Form Analysis Using Colour and Context

A thesis submitted in partial fulfilment of the requirements of The Nottingham Trent University for the degree of Doctor of Philosophy

Wing Seong Wong

October 2001

Department of Computing and Mathematics
The Nottingham Trent University
Burton Street, Nottingham, NG1 4BU, UK
wing@doc.ntu.ac.uk
Tel: +44-115-8484742; Fax: +44-115-9486518

ABSTRACT

Despite the advances in computer technologies, the automatic processing of form documents, especially those that are filled with cursive writing, remains an unresolved problem. In many cases, incoming forms must still be entered manually into the document management system before the data can be electronically processed.

This work aims to contribute ideas to improve the form data extraction and recognition processes to realize a more reliable automatic form processing system. It investigates the possibility of using colour to improve the data extraction process and Optical Character Recognition (OCR) to retrieve contextual knowledge to improve the Cursive Script Recognition (CSR).

An innovative colour reduction technique is proposed that can successfully reduce the colour content of form documents based on a direct comparison of the pixels' RGB value. Using these quantised forms, the use of colour to aid the extraction of the filled data is then investigated. Three experiments are conducted to assess the effectiveness of such a method. Experimental results show that an extraction system that utilizes colour information will improve the recall rate from 96.5% to 99% and accuracy rate from 97.5% to 99%, with an extraction speed that is up to 3 times faster than a black and white extraction system. The effectiveness of the new extraction method over a black & white technique is reflected in a significant improvement in the CSR rate (up from 49% to 58%) and at the same time as reducing the need for the commonly used text repair algorithms.

The novel concept of using OCR to aid CSR by extracting the contextual knowledge has also been demonstrated. OCR generated cues are used to reduce the CSR search space by limiting the lexicon size for a given field. The experimental results show that using current OCR technology, cues can be successfully located 99% of the time resulting in an improvement of the CSR rate by an average 12% (from 43% to 55%).

Finally, a further study has been conducted to investigate the feasibility of using the developed methods to process a filled form without using the equivalent blank form image. The experimental results show that although the extraction rate drops from 94.1% to 82.7% when the blank form is not available, most of this decrease is the result of miss-retrieved OCR text rather than the filled-in words. The actual CSR rates reduction only drops by around 1.8%.

The work described in this thesis is the author's own unless stated otherwise.

It contains – to the author's best knowledge – original material.

ACKNOWLEDGMENTS

Many people have shared in this endeavor and all of them deserve a great deal of thanks. As always, my first thanks must go to my wife Jiun, who made all the sacrifices during the writing of this thesis. She endured the late nights and many working weekends, and I thank her for her patience and encouragement during this time.

I want to record my great debts of gratitude to my supervisors, Dr. Nasser Sherkat and Dr. Tony Allen. They gave me lots of inspirations and valuable feedback during my research, helped keeping me on target and prevented me from dispensing what would otherwise be confusing advice. Special thanks must go to Tony Allen for his extra ordinary effort in helping me strengthen this thesis. His editing and sharp-eyed proofreading have saved me many times.

A heartfelt thanks to Mark and Annie for their kind deeds and constant encouragement.

I thank Kenny and Sharon for their support and helpful advice.

Finally, I thank my parents for all their contributions; without them this thesis would not have been possible.

TABLE OF CONTENTS

Indices and Tables

List of figures

LIST OF TABLES

1. INTRODUCTION

A form is a special type of document that is used to capture data and information. Once captured, this data must then be extracted and processed in order to fulfill the purposes for which the data was required. Traditionally, most of this data has to be manually 'keyed' into the computer system before it can be processed. This is especially so for form documents that are filled-in with cursive handwriting. Unfortunately, this manual capturing process is both tedious and prone to errors. The process also requires many staff hours and can be very costly. According to [1], it costs 7-8 cents to process a single payment (one bill and one Cheque) in most utility companies in Canada. In the United States, the cost of manually capturing data from a form has been estimated to be about $2.50 per form [2]. Considering the amount of payments or forms that need to be processed every year, there is a huge amount of money involved. Another problem associated with the manual capture of data from forms is that, even after the data entry process, the original forms need to be kept for legal or audit purposes. Without the automatic indexing of form images on the basis of their entered data, subsequent inspection of these documents can be a labour intensive process. Automatic form processing and data capture, with its potential to rationalize the situation, has thus become a major research area.

Ideally, an automated form processing system should be able to capture all of the data from a form. Unfortunately, this is not the case and most of today's form-processing systems still require the need for human intervention to verify and correct errors produced by the system. Proper form design has been shown to be effective in reducing the automated data capture errors [3] but in many cases, the re-design of a form is difficult, if not impossible. If the targeted forms have been in existence for some time and thousands of forms have been filled out well before the automated processing system has been implemented, then the potential for system improvement via form modification is limited. Moreover, even if the form can be redesigned, many of today's automated form-processing systems still fail to reliably process hand-filled data forms; especially those filled with cursive writing. This is due, principally, to the poor performance of the handwriting recognizer(s). These restrictions thus limit the potential of current form processing solutions and point to the need for a better and more reliable system to realize the full potential of office automation.

1.1 Overview of An Automatic Form Processing System

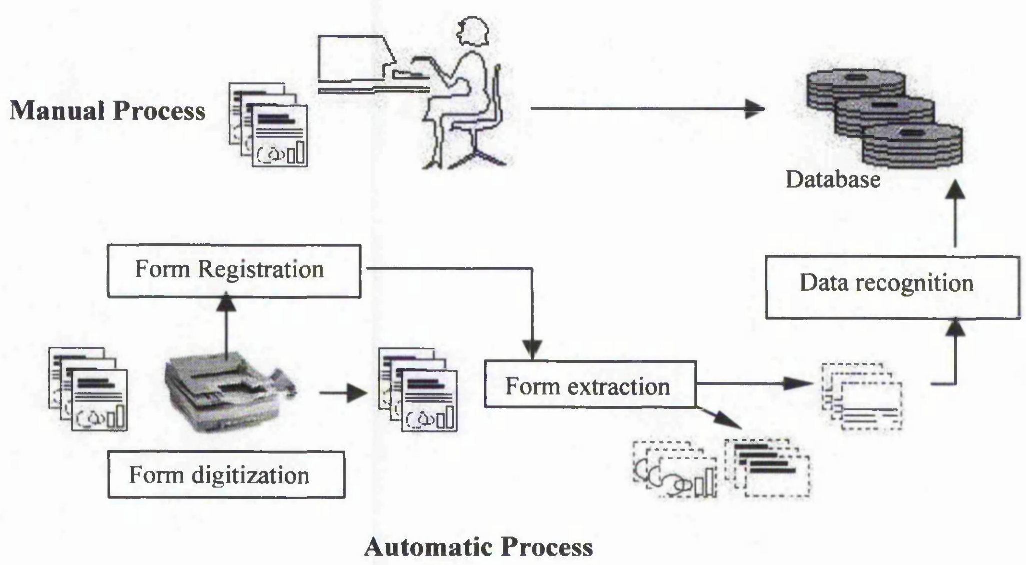
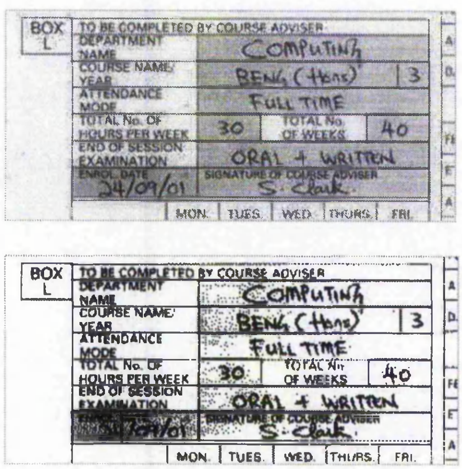


Fig. 1-1. A manual form processing system versus an automatic form processing system

In general, there are 4 main processes in an automatic form processing system. These include the digitization process, form registration, form extraction and data recognition (fig. 1-1). Unfortunately, each of these processes can produce error and noise effects that, when propagated through and multiplied by subsequent processes, can lead to problems or even eventual failure within the entire automation system.

1.1.1 Form Digitization

Form digitization is the process of transforming a paper form into an electronic format ready for computer processing. The most common device used to digitize a form is a scanner. Traditionally, forms are digitized into either gray-scale or black & white images for reasons of cost and computational overhead. Although these images require much less computer storage space than colour images, the loss of colour information from some forms can produce noise effects that could introduce errors into the automated system (fig. 1-2).



Colour form images

Black & White form images

Fig. 1-2. Scanning forms in black & white can produce noise and cause loss of some of the image details when compared to scanning forms in colour.

With the development and cost reduction of colour scanners and storage media, processing forms in colour has now become feasible. Currently, research into handling colour features in document analysis has become the focus for many researchers [4,5,6,7,8,9].

1.1.2 Form Registration

A typical filled-in form consists of the following three components:

1. Preprinted components – such as logos & machine printed text
2. Structure components – such as lines & boxes
3. Filled-in data – including machine-type, machine printed and handwritten data (cursive words or hand printed characters)

The first two components are regarded as the preprinted entities. Distinguishing these entities from the filled-in data is one of the fundamental problems in form processing. According to the work in [10,11,12,13], the characteristics of these entities can be classified as follows:

- Forms contain many straight lines oriented mostly in horizontal and vertical directions
- The information that should be acquired from a form is usually the filled-in data
- Pre-printed texts in form documents often contains a fixed set of known words, which can be recognized with current OCR software

Currently, a manual process is needed to determine the position and purpose of the fields on a form. Thus, for each form design that the system is required to process, a person must manually create a form description file, describing the geometry of the fields on the form and the contextual information associated with each of these fields [10,14,15,16]. Form registration is the process of mapping the filled form image to this form description file so that the information provided by the description file can be used to help extract the filled data. There are two processes involved in form registration - form identification and mapping. In a form processing system that is required to handle several different kinds of form, a form recognition process is needed so that the correct description file is invoked to process the form. Many techniques have been proposed that can reliably recognize a form based on geometric features such as lines and cells [17,18,19,20] as well as pixel level features such as block reduction & comparisons [21]. Several robust algorithms [2,22,23] have also been proposed to map the pre-printed entities from the template to the filled form using affine invariants [24], geometric hashing [25] and statistical analysis.

The major draw back in using a template to extract the data is that the description file must be re-created whenever there are changes made in the form design and different description files must be created for different forms. Interactive tools that provide a graphical interface for manually marking-up and creating these description files [26,27] do facilitate this process but do not solve this problem completely.

Another potential problem with this process is that when there are many form types that the system needs to process, form identification accuracy will be reduced and the wrong template could be invoked to extract the data. Although current form identification or recognition techniques can achieve accuracies greater than 96%, the recognition rate generally reduces as the number of form types increases. Thus, a more robust form recognition method might need to be developed as the total number of form types increases.

1.1.3 Form Extraction

Form extraction (or form dropout) is the process of extracting the filled data from a form. This process involves the removal of pre-printed entities (lines, boxes, logos and machine printed text) so that only the filled data is presented to the recognizer(s). Generally, there are two methods for extracting the filled-in data from a form - an image threshold method and a subtraction technique [28]. If the thresholding technique is considered, several colour dependent strategies must be adopted. Whilst this technique is simple to use and implement, it requires a careful design of the forms and thus limits its usage to applications where form re-design is allowed.

The subtraction technique on the other hand, is strongly dependent on the effects of noise, offset and skew between the template and the filled forms. It requires complicated and time consuming analysis to overcome problems such as capturing filled data that is outside the pre-defined boundaries and minimizing the distortion to an image which results from removing pre-printed lines that intersect the data. Fig. 1-3 shows a typical flowchart for a form extraction system that employs a subtraction technique.

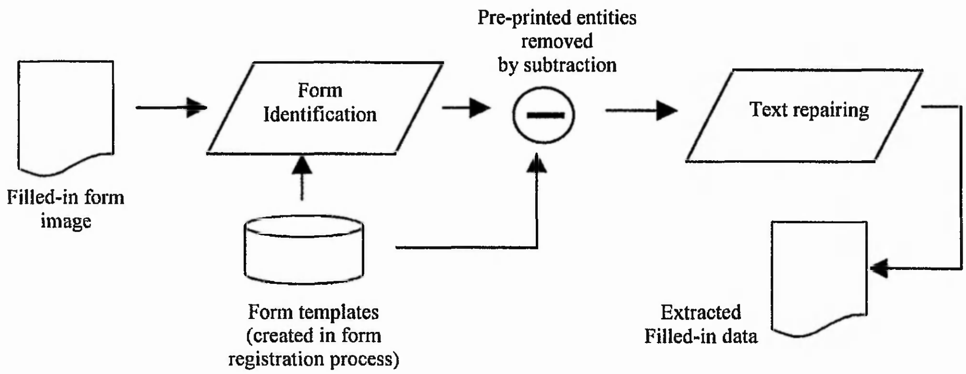


Fig 1-3. A typical form extraction system flowchart based on the subtraction methodology.

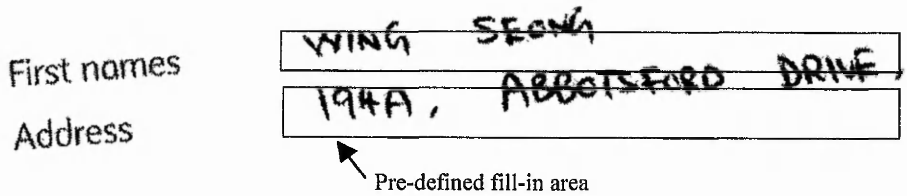


Fig. 1-4. Page skew will cause an entity mismatch with the corresponding entities in the template, reducing the extraction efficiency and accuracy.

