entry area 210 using the dimensions of the data area 210 as inputs. Written data in the form of the
word "Jameson" has been written partially within the data capture area 212 and partially outside the data capture area 212. The written data is captured by a computer via use of a digital pen, scanner, PC tablet etc. in a manner that has already been discussed for other embodiments of the invention.

The first part of the word, i.e. the text "JAMES" is written within the data capture area 212. The second part of the word, i.e. the text "ON" falls outside the data capture area 212. However this part of the word has still been captured by the computer. According to an embodiment of the invention, the software (computer logic) can be configured to recognise written data that is in the general vicinity of a data capture area 212 but falling outside that data capture area 212. Written data falling within this vicinity can then undergo OCR/ICR processing and be assigned to the same electronic/digital data entry field as the written data falling within the data capture area 212. In effect an additional data capture area 212' can be calculated for the second part of the word. The extent of the vicinity may in some embodiments be calculated from the size and configuration of the data capture areas 212 on the form 200, and in some embodiments by analysing where the markings are that are near, but outside, the data capture area 212 that has been generated using the dimension of the data input area as input.

In one scenario the letter "O" may fall within the calculated vicinity of the data capture area 212 and therefore be processed along with the rest of the data word that falls within the data capture area 212. The letter "N" may fall outside the calculated vicinity because for example this letter is too close to an adjacent data capture area and therefore the calculated vicinity cannot be extended this far. However the letter "N" may still be processed with the rest of the word if the software is configured to recognise that the letter "N" was written shortly after the letter "O" or if the letter "N" intersects the letter "O".
It will be appreciated that in one group of preferred embodiments a form
design system takes as inputs information relating to a plurality of data
entry areas (or a data entry area) and using a set of rules defines respective
data capture areas associated with each of the data input areas, without the
need for a user to position or size the data capture areas. The dimension of
the data capture area may be a percentage of a dimension of data capture
area. Overlapping of data capture areas may be avoided, or permitted. The
percentage dimension of the data capture area and the amount of overlap
between data capture areas may be calculated in a dynamic way for each
identified user, possibly based on their handwriting style. An analysis of
the handwritten markings on a form may be used to calculate the data
capture area (for example analysis may define a baseline for a word), data
entries, and from the position of the baseline, and perhaps its length, a data
capture area may be derived/determined. In the case of boxed data input
areas, each adjacent box may be considered a separate data input area.

The data entry area may comprise a box, or a line above which data is to be
written/user-applied to the form; some marking on the form to guide the
user as to where they should mark the form (e.g. with writing, or with a
tick, or with something. Alternatively, the data entry area may not be
delineated as such on the form, but rather more “pointed at”. For example,
“write name here:” can constitute a data entry area, or some
written legend and an arrow to indicate where to write an answer.

Embodiments of the invention may reduce/minimise a digital
transformation phase of data processing, reducing the time to enable a form
to be used for OCR/ICR workflows, by reducing the time taken to design
the form, design of data capture areas being guided by the graphical design
of the form.
Other embodiments, or perhaps an embodiment that also includes the above, provide dynamic adaptation of the definition of data capture areas depending upon user handwriting style.

The concepts of (i) using the date area markings on a form to influence the position of the data capture area, and (ii) using the user-applied markings on the form to determine the data capture area (or at least to influence a subset of captured data that is to be passed to a data set for processing), can be thought of as linked in the sense that they both use markings of some kind on the form to provide a guide as to where the data capture area is going to be. The ideas can be used separately or together.

Of course, it will be appreciated that the different aspects of the invention, and different features disclosed herein, may be used together in any number, and in any combination.
CLAIMS

1. A computer-implemented method of determining the area over which a data capture area extends, the data capture area being associated with a data entry area of a form, the method comprising using software to determine the data capture area using at least one of:
   (i) the size or
   (ii) position; or
   (iii) the shape of the data entry areas.

2. The method of claim 1 wherein the boundaries of said data capture area are determined using a dimension of the data entry area.

3. The method of claim 2 wherein the boundaries of the data capture area form top, bottom, left and right margins around the data entry area such that the top and bottom margins are each substantially half the height of the data entry area and the left and right margins are each substantially two-thirds of the height of the data entry area.

4. The method of claim 1 comprising determining the boundaries of a plurality of data capture areas associated with a respective plurality of data entry areas of a form using the local environment of each data entry area relative to adjacent data entry areas to determine the position of said data capture areas.

5. A method according to claim 1 practised on a form that has a first and second adjacent data entry areas, the method comprising using computer logic to determine a first data capture area associated with said first data entry area, and a second data capture area associated with said second data entry area, said first and second data capture areas having an overlap area where they overlap, and
using computer logic to allocate data acquired from said overlap area to either a first or a second captured data set corresponding to data acquired from said first or said second data capture areas respectively for character recognition processing of said captured data sets, said captured data set to which data from said overlap area is allocated being determined by said computer logic using information derived from the data from said overlap area.

6. The method of claim 5 comprising determining the centre of gravity of a stroke written on the form and assigning the stroke to a captured data set in accordance with the position of the centre of gravity of the stroke on the form.

7. The method of claim 5 comprising determining the position where a stroke written on the form crosses with a stroke previously written on the form and assigning the stroke to the captured data set containing the centre of gravity of the previously written stroke.

8. The method of claim 5 comprising determining the crossing position of a later stroke written on the form with a stroke previously written on the form and assigning the later stroke to a captured data set in accordance with the position of the crossing of the strokes.

9. The method of claim 5 comprising determining the captured data set to which data relating to a stroke is allocated using the chronological order of strokes.