Fig. 1-4 shows how page offset and skew, induced by the digitization process, can cause errors in the registration process and consequently reduce the form extraction efficiency and accuracy.

The following is a list of problems that could be found in a subtraction process:

- Broken text due to line removal process
- Distorted text due to imperfect text restoring algorithms
- Incomplete entities removal
- Correlation position errors between the blank (template) and the filled form

1.1.4 Data Recognition

In any automatic form processing system, data recognition is the most important process. After all, it is the filled-in data that an automated form processing system is trying to capture. The more data a system can 'read' correctly, the less human dependent the machine will be. Many recognition technologies had been developed over the past 40 years to help automate this data capturing process. These range from Barcode and Optical Mark Recognition (OMR) readers to Optical Character Recognition (OCR), Intelligent Character Recognition (ICR) and Cursive Script Recognition (CSR) systems. In terms of speed, accuracy and cost, barcode technology is well proven and is far superior to keyboard entry.

However, barcodes can only be pre-printed onto a form and, therefore, can only be used to help improve processes such as form identification and recognition. OMR technology, on the other hand, is widely used for data entry in multiple-choice applications. The key function of OMR is to detect and identify an entry in a particular location on a form and interpret the information according to a pre-defined set of rules and values. Major applications of OMR are the National Lottery, multiple choice exam paper marking, market research and mail order catalogues reading. However, form design is critical for OMR applications. Under controlled environments where the respondent has only a few options to choose from, OMR is the best choice for data entry and recognition.

OCR is a character-based recognition technique that is capable of recognizing machine printed fonts and alphanumeric handprint. This recognition is based on a matrix or template matching technique where each character is compared to a set of prototype characters in the database. The recognition rate varies, but for a well laid out document (such as a pure text letter or article), scanned at 300dpi resolution, a recognition rate of 99% can generally be achieved.

ICR is now the most commonly used technology to read handprint characters within an automatic form processing system. It normally incorporates neural network methods to cope with the high variation in human writing styles. However, as style varies dramatically from person to person, ICR can only classify the characters to within a given confidence level. In addition, ICR suffers from the same problems as OCR in that the recognition rate drops dramatically when hand-printed characters are touching or when low quality texts are to be recognized. The recognition rate for ICR is also much lower than OCR as human writing styles have a much greater variation in character formation than machine printed text where the character shape and spacing is relatively predictable and consistent. Table 1-1 gives a summary of the error rates for each of the key recognition technologies as reported by the industry in [29].

Recognition technique	Typical Error Rate (based on raw techniques without validations)	Comments
Barcode	0.05-0.5%	Depend upon code used
OCR (single font)	0.05-0.5%	Assume standard fonts and good printed quality
OCR (Omni-font)	2-5%	The more font there are, the greater the error rate
ICR (Constrained) - Numeric - Alpha - Alpha numeric	 10% (5%) 15% (8%) 20% (10%)	Brackets show the error rates after validation and edit checks process
Manual Key Input	0.5-2%	Involved verification such as manual re-key again

Table 1-1. A summary of the typical error rates for each of the capture methods as reported in the industry in [29].

Cursive Script Recognition (CSR) is, by far, one of the most challenging research areas in recognition technology. Again, this is due to the vast variability in human handwriting, both between different writers (inter-writers) and within the same writer (intra-writer). Although much research has been done in order to address the problems of reliable cursive handwriting recognition, results are far from satisfactory (see chapter 2). Fig. 1-5 (overleaf) shows an example where all these key recognition methods are applied to a filled form in an automatic form processing system.

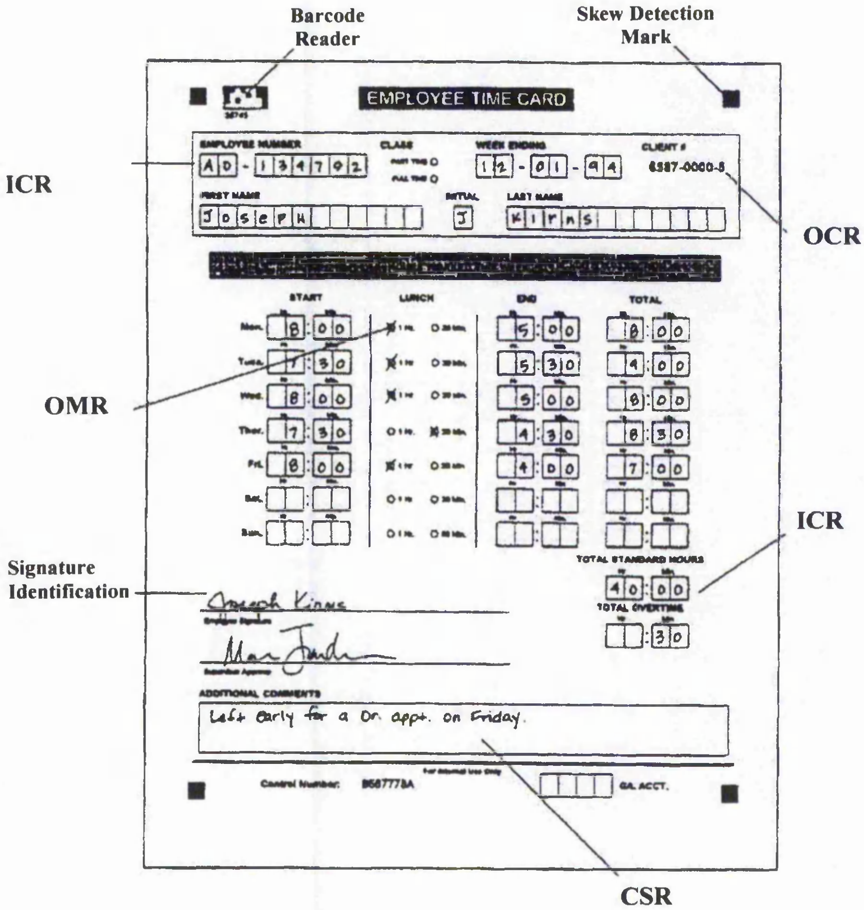


Fig. 1-5. A sample form that requires several key recognition technologies in order to automate the data capture process.

As stated earlier, one of the fundamental problems in handwriting recognition is the variability of handwriting. The characteristics of handwriting such as slant, height of ascenders, letter connectivity, word length etc. differ from person to person. Clearly, coping with such diversity is a challenge for most of today's recognizers. According to Tappert [30], handwriting can be characterized into 5 basic categories:

- Box discrete characters
- Space discrete characters
- Run-on discretely written characters
- Pure cursive script writing
- Mixed cursive script writing

Fig. 1-6 shows an example of each type of handwriting that could be found in a form document.

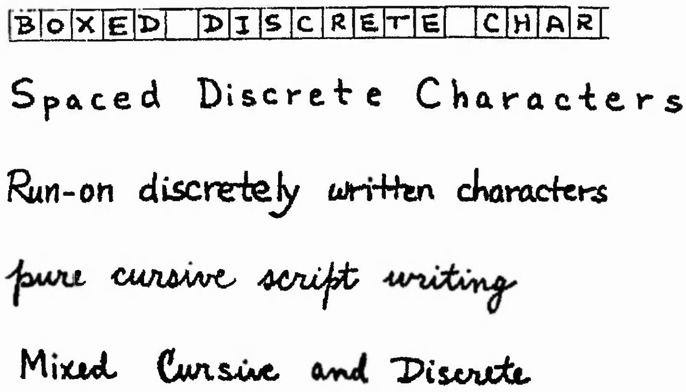


Fig. 1-6. Types of handwriting as defined by [22].

The difficulty of the problem can be described in terms of the character segmentation required. Generally, the more the characters are touching with each other, the lower the recognition rate will be. Most of today's ICR engines can recognize a boxed or spaced discrete character with an acceptably high accuracy rate. However, ICR will normally fail to produce good recognition results with run-on discretely written characters and cannot handle cursive script at all. The poor CSR performance on the cursive and mixed cursive discrete characters serves only to deepen this problem. To help over come this, most of today's forms contain boxes or lines to guide the writer into using boxed discrete characters. Unfortunately, as proven in [3], even with all these guidelines in place, writer idiosyncratic responses on forms continues to be one of the major sources of error in an automated form processing system

Besides variability, recognizers are also required to cope with ambiguity. Some pairs of characters, such as 1 & I, 0 & O, 5 & S, Z & 2, G & 6 are very difficult to disambiguate.

Many upper and lower case characters, such as C & c, K & k, P & p etc. can be identical when written by a human. In cursive words, some character-pairs are also confusing, in particular 'cl' and 'd', 'm' and 'nn' and 'uu' and 'w'. Another problem in handwriting recognition is the illegibility of poorly written handwriting. Sometimes, words are distorted to a point where they are illegible. Humans then make use of contextual information [31] and, in extreme cases, often resort to guessing when reading such distorted words.

Many techniques have been developed in the past to help improve the handwriting recognition rate. One commonly used technique in form processing applications is to manually define and link the filled-in data to the contextual information related to that field. In this way, different lexicons can be used to recognize the data. This information can also be used to choose different recognizers to perform the recognition task so that better recognition rates can be achieved. However, this technique is limited in its application to a specific form and is as sensitive to skew and page offset as is the form extraction process. In addition, every time there are minor changes made to the form design, the areas have to be manually re-marked-up and linked to the contextual information again. The same is true when a new form is added to the system.

1.2 Overview And Contribution Of This Work

This work investigates the problems involved with form extraction and handwriting recognition in a colour-based form processing system. The initial stage of the study focuses on developing an efficient method for handling colour images so that colour information could be utilized for form extraction. As processing a form in a 24-bit full colour format requires much higher computational overheads than an equivalent black and white system, a novel colour reduction technique is introduced. The speed, recall, accuracy rate and extracted data quality improvement of this colour-based form processing system has been fully investigated and compared to its black and white equivalent.

The later stage of the research concentrates on studying the feasibility of **automatically** obtaining high-level contextual knowledge from the pre-printed text in a form (using a commercial OCR+ engine) to assist the CSR. A novel linking algorithm has been proposed to link the cue words to the filled-in words. The method has been tested with a developmental CSR engine [32] and the recognition performances are reported.

The final part of the work then attempts to investigate the possibility of using the developed techniques to process a filled form without the use of blank form knowledge. The work demonstrates a feasible method for using just an OCR engine and a line searching method to extract the filled-in data without using a blank form template. The

+The OCR package chosen for this work is TextBridge Pro from Xerox Imaging System. It can perform recognition task at different resolutions and can handle most images formats (from black & white to colour).

extraction performances has been assessed and compared to the results obtained from a form extraction system that utilizes blank form knowledge.

The following publications describe the novel methods that have been devised as a result of this work:

[33] Wing Seong Wong, Nasser Sherkat, Tony Allen, "Use of Colour in Form Layout Analysis", IEEE proc. in International Conference In Document Analysis and Recognition, pp. 942-946, 2001

[34] Wing Seong Wong, Nasser Sherkat, Tony Allen, "Contextual Focus for Improved Recognition of Hand-Filled Forms", IEEE proc. in International Conference In Document Analysis and Recognition, pp. 748-752, 2001

[35] Wing Seong Wong, Nasser Sherkat, Tony Allen, "Form Extraction by Colour Information", 4th IAPR International Workshop on Document Analysis Systems, pp. 109-120, 2000

1.3 Outline Of The Thesis

This chapter sets the scene for this work by introducing the issues involved in automatic form processing. Chapter 2 provides a review of the state of the art in colour handling, form processing, form extraction, handwriting recognition and other related areas. Chapter 3 looks into the concept of using colour information to aid the extraction of handwriting data from colour forms. It explains, in detail, how the colour content of a form can be reduced and how form colour can be used as a tool to aid the form extraction process. The effectiveness of the developed method is shown by comparing the extraction recall, accuracy and CSR rate to that of a black and white extraction method. Chapter 4 provides a comprehensive explanation of the methods used to automatically link the contextual knowledge contained within a form to the handwriting data in order to improve the CSR performance. Quantitative findings of the advantages gained by using such a method are presented in the latter part of this chapter. Chapter 5 investigates a possible method for the processing of a filled form without using the blank form knowledge. Chapter 6 concludes the work and provides suggestions that emanate from the experiences gained whilst conducting this research and the findings that this work has produced. These suggestions also serve to link this PhD work with other potential projects in the future.

2. LITERATURE REVIEW

This chapter provides an overview of the methods that have been proposed to address the problems involved in an automated form processing system; in particular those appropriate to the areas of form extraction and handwriting recognition. In addition, as this thesis deals with form extraction utilizing colour information, problems associated with colour reduction and colour-handling techniques are also discussed.

The techniques and problems involved in form extraction in an automated form processing system are first described in detail in section 2.1. This is followed by the colour handling techniques that have been reported in the literature in section 2.2. Section 2.3 then provides an overview of the principal approaches to Cursive Script Recognition and their application together with a summary of results.

2.1 Form Extraction

There are two fundamental approaches for extracting the filled-in data from form images – Model based and Model-less. In a model-based approach, the filled-in data is extracted by using a reference template created from a blank form image. This template can be created either manually, semi-automatically or fully automatically. Fig. 2-1 shows a typical model-based form processing system diagram [2]. Note that in a manual template creation and recognition system, the form modelling process might not be present. However, in each of these cases, a blank form sample must be available. A model-less approach, on the other hand, is a form processing system that does not utilize any blank form knowledge to extract the filled-in data from the filled form images. Whilst it is much more difficult to achieve good results using this approach, the model-less approach gives far greater flexibility to the system when dealing with different kinds of form design. This is especially important if there are a large number of form types that have been produced at different places and times (for example scaling and design variations within the same form type). Most of the time, methods that are developed in a model-less system can be applied to a blank form to automatically generate the form template for the extraction process in a model-based system.

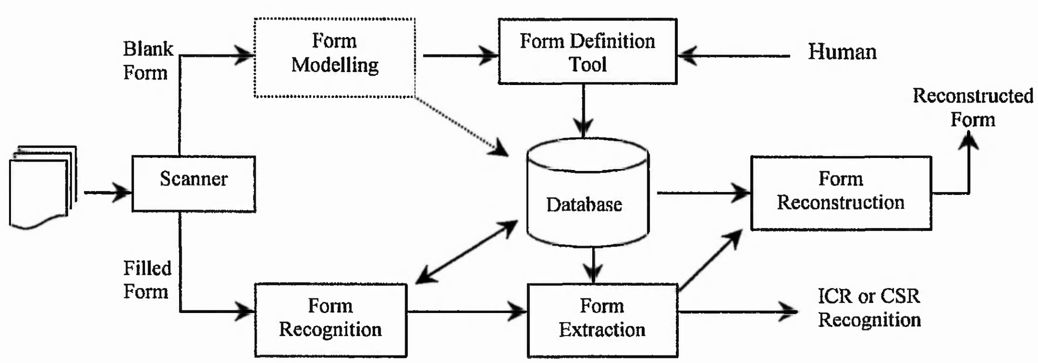


Fig. 2-1. A typical model-based form processing system diagram [2].

2.1.1 Form Modelling and Recognition

The form modelling and form recognition processes are sometimes referred to as form classification in a form processing system. Most of the time, the techniques developed for the form modelling process are the same as those for the form recognition process. This is due to the fact that the aim of a form modelling process is to extract the features of a blank form to create the template for registration and extraction purposes whilst the form recognition process is required to compare the features of a filled form to the template database in order to choose the right template for extraction. Obviously, the features that were used to create the template and the features that are used to distinguish between forms must be the same.

Generally, there are 3 form classification strategies – semantic (extraction of a region of the image that provides discrimination information amongst the form types), pixel-level (block reduction and comparison) and geometric (analysis at a higher level of abstraction than the pixel-level but below the semantic level) approaches. The semantic approach is normally implemented using OCR or barcode recognition methods. A fixed identification string or barcode is used to label and distinguish between forms. In [36] however, attempts have been made to use the purposely-added solid black squares and fixed black lines (at given locations) in a form to identify the topology of a form and locate the targeted fields. Unfortunately, no quantitative data on the field registration accuracy and extraction rate are reported. The major problem using this semantic approach is that the position of this semantic information must be consistent and that the printed information (either barcode or machine printed identification number) must be of good quality. This limitation has thus reduced the usefulness of such an approach in real life applications, where form re-design is sometimes difficult and impractical.

The pixel-level approach on the other hand is normally implemented through a block reduction method. The form image dimension is normally reduced by averaging the pixel values within each of the blocks whilst ensuring that enough detail is left to discriminate between forms. However, such an approach has proved to be inadequate and difficult to optimise [23] for effective classification.

The third approach in form classification is by geometric analysis. There are 2 different interpretations of geometric information; one is by geometric structure - which corresponds to the physical characteristics of each entity found within a form, the second is by logical structure - which describes the inter-relationships between each of the entities in a form.

One of the most common structure-based form classification techniques makes use of the line features that exist within a form. This is due to the fact that lines are often the most prominent and important features within a form. In [2], the locations of long horizontal and vertical lines are used to distinguish form images and select the appropriate template. The matching process is based on a feature-matching threshold, which is proportional to the number of features that exist within a form. A template is chosen based on the number of feature matches between the template and the filled form, which must be over the threshold setting. A similar approach was also developed in [37] where the length, width and position of horizontal and vertical lines is used to facilitate the classification process. Both methods are sensitive to physical structure variations in the form and are prone to failure if the lines are broken. To counter the broken line problem, a more robust line-crossing feature was introduced in [23]. A neural net classification engine was employed that uses nine types of line crossing feature to distinguish between form images and a success rate of 98% was achieved. However, this approach requires a training stage that is not conveniently implemented in a practical setting. In addition, the classification performance is greatly affected by the number of filled-in data words that cross the form lines, to create nonessential intersections. In [38], features such as line segments, regions and landmarks were used to perform the form classification and registration process. The method is implemented successfully but unfortunately no quantitative results on the field registration accuracy are reported.

Logical structure on the other hand defines the inter-relationship between each form entity and thus represents a higher-level feature than a single entity such as a line or intersection. Data structures that are commonly used include strings and trees. The use of strings for form classification has been proposed in [39,40,18]. In these works, the geometric layout of objects such as lines, text and spacing on a form is converted into a linear string representation. This approach is claimed to be quicker and more robust than any of the other methods and, with optimum setting, can achieve nearly 100% accuracy. However, such an approach will not be useful if the geometric layout is liable to change with time and if a higher level of layout knowledge is needed. To retrieve higher level layout knowledge, a generic top down approach was proposed in [41]. Three binary trees are constructed, two of them are for global and local structure and one is for classification. Besides lines, some of the pre-printed data is also used. Unfortunately, the detail of the implementation is not given and no experimental results were provided. The hierarchical representation of form documents was then used in [17] to perform data extraction and form retrieval. Promising results had been claimed but again no further details were disclosed.

2.1.2 Form Extraction

In a typical form extraction system, the processes involved in extracting the filled-in data from a filled form are form registration, pre-printed entity subtraction and data restoration.

Form registration is the process of mapping the template created in the form classification and definition process to a filled form so that the filled-in data can be located. This process is needed due to the skew and page offset that could be induced during the scanning process. Many robust methods had been proposed and developed that can produce a skew angle estimation accuracy of between +/-(1-3) degrees. Methods proposed include connected component projection profiles [42], Run Length encoding & Hough transform [43] and histograms analysis [44]. Skew detection and correction is considered a fully resolved area now and many skew detection methods have even been implemented within the scanning devices themselves. The methods that are used to determine the vertical and horizontal offset between the template and the filled form rely on the same features that are used in the classification process. For example, in [23], line-crossing features are used to measure the displacement. Within a +/- 25 pixel region of each of the template positions, the registration system will try to find a mapping offset value. A threshold of 3 pixels value is used to map a vertical or horizontal displacement point to the template. Mapped points are then summed and the scores used to determine the best-offset value for the form. Alternatively, in [2], a simple line location comparison is used to calculate the offset value. In this case, the form is first de-skewed and then the line positions found in the blank and filled forms are compared. The difference value obtained is then the offset value used for the form.

Once the page offset has been determined, the filled-in data is then extracted using the information supplied by the template. This information is normally in the form of pairs of corner points indicating the positions of the filled-in areas or the pre-printed entities. In the case of filled-in areas, every pixel enclosed with the boxes formed by the corner points will be extracted and a further non-data removal process applied. The non-data removal process is needed because, in some cases, pre-printed entities are enclosed within the boxes and are, thus, accidentally extracted by the system. If the template information is in the form of pre-printed entity location, a subtraction process is employed. All the pixels enclosed in the boxes will be removed and the remaining pixels should be the filled-in data.

In both of these cases, some portions of the filled-in data will be inevitably removed, especially when the filled-in data is touching with the lines or pre-printed text. Text restoration is a process developed to resolve this degradation issue. In [2], a method is proposed to minimize the distortion of the filled-in data during the line removal process by using an intersection point analysis. In this process, pixels that touch form lines are recorded and when the intersect points are bigger than a given threshold, the line portion at the intersect points is retained. In [45], several patching techniques were employed to restore all of the broken characters that result from the line removal process. Arc patching, binding arc patching and quadrilateral patching were employed to restore the detected intersection points between the lines and the text. The preliminary experimental results were promising but several issues were left unresolved, namely: filled-in words touching pre-printed text, confusion between filled-in characters and small check box areas and the differentiation of filled-in data from the pre-printed text. These problems were addressed later by X. Ye et al in [46] where 97.4% of the characters that were touching with pre-printed text were successfully separated using stroke width comparison. Unfortunately, the limitation of this method is that the filled-in stroke width must be different to the pre-printed text. This thus limits its scope to handwritten filled-in data that has a different stroke width to the pre-printed text. In [47], Hidden Markov Models approach was employed to distinguish the handwritten text from the machine printed text. A 72.2% recall rate and a 92.9% accuracy rate were reported. Unfortunately, in this work, there is no text separation process employed and thus, the reported results were the overlapping identification rate rather than the actual separation rate.

Application Name (Approach/template creation method)	**Method**	**Achievement**
US IRS form processing system (Model-Based/Semi-Automatic)	Extraction by line intersections features [48] (Geometric Structure) Extraction by using ten line junctions features [23] (Geometric Structure)	Field Registration = 95% Form Recognition = 100% (Test set size = 25 forms) Extraction rate = 99.5% Form recognition = 98%
Intelligence Form Processing System (Model-Based/Manual)	Extraction by form mapping using line and boxes features (Geometric Structure) [2]	Extraction speed = 10 seconds
Generic Form Dropout System (Model-based/Automatic)	High level morphological subtraction [49] (Logical Approach) Four Directional Adjacency Graphs to locate form fields (Logical Approach) [50]	Very robust method for extracting data from grey-scale images Preliminary results are claimed to be promising but no detailed results are given
Generic Form Dropout System (Non-Model based)	Extraction by Block Adjacent Graph (BAG) method [51,52,53] Connected Component Analysis method with dilation and erosion for extraction of lines component [54] Extraction by using types of line segments features and fuzzy matching for form recognition [55] Form structure detection using strip projection [56,57]	Solves most lines/text overlapping issues 98.5% of fields detected 99.1% of lines detected 0.7 to 1.6 seconds Feature Extraction time = 1.77s to 6s Fast and robust, better than Hough transform and run length based algorithm

Table 2-1. Summary for some of the proposed form extraction systems, their proposed approaches and reported performances.

Table 2-1 shows a summary of other proposed systems, detailing their approaches and reported performances. From the reported work, it appears that problems related to finding vertical and horizontal lines are well addressed and close to being fully resolved.

The text degradation problems that are caused by the line removal process have also been fully investigated by many researchers and some very efficient methods have been developed.

Line features are the strongest and most prominent characteristics in form documents, thus it is no surprise that most of the systems proposed are based around these features. However, as this feature has been fully explored it can no longer be used to help improve the system performance and so other features must be exploited. As demonstrated in [58], one of the possible features to be used is colour. In this work, colour has been used successfully to extract the signature and seal imprint from cheque images. It is thus believed that this concept can be extended to a form-processing environment where it has a greater potential for further improvements in the current form processing systems.

2.2 Colour Reduction

Basically, there are two reasons for reducing the colour content in an image – first, to save memory consumption and second, to reduce the system complexity and computational cost [59]. The first approach is to reduce the colour information in an image as much as possible whilst minimizing the visual degradation that could be caused by this action [60,61,62,63]. Its main advantages are: smaller images file size for faster transmission and lower storage cost for image storage and retrieval systems. The most common procedures for such applications involve a transformation of the colour format from a RGB colour domain to a human visual model domain [62]. These techniques attempt to keep the images as close as possible to human perception even after the colour reduction process. The second approach, which is most closely aligned with this work, attempts to reduce the complexity of the document analysis and processing system. Many techniques [4,5,6,7,8,64,65] have been proposed that attempt to address the colour issue for text extraction from colour documents. Some of these work directly in the RGB colour model [4,6,64,65], while others convert the images to a HVS (Human Visual System) colour model [5,7,8]. The disadvantages of using a HVS model for such applications are the accuracy and efficiency loss during the conversion [66]. In fact, as suggested in [4], a better segmentation result can be obtained for colour text document when applying the colour reduction method directly onto the RGB space. Fig. 2-2 shows a summarised process flow for each of the techniques proposed. It is important to point out that all of the techniques reported so far make very limited use of colour information for document analysis. Their main purpose is to reduce the colour content in order to simplify the text extraction process.

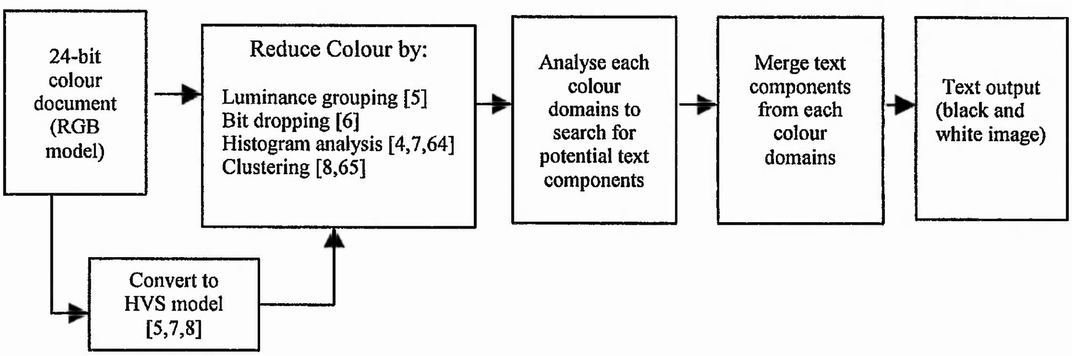


Fig. 2-2. The process flow for the techniques that have been proposed to extract text components from colour documents

The colour reduction process generally consists of two steps. First is the palette design, in which the total number of palette colours (i.e. the targeted number of colours that the image is intended to be reduced to) is defined. Second is a pixel mapping, in which each colour pixel is assigned to one of the colours in the palette.