10. The method according to any of the above claims, wherein data capture areas for a particular user are generated according to a rule associated with said particular user.

11. The method of claim 10 wherein the rule is based on an analysis of strokes applied by the user when the user has previously completed forms.
12. A system comprising a processor adapted to process data acquired from a data input form, a data input device for inputting data from a form into the processor, and software which when run on a control processor of the data processor is adapted to allocate input data to a determined one of a plurality of data input fields, said software being adapted to determine boundaries of data capture areas adjacent data input areas of said form, said boundaries being determined automatically by said software using the positions of said data input areas, data derived from said data capture areas being allocated by said software to associated said data input fields automatically.

13. The system of claim 12, wherein the data input device is chosen from the group:

(i) a digital pen;
(ii) a scanner;
(iii) a PDA and stylus system
(iv) a tablet PC and stylus system; and
(v) a touch sensitive screen.

14. The system of claim 12 which comprises a device from the group:

(i) a digital pen;
(ii) a scanner;
(iii) a PDA and stylus system
(iv) a tablet PC and stylus system; and
a touch sensitive screen.
Create data entry area

Create data capture area

Print/Display form

Write on printed/displayed form

Capture data falling within data capture areas

Perform ICR/OCR processing on captured data

Assign processed data to digital data entry field associated with data capture area

Process digital data entry fields
Create data entry areas

Automatically generate data capture areas according to a dimension of data entry areas

Data capture areas overlap?

Assign data capture areas to data entry areas

Move data entry areas

Reduce size of data capture areas

Reposition data capture areas in accordance with repositioned data entry areas

Move boundaries of overlapping data capture areas 212 to be equidistant between boundaries of respective data entry areas 210

Move boundary of one an overlapping data capture areas more than the boundary of other of overlapping data capture area

Assign data capture areas to data entry areas

END

Figure 13
User enters written data on to form

Determine centre of gravity (CoG) of strokes forming the written data

Does CoG fall within a data capture area?

Does CoG fall within a single data capture area?

Assign stroke to data capture area

Assign stroke to the data capture field that has a CoG nearest to the CoG of the stroke

Assign stroke to nearest data capture area to the CoG

Figure 20

Figure 21
Figure 28

Figure 29
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

<table>
<thead>
<tr>
<th>IPC</th>
<th>Classification Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 G06K9/20</td>
<td></td>
</tr>
<tr>
<td>9 G06K9/22</td>
<td></td>
</tr>
</tbody>
</table>

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

**Minimum documentation searched (classification system followed by classification symbols)**

<table>
<thead>
<tr>
<th>IPC</th>
<th>Classification Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 G06K</td>
<td></td>
</tr>
</tbody>
</table>

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

**Electronic data base consulted during the international search (name of data base and, where practical, search terms used)**

EPO-Internal, PAJ

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>US 5 652 806 A (FRIEND ET AL) 29 July 1997 (1997-07-29) abstract 29</td>
<td>1,2,5,14-16</td>
</tr>
</tbody>
</table>

| X        | US 5 455 901 A (FRIEND ET AL) 3 October 1995 (1995-10-03) abstract 30 | 1,2,5,14-16 |

| X        | US 6 614 929 B1 (YOKOTA KAZUAKI) 2 September 2003 (2003-09-02) abstract; figures 5-8 column 4, line 36 - line 63 | 1,14-16 |

**X** Further documents are listed in the continuation of box C.

**X** Patent family members are listed in annex.

* Special categories of cited documents:

- **A** document defining the general state of the art which is not considered to be of particular relevance
- **E** earlier document but published on or after the international filing date
- **L** document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- **O** document referring to an oral disclosure, use, exhibition or other means
- **P** document published prior to the international filing date but later than the priority date claimed

**T** later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

**X** document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

**Y** document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

**X** document member of the same patent family

**Date of the actual completion of the international search**

14 July 2005

**Date of mailing of the international search report**

01/08/2005

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fac. (+31-70) 340-3016

Authorized officer

Müller, M
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
</table>
| X        | US 5 717 939 A (BRICKLIN ET AL)  
10 February 1998 (1998-02-10)  
column 6, line 49 - column 8, line 12  
column 12, line 17 - column 13, line 12  
abstract; figure 13                  | 1, 2, 5, 7-11, 14-16 |
| X        | US 5 631 984 A (GRAF ET AL)  
20 May 1997 (1997-05-20)  
abstract; figure 1                   | 1, 14-16             |
| A        | PATENT ABSTRACTS OF JAPAN  
vol. 014, no. 183 (P-1035),  
12 April 1990 (1990-04-12)  
& JP 02 0209885 A (TOSHIBA CORP),  
31 January 1990 (1990-01-31)  
abstract                            | 10                   |
<table>
<thead>
<tr>
<th>Patent document cited in search report</th>
<th>Publication date</th>
<th>Patent family member(s)</th>
<th>Publication date</th>
</tr>
</thead>
<tbody>
<tr>
<td>US 5652806</td>
<td>A 29-07-1997</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td>US 5455901</td>
<td>A 03-10-1995</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CN 1112759 A ,C</td>
<td>29-11-1995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69414530 D1</td>
<td>17-12-1998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE 69414530 T2</td>
<td>10-06-1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES 2126072 T3</td>
<td>16-03-1999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 7200720 A</td>
<td>04-08-1995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SG 46207 A1</td>
<td>20-02-1998</td>
</tr>
<tr>
<td>JP 02029885</td>
<td>A 31-01-1990</td>
<td>NONE</td>
<td></td>
</tr>
</tbody>
</table>