Bit dropping is the simplest method to use to reduce the number of colours in an image. The idea is that only a given number of significant bits are important for the representation of a pixel colour. For example, in a 24-bit RGB colour image, each pixel is represented by an 8-bit R, G and B value. If we used only the first two significant bits to represent the colour and ignored all the other bits, then the image will effectively be reduced to a 6-bit colour image. This method is fast and effective, especially for colour images that have just a few very distinct colours. However, the major drawback for such method is that the total number of targeted colours that the image can be reduced to cannot be less than 64 (2 bits for each R, G and B component) without losing a significant amount of information. In addition, the number of colours that it can be reduced to is rigid.

Colour histograms are one of the most commonly used techniques to facilitate palette design. Palettes are first found by searching for the dominant peaks in the image colour histogram and assigning each of these peaks as one of the palette colours. The number of peaks defines the total number of colours that the image will reduce to. The main problem for this technique is that not all of the colours in an image give a clear peak in the histogram. This is especially so for complex colour images, where most of the time, text will be separated into several colour domains after this process. A text searching and merging process is then needed.

Luminance and chrominance distances have been used in [5] to classify pixel colours into one of 42 pre-defined colours. A basic group of 21 quantized colours is first defined using the combination of chrominance (Red, Green, Blue, Yellow, Magenta, Cyan and Grey) and luminance values (Dark, Middle and Light). A derived group of another 21 quantized colours is then defined at run-time. This method suffers from the same problem as the histogram method, in that a separate technique is needed to re-merge the text that has been separated into different colour domains.

Pixel mapping is the process of merging the nearby pixel colours into one of the pre-determined palette colours. This process is sometimes referred to as clustering, where a cluster is used to represent each of the palette colours and a mapping process is applied to merge the pixels colour value to its nearest cluster. A frequently used clustering algorithm is the C-means clustering algorithm (CMA) [67], where cluster representatives are iteratively updated and labelled. Other clustering algorithms include Fuzzy C-means clustering [68], learning vector quantization [69] and Kohonen self-Organizing Maps (SOM) [70]. All these techniques have been implemented successfully and all have been shown to be effective in mapping the pixel colours into the palette colours. Unfortunately, for all of these cases, no quantitative results have been given and the evaluation is mainly based on the image visual output. Thus, the performance for each of the methods is very subjective and there is no simple way to compare their effectiveness.

2.2.1 Colour Issues in a Form Processing System

Whilst most of the current form-processing solutions deal with images in black and white or grey-scale, colour has been widely used in the form design stage to enhance the data capturing process. With a careful selection of the colours used in a form, the capturing process can be improved. This is due to the fact that a scanner is sensitive to certain colours under a given operating light source (fig. 2-3). Thus, when colour is used correctly in the form design, certain entities (such as lines & boxes) will be invisible to the scanner and, when digitized, will dropout from the form instantly. However, this method is restricted to applications where form design is under full control. Furthermore, the form colour information is no longer available to the system beyond the digitization stage. This therefore limits the potential for using the colour to help the processes in the subsequent stages.

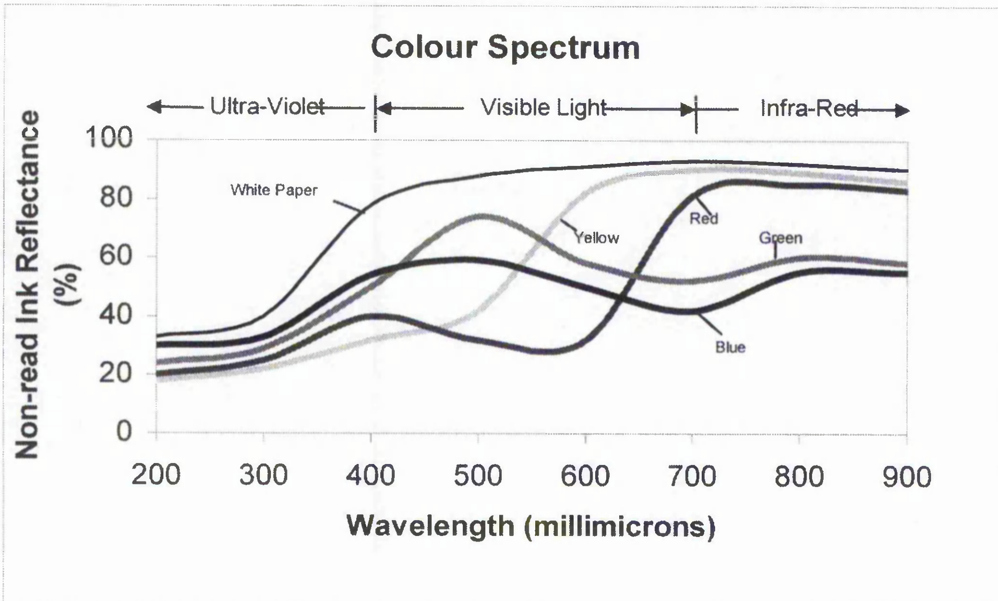


Fig. 2-3. Optical scanners work on the relationship between the type of light, paper reflectance and non-read ink reflectance. For example, if a scanner's light source is at 500 millimicrons, the scanner will be more sensitive to red and yellow colour than green and blue [71].

As mentioned earlier, some of the colour reduction techniques concentrate on reducing the amount of colour present in a given image whilst attempting to preserve most of the visual features that are noticeable to the human eye. Although this is acceptable in cases where file size and visual quality are the only concern, such methods are not of much help when it comes to automatic character/word recognition. This is due, principally, to the fact that a machine vision system 'reads' differently when compared to its human equivalent. This is a direct consequence of the way in which the optoelectronic sensors operate compared to the human eye.

The main difference in behaviour can be viewed as follows:

1. Sensitivity to colours
2. Sensitivity to brightness levels

In a fully automated form processing system, the transformation of a raw RGB colour space to a human visual perception model is thus inappropriate and unnecessary. Current quantisation techniques also effectively segment a document into several colour domains, causing problems for text recognition systems when they attempt to separately recognize the words in a given colour domain. This problem has been explored in [9] where a conventional colour reduction method is employed to handle a colour form image. Currently this effect is not very pronounced as existing recognition systems do not make use of the colour information for text recognition and documents are always converted to black and white prior to the recognition process.

However, if a system were to attempt to utilize colour for processing purposes, the effect could be disastrous. Take for example a 2-colour document (blue text and white background) that has been scanned as a full colour image. With the conventional colour reduction methods, the number of colours that would be used to quantise the document could be 3, 4 or even more. This is due to the fact that, the colour of the text edges can often be miss-classified as one of the targeted colours in the image. This can effectively segment the text into different colour domains (see fig. 2-4). In a conventional form processing system, documents are converted to black and white prior to any recognition process; hence the different blue colour pixels (light or dark) need to be re-united back to the black domain before processing can begin. However, if the system attempted to process the document in colour, the separation of the text into several colour domains could cause a failure in the text recognition process. A colour reduction process that is more amenable to CSR is therefore still needed.

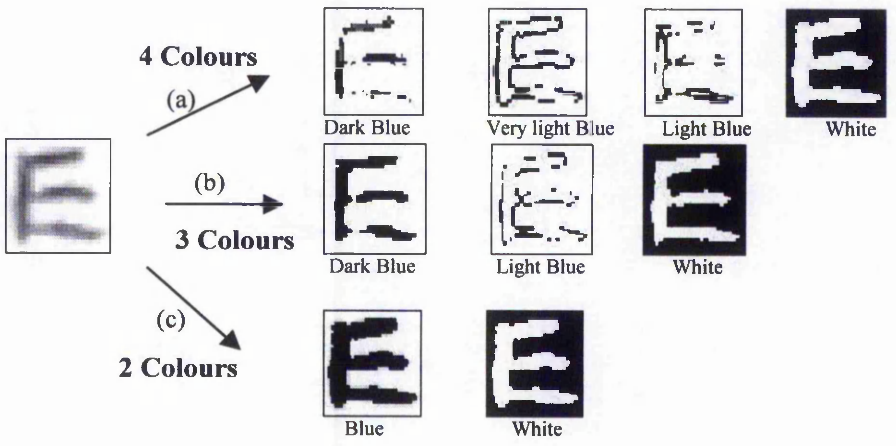


Fig. 2-4. The number of colours chosen to reduce the images will significantly affect the quantization results. The number of colour used in (a)(b)(c) above is 4,3 & 2 respectively.

There are two possible methods for resolving this separation issue:

1). Employing a merging process after the quantization process

2). Defining a smaller number of palette colours before the quantization process

Since a form usually contains only a limited number of colours, the second method is more appropriate if speed of processing is also taken into account.

2.3 Handwriting Recognition

Off-line handwriting recognition remains an extremely challenging and difficult task to accomplish due, primarily, to the vast variability in human handwriting. Although much research has been done in order to address reliable cursive handwriting recognition, current results are far from satisfactory.

2.3.1 Recognition Techniques

There are 3 general approaches for recognizing cursive words:

1. Segmentation-based (character) recognition.
2. Segmentation-free (word) recognition
3. Perception-Oriented recognition

In segmentation-based methods the recognition process is based on an attempt to find the best match between blocks of primitive segments in the word image and the word's letters. It is a difficult and inherently error prone process (see Sayre's paradox* [72]) if the isolated primitive segment is expected to be a character and, until now, no method has been developed that can successfully segment a handwritten word exactly into individual characters [73,74]. Thus, words are always either over or under-segmented. However, the major advantage of using a segmentation-based approach is its flexibility with respect to the size and nature of the lexicon. This is a direct result of the method being character-oriented, i.e. each of the primitive segments could only be one of the 26 alphabetical letters.

* According to Sayre, a letter cannot be segmented before having been recognized and cannot be recognized before having been segmented.

There are 3 basic approaches to achieving an optimal recognition performance in a segmentation-based recognition system, these are:

- Dynamic Programming [75,76,77,78]
- Shortest Path [79,80]
- Hidden Markov Models (HMMs)+ [81,82,83,84,85]

Dynamic Programming is mostly applied to explicit segments (i.e. mapping word segments into individual letters). This requires precise segmentation and minimum spurious ligatures. By mapping a compatible graph to the dynamic programming algorithm and identifying the shortest path from left to right using the Viterbi algorithm, it is possible to handle under or over segmented words (implicit segmentation). HMMs on the other hand can cope with more variation and noise, and are thus able to work on fragments that are not individual letters. HMMs algorithms are therefore most commonly applied in implicit segmentation methods.

In segmentation-free recognition methods the recognition process is based on an attempt to find the best interpretation possible by comparing a sequence of observations derived from a word image with ideal models of the words in a lexicon. There are 3 classes of features that could be used to construct the sequence of observations. These 3 classes are characterized by the features that they extract from the word image, i.e. whole word (high level) features, letter (medium level) features and sub-letters (low level) features. High-level (holistic) features such as loops, ascenders, descenders, t strokes and i dots are structural elements of a word, and are thus less affected by the writing style and cursive variability. These features can also be sub-classified according to size, location or

+ HMMs make use of significant segmentation, where each transition is associated with an observation symbol that has a semantic meaning of a certain fraction of the respective character.

orientation. Different algorithms have been developed for comparisons between a pair of observational sequences. The most common methods are (i) a minimum edit-distance calculation based on dynamic programming using low [86,87], medium [88,89] and high level [32,90,91] features, and (ii) resemblance estimation by HMMs for low [92,93,94,95] and high [96,97] level features. The strong theoretical framework provided by HMMs make it the most frequently used technique in the recognition task. However, the recognition performance is more affected by the feature set involved than the implementation methods used. At the moment, there is no single class of feature set that can be considered optimal. High-level features cannot distinguish among similarly structured words whilst low level features are sensitive to noise and writing style.

In perception-oriented recognition methods the recognition process is based on an attempt to model the human reading scheme. When using this method, a bottom-up manner is adopted to identifying letters anywhere in the observation sequence derived from a word image. This is similar to a human reader trying to match consecutive segments with possible characters without searching a word image from left to right. Humans usually seek for the most reliable characters that can be recognized and then match the remaining gaps with candidates from the lexicon that agree with the preliminary recognized characters [98]. Because of its similarity to the human reading scheme, this method is often referred to as a human-like reading system. Several methods have been developed using this recognition approach based on alignment of letter prototypes [99], cyclic bottom-up/top down neighbouring features excitation [100,101,102,103], word superiority effects [104], and segment aggregation [91]. Because of its nature, this method is more robust and independent of segmentation issues since the recognition of all letters in the word is not necessary. However, the method is based more on a discrimination between the words in a lexicon than on their recognition. Thus, perception oriented recognition is only really suitable for applications that involve a small (20-30 words) static lexicon [74].

2.3.2 CSR Applications and Performance

Despite the generally lower performance of CSR, as compared to other recognition technologies such as OCR and barcode recognition, many applications based on CSR have been successfully implemented, in particular bank check reading and postal applications. Unfortunately, it is almost impossible to compare the recognition performance achieved by each of these different systems due to the following reasons. Firstly, many methods use proprietary databases or are tested on relatively small lexicons. Secondly, the recognition performance of a system relies on many factors such as pre-processing, post-processing, segmentation etc. The chosen lexicons and recognition methods used also inevitably affect the final recognition performance. For instance, a segmentation-free recognition system using holistic (high-level) features will perform better with lower-case or mixed-case lexicons than with an upper-case word lexicon due to the fact that upper case words have less features (no ascender or descender) than lower and mixed case words. Finally, some of the methods have not been fully investigated or are still under development. Thus comparing such preliminary results to others in the literature would be misleading. The availability of public domain data such as NIST [105] and visual toolkit [106] is part of the community's effort to produce a convenient platform for comparing recognition performances among researchers. However, until a complete data set that encompasses all of the different application and research requirements is available, many researchers will continue to use their own data making the task of comparison almost impossible. That said, whilst we cannot directly compare the performances of the recognition systems at the moment, the reported results can be viewed as an indicator of the state of the art of the recognition system performance under specific application or conditions. Papers that attempt to comprehensively compare recognition system performance can be found in [74,107,108].

One of the most important commercial applications for off-line CSR is the machine reading of bank cheques. The main reason for this is that there are millions of cheques passing through the banking system each day. Even if a recognition system is only able to confidently verify half of the cheques requiring processing, this would save much labour being expended on a tedious and often unpleasant job. There are several key factors that allow for CSR success in a check reading application. Firstly, the lexicon size is small - the reading of legal amounts typically consists of only 25 to 30 words. Secondly, the presence of a courtesy amount written in digits. By reading both the courtesy and legal amount, the recognition system has a better chance of recognizing the words correctly - courtesy recognition is easier than legal amount recognition. As described in [109], the developed CSR methods are good enough to be used in commercial products, and a family of systems that can work on French, English and American cheques has been developed that claim to have a performance close to that of a human reader albeit with a rejection rate of 30-40%.

Table 2-2 shows a summary of some of the reported results in the literature for bank cheque reading. As mentioned earlier, no direct comparison can be made between each of these systems but the best recognition performance that has been achieved so far is from Dimauro [114]. This result is based on 300dpi resolution images with well-separated digit and character data sets. In [109], the human like performance of 99% accuracy can only be achieved by using manual human intervention to discard the ambiguous samples. From this, it is clear that the problems involved in cheque reading are still far from being fully resolved. Until a more robust CSR system can be produced, research into cheque reading methods will continue be the focus for many researchers.

Task	**Data size**	**Performance**	**Conditions**
Courtesy Amount [110]	10,000 cheques	60%	300dpi, 26% confusion rate
Courtesy Amount [111]	500 cheques	85%	unknown rejection rate
Courtesy Amount [112]	3374 cheques	74%	5.52% rejection
Courtesy Amount [113]	503 amounts	71%	With Verifier and Post Processor
Both [114] Courtesy	1500 samples	98%	300dpi, well separated digit
Legal	1000 words	95%	Italian basic words in check
Whole field		80%	Whole worded amounts
Both [115]	N/A	80%	3.5% misread, commercial system
Month Word Only [116]	402 words	91%	12 lexicons
Legal Amount [117]	500 words	60%	12% confidence, 0.2% rejection
Legal Amount [118]	2515 words	72%	32 lexicons
Legal Amount [119]	N/A	97%	76% rejection
Legal Amount [120]	1083 words	44%	2-10 lexicons
Legal Amount [121]	2378 words	80%	26 lexicons, 6.9% rejection

Table 2-2. Summary of the overall recognition results that have been achieved in check reading application.

The next most popular CSR application is postal processing. Again, like the cheque reading systems, even a recognition system that is able to confidently read only half of the envelope address/zip codes, would save much labour time and cost. Mail sorting can be seen as an ideal application for CSR since it is tolerant to relatively large amounts of error and allows a large rejection rate. Many mail reading/sorting systems have already been installed in many post offices around the world. By using the machine to locate and read the postcode on an envelope, many of the mail pieces can be directed automatically. Table 2-3 shows a summary of some of the reported results in the literature.

Application	Data size	Performance	Conditions
Address Reader [122]	451 address	50.6%	50% confidence threshold
Address Reader [123]	908 samples	90.6%	Manual word extraction
Address Reader [124]	450 samples	71%	0.7% error rate, 28.5% rejection
Zip Code Reader [125]	930 zip codes	49.8%	49.3% rejection
City Name Reader [126]	2500 mail	61.7%	1.3% error rate

Table 2-3. Summary of the overall recognition results that have been achieved in postal reading application.

As expected, the recognition results are not as good as that in the Cheque reading application. This is due to the fact that address reading requires a far more sophisticated CSR system that has to cope with larger lexicon sizes and a mixture of words and digits.

With successes in cheque and postal applications, it is no surprise that the next target application for CSR researchers is form processing. Several form processing applications have already been proposed and developed. For example, the tax form reader developed at CEDAR [127] (modified from a postal address reading machine), invoice-processing system proposed by Bayer *et al* and Kosiba *et al* [128,129], flight coupons processing developed at IBM [130], USA Census form processing system [36] and credit card slip processing system proposed by Paik *et al* [131]. Some of these systems deal with machine printed text only, while others deal with specific fields such as name, address and hand printed digits. Many of the form processing systems are still in their preliminary stage and many issues have yet to be fully resolved. Generally though, when recognition of handwritten data is the primary concern, the recognition performance is poor. For example, the average recognition rate reported on the tax form reader in [127] is only 36.6%. While waiting for the ultimate breakthrough in CSR performance, most of the form-processing packages in the market today use a degree of human intervention in several stages of the processes (for instance form registration and data verification) to increase the overall performance. It is obvious then that if a fully automated system is ever to be achieved, a more robust CSR method is needed so that this human intervention can be eliminated.

2.3.3 Recognition Improvement – Future Direction

There are many ways that have been proposed to improve the CSR performance. These range from front end (digitization, noise removal, segmentation, skew/slant detection and features extraction) to back end processing (recognition methods, use of style, linguistic and other knowledge).

Table 2-4 summarizes some of the strategies and methods that have been published in an attempt to help improve the recognition performance. All of the reported techniques have been proven to be able to improve the overall system recognition performance but again, due to the nature of their testing environment, a direct comparison between them is impossible. Some of these ideas and concepts are application specific; for example trigraphs, linguistic and semantic strategies [139,140,146,59] are more suitable to handwritten page or sentence recognitions environment than form or cheque processing applications. There are also outstanding issues that need to be resolved before these concepts can be applied in a real life application.

	Strategies to improve the system performance
Pre-processing	Word length estimation [132,133]
Pre-processing	Segmentation enhancement [134]
Pre-processing	Word quality enhancement (tilt & slant correction) [135]
Recognition	Used of position, contour and bending point features [136]
Recognition	New normalization & segmentation strategy using HMM [137]
Recognition	Multiple recognizers combination [138]
Recognition	Used of trigraphs in context dependent recognition [139,140]
Recognition	Combination of word level and feature level recognition [141]
Recognition/ Post-processing	Lexicon reduction using recognizered characters [142] and recognized features [143,144]
Post-processing	Used of semantic information from corpus [145]
Post-processing	Used of linguistic information incorporating recognizer's confidence and word sequence [146]

Table 2-4. Summary of the techniques that have been proposed to help improve the recognition performance.

A specific example of using high-level knowledge to help CSR in a form-processing environment is the use of contextual information from the template [14]. In this case, contextual information is manually associated to the filled-in word in a form processing system during the form description process. Whilst this contextual information does provide extra information to the recognition system to help improve the recognition performance - either limiting the search domain of the recognizer or validating the recognition results - the manual mark-up method is somewhat tedious if a large number of different forms are to be processed by the system. Thus, a method needs to be developed that can automatically link this contextual information to the filled-in word.

2.4 Summary

This chapter provides an overview of existing form extraction system techniques together with a discussion of the colour handling techniques and the CSR methods & applications presented in the literature. In the first section, the two basic approaches used to extract data from a filled form are presented. The first approach uses a template, which can be created manually or automatically, to facilitate the extraction process. The second approach uses the techniques developed in DLA (Document Layout Analysis) to remove the form frames without using a template. The most commonly used features to represent a form structure are lines. Most of the systems reviewed were able to identify and extract the line components with a very high degree of confidence. A lot of systems also employ a fairly reliable text-repairing algorithm that can restore most of the damaged text that results from the line removal process. Recent efforts have concentrated on reducing the system complexities and increasing the system efficiency (speed and extraction rate). Colour has been shown as one of the potential features that could be used to further enhance the system performance. However, section 2.2 has also shown that current colour handling techniques are not efficient enough to handle colour document images. Many techniques developed in the past reduce the number of colours in an image to a given number of colour domains and then search through each domain for possible text or characters before converting the colour image into a black and white image. Consequently, the form colour information is lost after the text searching process. This limits the potential for using the colour information in the subsequent processing stages.

The third section introduces the 3 general methods used in CSR – segmentation based, non-segmentation based and perception oriented. The implementation techniques were briefly discussed, followed by examples of real life application with their respective performances. Despite a low recognition rate, CSR has been successfully applied in both cheque and postal applications. Thus, focus has now moved on to form processing and improving the CSR performance. Currently, many of the processes involved in form processing require human intervention – i.e. template creation and recognized data verification. In order to process form documents efficiently and fast, human intervention has to be minimized. The desire to produce a fully automated form processing system has therefore motivated many researchers to develop more reliable CSR and data extraction techniques, such as the work presented in this thesis.

3. COLOUR BASED FORM EXTRACTION

The concept of using colour to improve the performance of an automatic form processing system is not new. For instance, **non-read inks*** are extensively used in the form design & printing industry to help improve form capturing efficiency by allowing the automatic removal of pre-printed entities from the filled form. However, aside from taking advantage of the 'colour-blind' behavior of the scanner, **colour knowledge** has never been used in any of the post scanning form-processing techniques.

This chapter introduces the concept of using colour information to improve the extraction of handwriting data from forms. Section 3.1 describes a new colour handling technique that can successfully reduce the colour content of form images using a direct comparison of the pixels' RGB value. The method forces the form image colour content to less than or equal to eight colours – a balance set between reasonable computation overheads and the need to provide sufficient information to aid the form extraction process. Section 3.2 presents an overview of a new colour based form extraction system that incorporates colour information to produce a better and more efficient extraction system. Section 3.3.1 then presents the results of an experiment investigating the effect of using such a colour reduction technique on the machine printed text quality. This is followed by another two experiments that quantify the performance gain obtained by using such a method in terms of extraction efficiency (section 3.3.2) and extracted text quality (section 3.3.3).

***non-read inks** are the colours that are invisible to the scanner device. Different light sources used in each scanner will have different non-read inks and colour responses [refer section 2.2.1].

3.1 A New Colour Handling Technique

Unlike other types of document, a form contains relatively few colours. This is due to the nature of the form function. As a form is used to capture data from people, the area of interest (filled-in data area) often contains simple or single plain colours as the background. Empirical results show that the majority of forms in use today can be adequately represented using just a few colours from a group of 8 pre-defined colours.

As a scanner digitizes colour documents using an RGB model, methods that work directly on the RGB model will eliminate the extra processes needed to convert this colour model to other models. Processing time is crucial in any automated form processing system, thus it is important to process the colour image in its original colour model in order to minimize the extra burden imposed on the system as a result of the inclusion of colour into the process.

In the RGB colour model, colour is represented by the amount of red, green and blue components. A full colour image uses 8 bits per colour component to produce 16 million possible variations of colour in an image. However, as mentioned earlier, form images need only a few of these colours. Therefore, an aggressive colour reduction method is proposed that can reduce the total number of colour variation in an image from 8 bits per colour component to 1 bit per colour component. This not only reduces the colour variation of an image from 16 million colours to just eight colours but also reduces the memory consumption of the image and makes the image much easier to manipulate than the full colour original. The only draw back for using such an aggressive colour reduction technique is when the original form document contains more than 8 colours; then colour information will be lost in the colour reduction process. However in practice, this rarely happens and even when it does, the 8 colour domains still provide more information than the conventional black and white alternatives. Fig. 3-1 shows how these eight colours are formed in the RGB colour model.

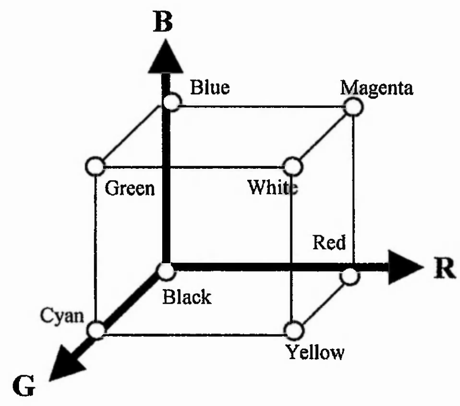


Fig. 3-1. In the RGB colour model, each colour is represented by specifying the amount of red, green and blue components. In an 8-bit per component system, these values will range from 0 to 255.

In an RGB colour model, a pixel's colour can be determined by comparing the RGB component values. If all these RGB values are close to each other in value, then the pixel will appear as a black, gray or white colour. When one or two of the component values are significantly higher than the other component value(s), a pixel's colour will be more prominent and distinct. For example, a pixel will appear as red colour when the R component value is significantly higher than its G & B component values. Similarly, when the R & G component values are significantly higher than the B component value and the R & G values are close to each other, then the pixel will appear as a yellow colour. Fig. 3-2 illustrates how the eight pre-defined colours can be formed by comparing the RGB values of a pixel.

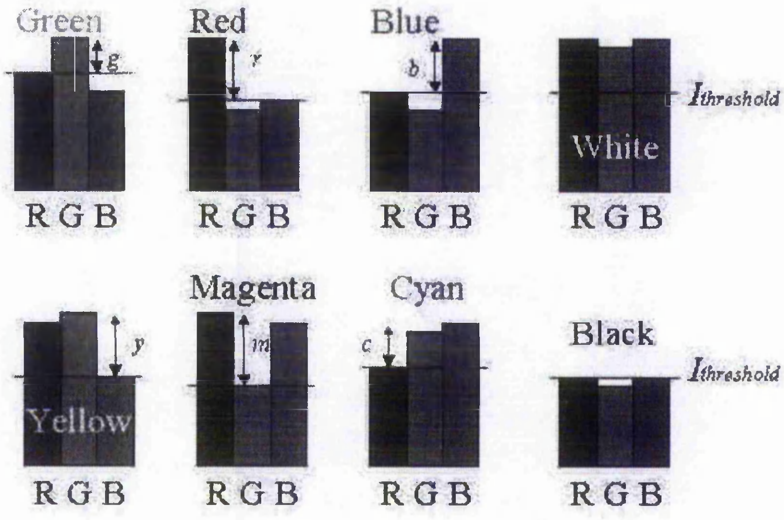


Fig 3-2. A pixel's colour can be determined by comparing the R, G & B component values within that pixel.

A pixel's colour can thus be determined by using the following expressions:

GREEN when $G>R+g$ and $G>B+g$

RED when $R>G+r$ and $R>B+r$

BLUE when $B>R+b$ and $B>G+b$

YELLOW when $R>B+y$ and $G>B+y$

MAGENTA when $R>G+m$ and $B>G+m$

CYAN when $G>R+c$ and $B>R+c$

<u>If the pixel's RGB values do not meet any of the above conditions</u>

BLACK when $(R+G+B)/3 < I_{threshold}$

WHITE when $(R+G+B)/3 > I_{threshold}$

Depending on the brightness characteristic of the scanner, the $I_{threshold}$ value that determines a pixel's colour as either a black or white colour may range from 170 to 220. The r, g, b, y, m and c values also vary according to the colour characteristic of the scanner [see Fig. 2-3, section 2.2.1]. For a scanner with an operating light source wavelength of 700-800 millimicrons, the scanner will be more sensitive to the white, yellow and red colours compared to the blue and green colours. Thus a red colour will be digitized to a pixel that has an R value that is significantly higher than its G & B component values when compared to another scanner that employs a 500-600 millimicrons light source. This scanned image thus requires a different (higher) r value compared to images scanned using the later light source. Such a system will also need significantly smaller g, c & b values than the r, y & m values. Empirical results[+] show that the r, y & m values may range between 35-50 and that the g, c & b values may range between 10-25. These values were determined by repeatedly quantizing several different kinds of colour images using different threshold values; starting from 0 and increasing to a point where the quantization output fails to quantize the image colour correctly. For example, with a digitized image that contains only red and black colours, the acceptable threshold range for the r-value will be those r-values that allow all the distinguishable red colour pixels in that image to be quantized into red colour pixels. The test images in this experiment were carefully selected (each contains 2-3 colours) so that the colour content in these images covers all of the pre-defined eight colours and represents all of the most commonly found colours in form documents. Table 3-1 shows the experimental results for this test and a summary of the ranges of threshold values that produce the correct quantization results.

	Threshold range
r	35-50
g	10-15
b	15-25
y	38-50
m	35-50
c	15-20
$I_{threshold}$	165-215

Table 3-1. Depending on the brightness and colour characteristics of the scanner, different threshold values are needed to quantize the colour images correctly. These are the workable threshold ranges found using 3 different scanner models.

[+] Based on experiment performed on several different colour images that contained different colour content, digitized with 3 different scanners (Hewlett Packard 5200C, Hewlett Packard 3690C and Epson GT 6000)

Since it is not possible to determine the threshold ranges for all of the colour variations found in all documents, these thresholds will fail at some point. However, within the bounds set by this experiment, it is shown that the quantization method will perform its job correctly using the given range of threshold values shown in table 3-1. Therefore, by choosing the mid-value of these ranges, the quantization method will perform its task correctly as long as a form does not contain a very large colour variation. The r, g, b, y, m, c and $I_{threshold}$ values chosen for this work are thus 43, 13, 20, 44, 43, 18 and 190 respectively.

Besides the scanner brightness and colour characteristic considerations, the scanning resolution must be taken into account as well, since it plays an important role in a pixel's colour generation. The effect of scanning resolution on the colour pixel production is illustrated in fig. 3-3. Note that as the resolution reduces, an image starts to lose the detail of the original picture (in this case the black line) and, at 100dpi, a black line is merely represented by a row of dark yellow pixels.

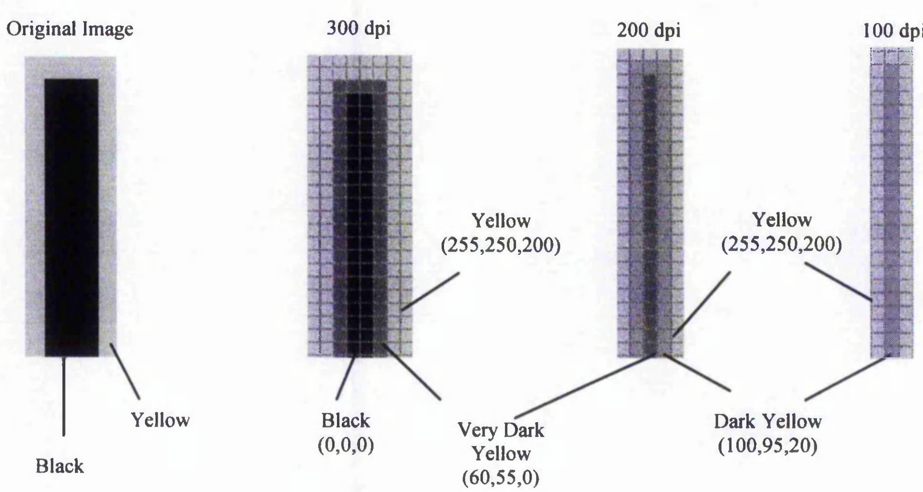


Fig. 3-3. The effect of different scanning resolutions on the generated pixels' RGB values.

Thus, if the RGB comparison method is used alone to quantise colour images, then, at 100dpi, the dark yellow line would be quantised to a yellow colour causing information to be lost in the quantised image. To solve this problem, a maximum reference value is first determined for each of the colours (up to maximum of eight as pre-defined earlier) present in an image. This can be done by applying a first pass pixel scan before the quantization process is carried out. In this process, each of the pixels is provisionally labeled as one of the eight possible colours based on the expressions mentioned earlier. The highest relevant value of each of these obtained colours can then be determined. For example, the red colour reference value will be the highest R-value $(R)_{max}$ found for a red-labeled pixel within the image, whilst for yellow it will be the highest mean value of the R and G components $((R+G)/2)_{max}$. The only exception is for the black colour reference value, which uses the minimum average value of the R,G and B components found in the image.

In a second pass, each of the pixels' colours is then decided by comparing its RGB values to their respective colour reference values using a distance threshold value D*. Take fig. 3-3 for example, by using the comparison technique on the 300dpi image, two colours will be identified (Yellow & Black - with a reference value of 252.5 (from 255,250,200) & 0 (from 0,0,0) respectively). Using these reference values, each of the pixels in the image will then be quantized into either a black or yellow colour depending on the differences between the pixel and reference RGB values. For instance, a very dark yellow pixel with an RGB value of 60,55,0 will be quantized to black as the difference between the very dark yellow pixels R & G values ((R+G)/2=(60+55)/2=57.5) and the yellow reference value ((255+250)/2=252.5) is greater than the threshold.

*By experiment, this D value is found to be in the range of 50-120 in order to obtain a correctly quantized output. The value chosen for this work is the mid-value of this range, i.e. 85

Similarly, for the 100dpi image in fig. 3-3, the dark yellow pixels will be quantized to a black colour while all other bright yellow pixels will be quantized to a yellow colour. Fig. 3-4 shows an example of the quantization result using such an approach.

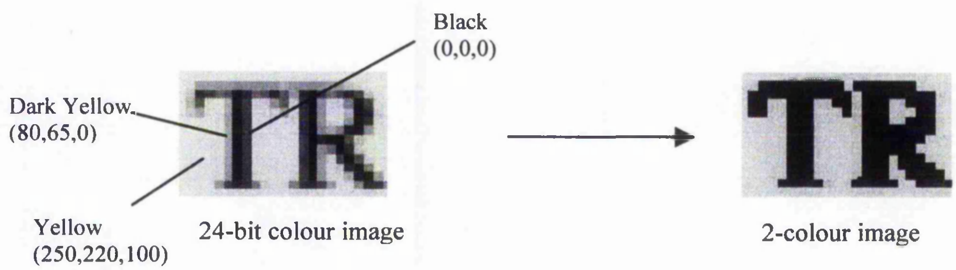


Fig. 3-4. Example of how dark yellow pixels are quantised to black colour when using the value referencing method on a 200dpi image

This two pass method does successfully cope with the problems of low resolution scanning, however, as processing time is important in form processing automation, the use of a two pass pixel may seem computationally expensive. To overcome this, the algorithm is improved by taking into account the colour visibility characteristic of the human vision system. As explained in [147], a human being is only able to distinguish colour when there is a sufficient amount of light (luminance) present. Thus, if a pixel's R, G and B component values are very low, then the colour is invisible to human vision and can thus be considered as a black colour. For instance, even when a red pixel's RGB values are 80,0,0, it still appears as black colour to human eye. Therefore, to speed up the quantization process, when a pixel's RGB values are very low (empirically found to be R & G & B < 90), it is safe to quantize these low value pixels to a black colour without applying the above-mentioned RGB comparison method. This effectively reduces the number of pixels that are required to be compared to the reference values thereby increasing the system performance.

3.2 Colour Based Extraction System

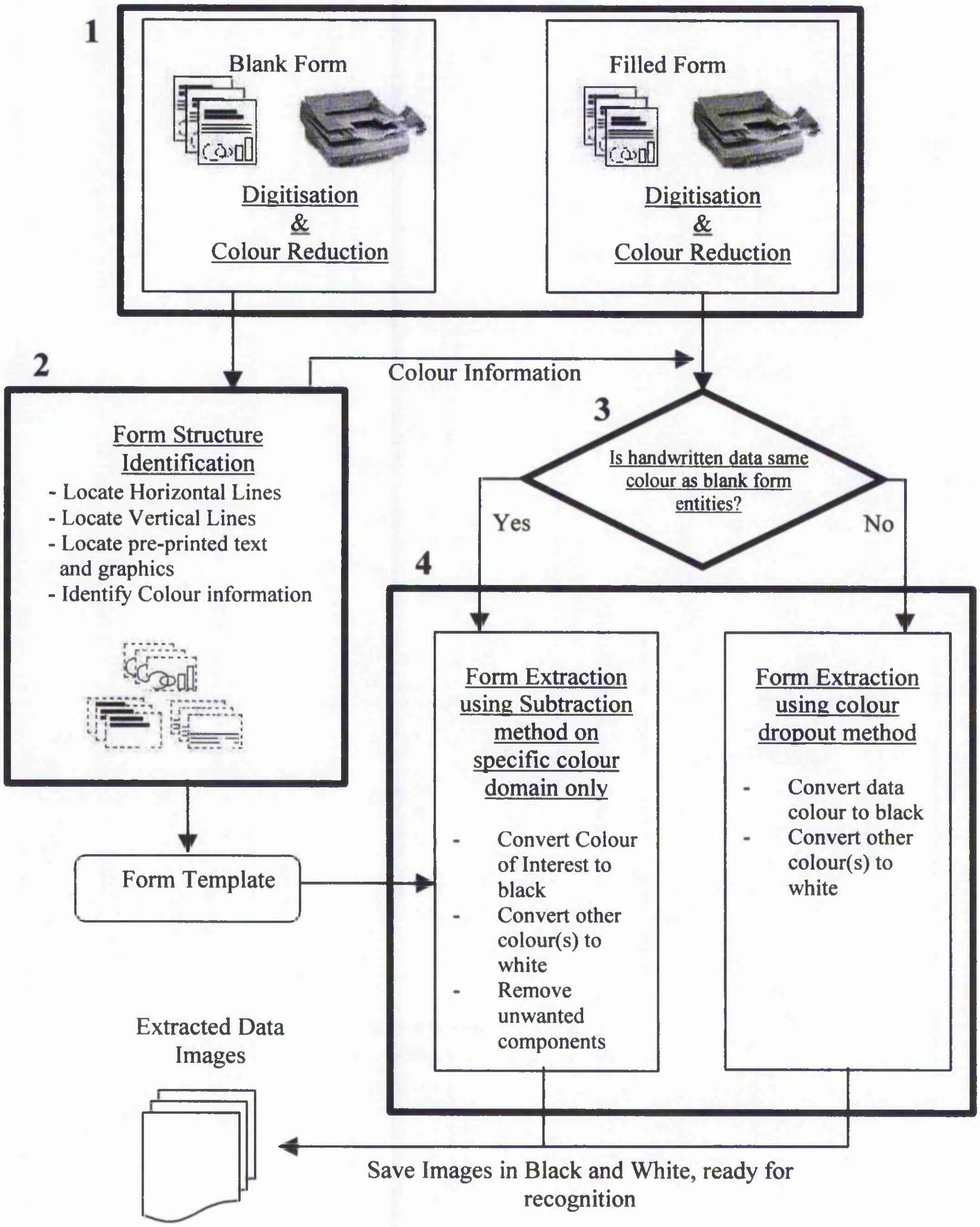
A new hybrid approach for extracting the filled-in data from form images is proposed. This novel method combines a colour dropout approach & subtraction techniques with the incorporation of colour information to provide a better form extraction solution.

Some assumptions have been made in this study to ensure that the focus of the experiments is on the use of colour to extract data from form images:

- The skew angle of the scanned images is assumed to be less than 3°
- Blank (unfilled) forms are available to the system
- Skew marks are not available

Fig. 3-5 shows the proposed colour-based form extraction system. It consists of 4 major parts:

1. Form digitization and quantization
2. Blank form structure analysis and identification
3. Colour analysis
4. Form extraction



1. ***Form Digitization and Quantization***
2. ***Blank Form Structure Analysis & Identification***
3. ***Colour Analysis***
4. ***Form Extraction***

Fig. 3-5. Block diagram of the proposed colour-based form extraction system.

3.2.1 Form digitization & Quantization

The process flow of the system is as follow: first, a blank form is scanned and quantized with the previously mentioned method. This effectively segments the form into several colour domains (maximum 8 as pre-defined in section 3.1). Vertical/Horizontal lines as well as pre-printed text and graphics components in each of the colour domains are then located.

3.2.2 Blank Form Structure Analysis and Identification

As the skew angle of the form images is assumed to be very small, a run length pixel count method can be employed to locate any possible lines within the blank form. In a black and white image, a line can be defined as a long series of black pixels connected to each other in a row. Similarly, in a colour image, lines can be located in the same way as in bi-level images by examining each of the colour domains individually. However, in both cases, this method will fail to locate broken, dotted or dashed lines. Therefore, a Run-Length Smooth Algorithm (RLSA) [148] is needed to join the dotted lines together before the pixel count method is applied. For the purpose of defining line connectivity in this study, a line is assumed to have a maximum of 5 pixels gap between two line segments. This is based on the assumptions that dotted, dashed and broken lines are normally very close to each other and when scanned at 200dpi, should produce a gap of less than 5 pixels width. Thus, when locating horizontal lines, a horizontal RLSA with a constant C value of 5 is applied throughout the image.

There are 3 possible types of lines in a form image – long run length lines that are used to form tables, boxes and fill-in spaces, short run length lines that are used to underline words or sentences and very short run length lines that are used as dashes or minus signs in a sentence. By using a run length pixel count method, long lines are easier to identify than short lines. This is due to the fact that text will often appear as short lines after the RLSA process. Hence, to avoid wrong classification, only lines that are obviously longer than 1 or 2 words are considered as targeted lines. In a 200dpi image, this is equivalent to a run length of approximately 100 pixels. Therefore, when there is a long series of black pixels connected to each other in a row for more than 100 pixels, this portion of pixels is retained as a potential line.

A line verification process is then applied to all these identified long run pixels using a contour tracing method [149]. Using this method, all the connected pixels are grouped together and enclosed in a bounding rectangle. As a line is always just a few pixels thick and is usually much thinner than text or other entities, a line can be ascertained by examining the height of this bounding rectangle (a typical horizontal line in a 200dpi image is not more than 4 pixels height). Figures 3-6 (a-d) shows an example of a blank form image with its identified lines and its corresponding output image after the line verification process. Once all the horizontal lines have been located and removed from the image, a vertical RLSA with a constant C of 5 is then applied across the image again followed by a run length pixel count in the vertical direction (assumed vertical lines' gap are the same as horizontal lines). This process is necessary because when the horizontal lines are removed from the image, some of the vertical lines become broken. By applying the RLSA before the vertical pixel count process, all the broken vertical lines are re-connected back together again. The identified vertical lines are then also removed from the image using the same approach as described for horizontal lines.

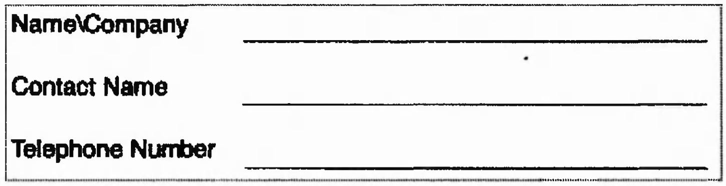


Fig. 3-6(a). A blank form image

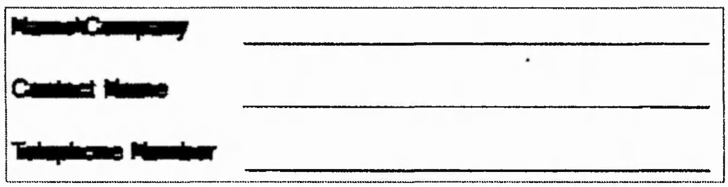


Fig. 3-6(b). Image after RLSA (C=5) process

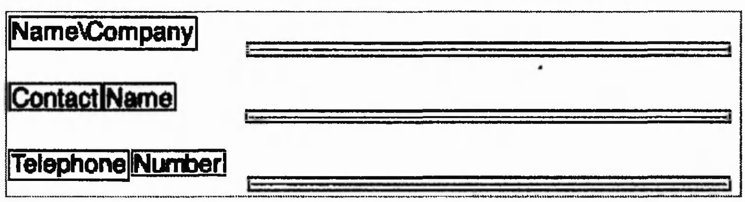


Fig. 3-6(c). Image after contour tracing method applied

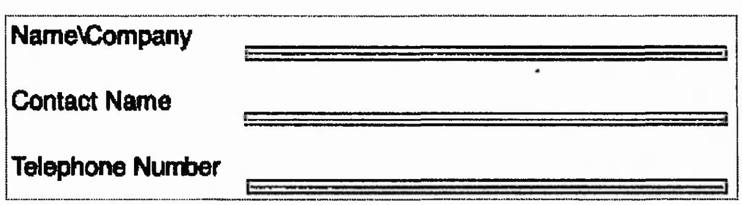
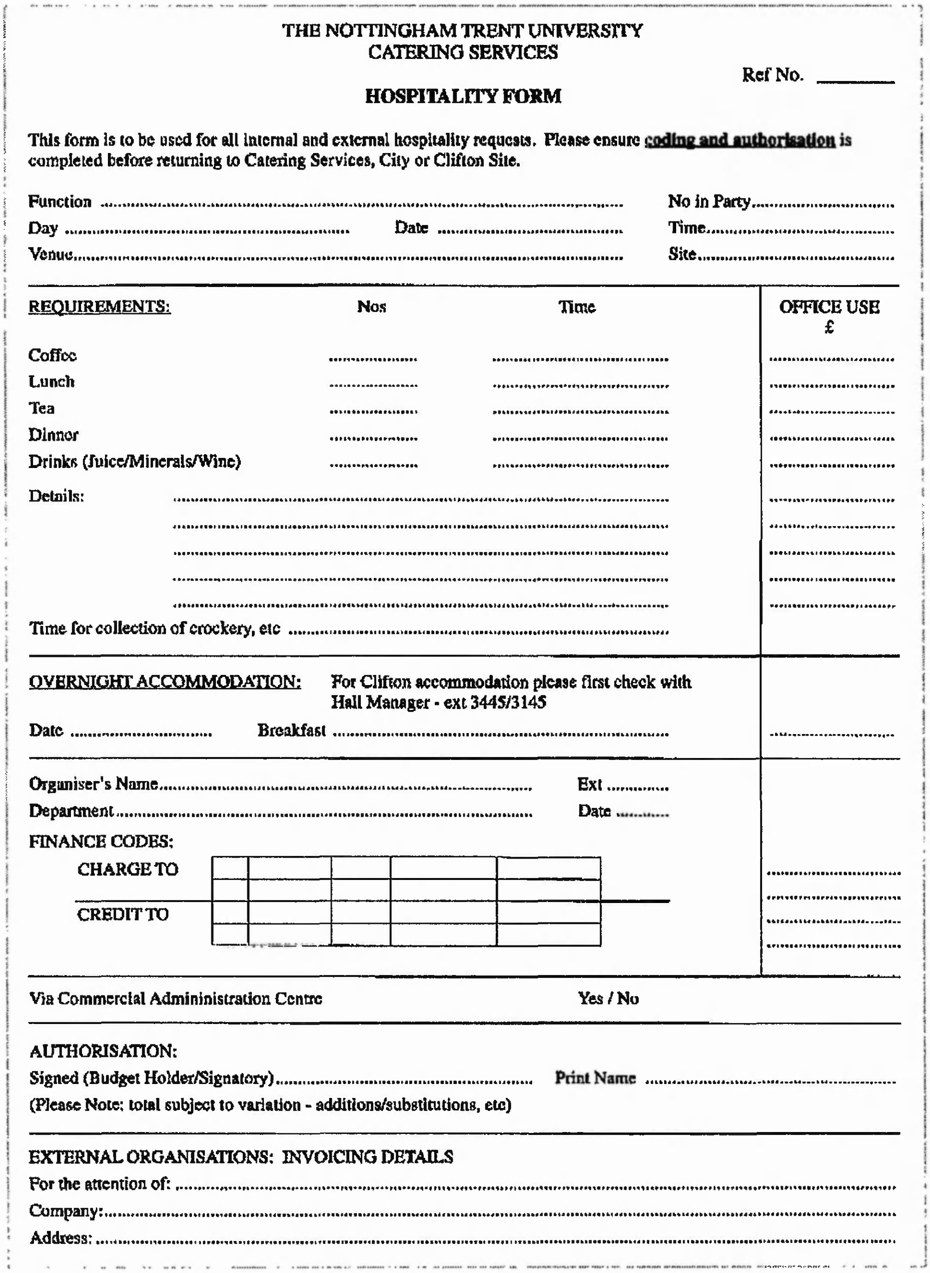


Fig. 3-6(d). Lines identified after the verification process

Once all the true vertical and horizontal lines have been located and removed from the image, the remaining connected groups then represent the text and graphic components for that form. These groups of rectangular bounding box locations, together with the line bounding boxes are then recorded as a template for the form removal process. Fig. 3-7(b) to fig. 3-7(d) show the identified pre-printed entities (vertical lines, horizontal lines and other pre-printed entities) that have been processed by this proposed form structure process on a form image shown in fig. 3-7(a).


THE NOTTINGHAM TRENT UNIVERSITY
CATERING SERVICES

Ref No. ________

HOSPITALITY FORM

This form is to be used for all internal and external hospitality requests. Please ensure coding and authorisation is completed before returning to Catering Services, City or Clifton Site.

Function .... No in Party ....
Day .... Date .... Time ....
Venue .... Site ....

REQUIREMENTS:	Nos	Time	OFFICE USE £
Coffee	....	....	....
Lunch	....	....	....
Tea	....	....	....
Dinner	....	....	....
Drinks (Juice/Minerals/Wine)	....	....	....

Details: ....

Time for collection of crockery, etc ....

OVERNIGHT ACCOMMODATION: For Clifton accommodation please first check with Hall Manager - ext 3445/3145
Date .... Breakfast ....

Organiser's Name .... Ext ....
Department .... Date ....

FINANCE CODES:
CHARGE TO
CREDIT TO

Via Commercial Admininistration Centre Yes / No

AUTHORISATION:
Signed (Budget Holder/Signatory) .... Print Name ....
(Please Note: total subject to variation - additions/substitutions, etc)

EXTERNAL ORGANISATIONS: INVOICING DETAILS
For the attention of: ....
Company: ....
Address: ....

Fig. 3-7(a). A typical blank form after the quantization process.

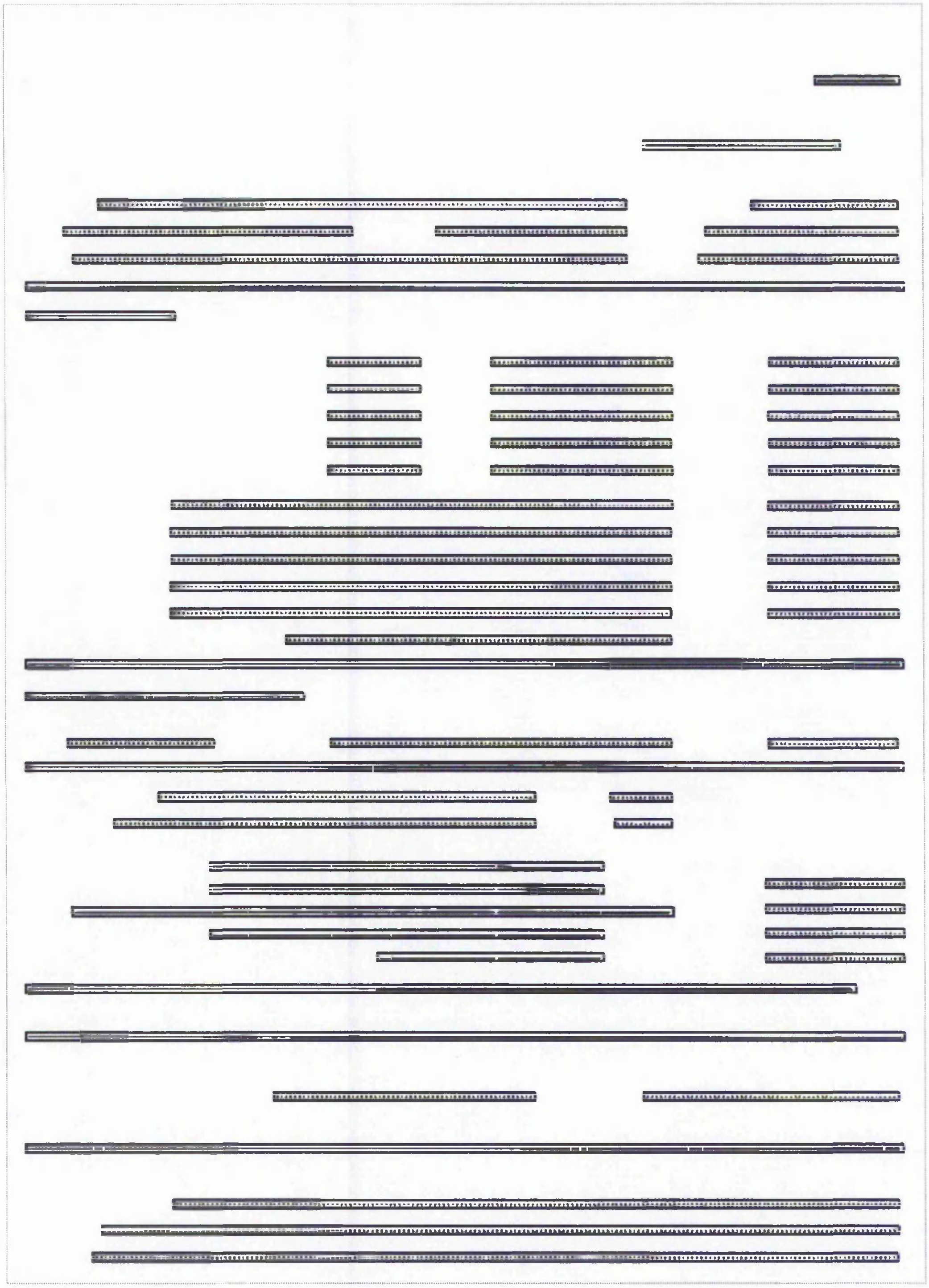


Fig. 3-7(b). The identified horizontal lines for the form image shown in fig. 3-7(a).

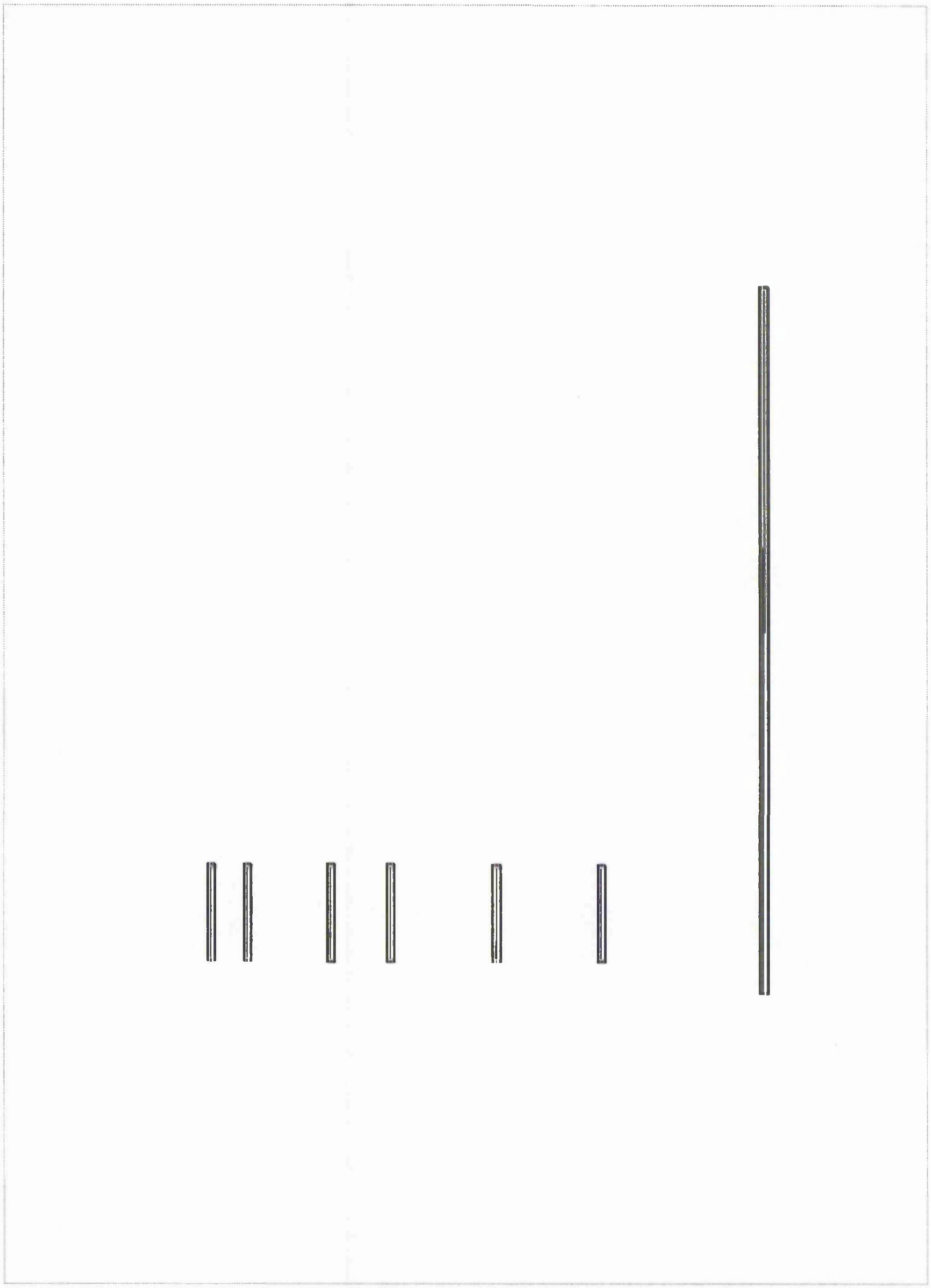
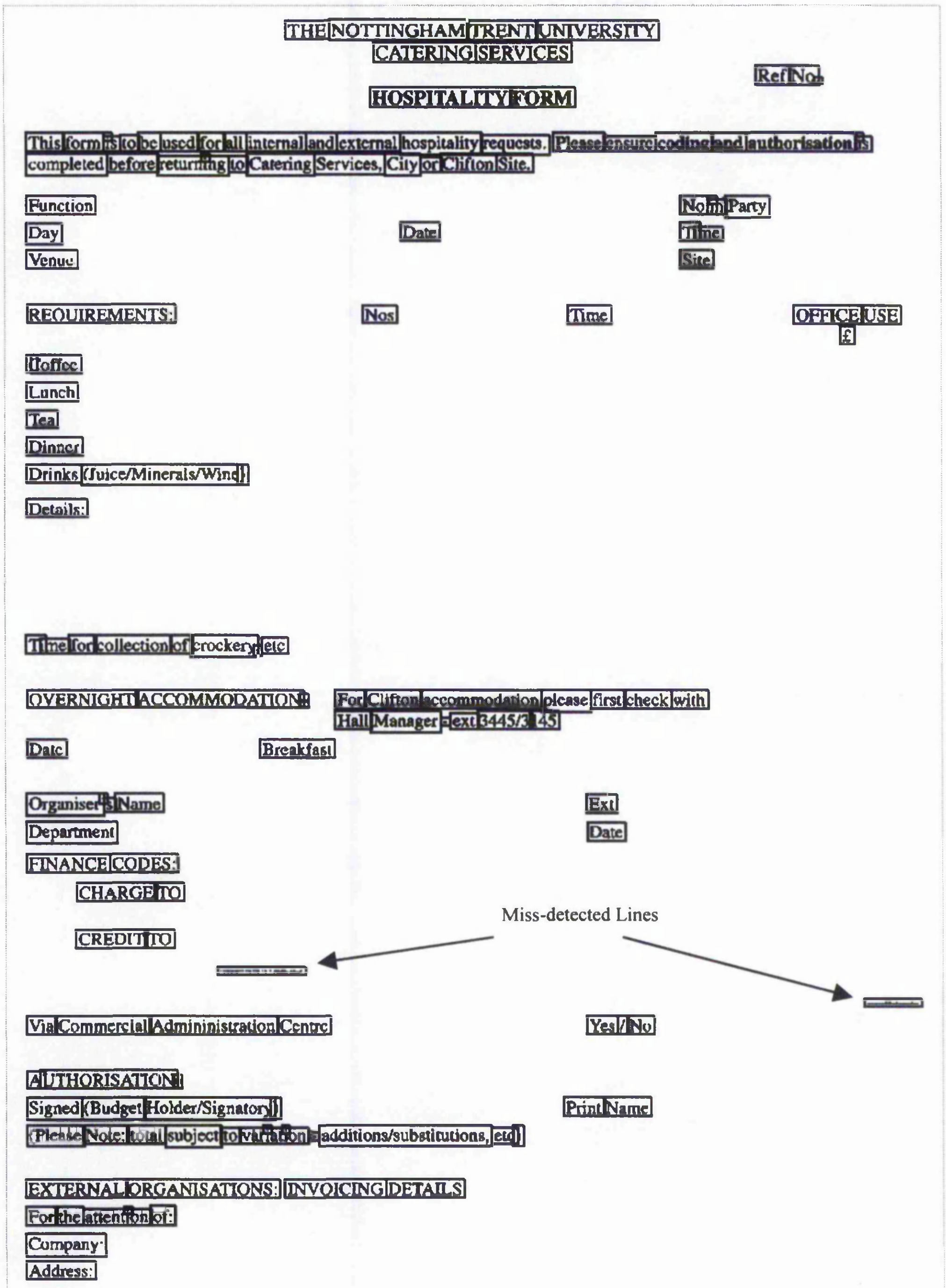


Fig. 3-7(c). The identified Vertical lines for the form image shown in fig. 3-7(a).


THE NOTTINGHAM TRENT UNIVERSITY
CATERING SERVICES

Ref No.

HOSPITALITY FORM

This form is to be used for all internal and external hospitality requests. Please ensure coding and authorisation is completed before returning to Catering Services, City or Clifton Site.

Function | No in Party
Day | Date | Time
Venue | Site

REQUIREMENTS: | Nos | Time | OFFICE USE £

Coffee
Lunch
Tea
Dinner
Drinks (Juice/Minerals/Wine)
Details:

Time for collection of crockery, etc

OVERNIGHT ACCOMMODATION: For Clifton accommodation please first check with Hall Manager - ext 3445/3145

Date | Breakfast

Organiser's Name | Ext
Department | Date
FINANCE CODES:
CHARGE TO
CREDIT TO

Via Commercial Admininistration Centre | Yes / No

AUTHORISATION:
Signed (Budget Holder/Signatory) | Print Name
(Please Note: total subject to variation - additions/substitutions, etc)

EXTERNAL ORGANISATIONS: INVOICING DETAILS
For the attention of:
Company:
Address:

Fig. 3-7(d). The identified pre-printed components (other than lines) for the form image shown in fig. 3-7(a).

Although the line-searching algorithm just described may seem rather too simple to be able to identify all of the possible lines in a document image, the method does provide a quick solution for identifying long lines in non-complex documents such as form documents. It is by no means a robust method for finding lines in complex documents when compared to other methods reported in literature [150,151,152], however, the main advantage of using such a method is its ease of implementation. Besides, the main reason for finding lines in this case is to perform subtraction at a later stage. Thus, even when a line is not identified in the line searching process, it will be treated as one of the other pre-printed entities and will still be subtracted from the filled form at the extraction stage. Hence, the line-searching requirement in this system is less stringent than in other applications and thus the line-searching algorithm described appears as a viable method.

3.2.3 Colour Analysis

After the filled forms have been scanned and quantized, the colour information in the filled form is then compared to the blank form colour information in order to decide whether the colour drop-out or the subtraction technique is to be adopted in order to extract the filled-in data from the image. As the number of colours and their pixel percentages are known from the quantization process, the filled-in data colour can be easily identified by comparing the differences between the blank and filled form. For instance, suppose a 3-colour blank form contained white (87%), black (9%) and red (4%) colours after the quantization process, the filled-in data colour on a completed form can be known instantly by comparing these figures against the quantized filled-in form values. This is shown in table 3-2.

		Filled-in colour
Blank form	White (87%), Black (10%), Red (3%)	-
Filled form #1	White (85%), Black (10%), Red (3%), Green (2%)	Green
Filled form #2	White (85%), Black (12%), Red (3%)	Black
Filled form #3	White (85%), Black (10%), Red (4%), Blue (1%)	Red & Blue

Table 3-2. The percentage changes for each of the colours contained in a form can be used to identify the filled-in data colour.

3.2.4 Form Extraction

When the filled-in data has been entered using a different colour to that present in the blank form, data can be extracted immediately using the colour dropout process. In this process, the filled-in data colour pixels are converted to black whilst all the other colour pixels are converted to white. The resultant black and white image is then ready for recognition. This method effectively eliminates problems such as page skew, page offset and abnormal filled data conditions (such as filled data out of the designated area) seen in the subtraction approach.

When the filled-in data has been entered using one or more of the colours that are used in the blank form, then a subtraction technique must be adopted. However, unlike other form extraction systems, only the filled-in colour domains need to be processed. This eliminates the necessity to process the whole image and hence improves the system efficiency.

As none of the commercially available forms used in this study contained marks that could be used to correlate the template entities locations to their respective filled-form equivalents, a pixel count method is used. This method identifies the first X- and Y-positions that contain more than a given threshold number of pixels, starting from the origin (0,0) position. For example, suppose a blank and filled form image have their first black pixel counts of more than the threshold number of pixels at coordinates 20,10 (x, y) and 18,15 (x, y) respectively, the offset value for these two forms is –2, & 5. The actual threshold value chosen for this work (50) is something of a compromise as the number of pixels count in one image at a particular value of x will usually vary from image to image due to the presence of image scanning noise, skew and offset. If the threshold value is set at a very low value (10 pixels for example), any slight noise added to the image will give a sufficient number of pixels to trigger the detection resulting in an incorrect offset. On the other hand, setting the value too high will impose unnecessary processing time on the form processing system and in some cases cause it to fail to locate the correct offset position due to the number of pixels in any x-position in the image being less than the threshold. Experimental results show that if this threshold is set at between 50 to 100 pixels, these problems can be avoided and a true offset obtained. Using these offset values, the pre-printed entities on the filled form can be removed from the filled form using the position information provided by the blank form template obtained in the form structure identification process.

Although the image-offset problem can be resolved by employing this pixel count method, image skew can still cause problems for the subtraction process. Accurate skew detection and correction can be very computationally expensive and time consuming [153,154] whilst fast skew correction methods generally produce inaccurate and inefficient restorations (de-skew). Nevertheless, most of the current image skew detection and correction techniques can correct the skew angle to within a range of $+/-(1\text{-}3)^{\circ}$ using a relatively fast and coarse approach [155,156,157,158]. Thus, when a subtraction approach is employed in a form extraction system, this skew angle must be taken into consideration otherwise the removal of the pre-printed entities will be incomplete. If we assume that the maximum skew angle is +/- 3°, the correlation position errors (PE) will then be +/- (Tan 3° x W) pixels in the Y-axis direction and +/- (Tan 3° x H) pixels in the X-axis direction. Fig. 3-8 shows this for a form's line entity.

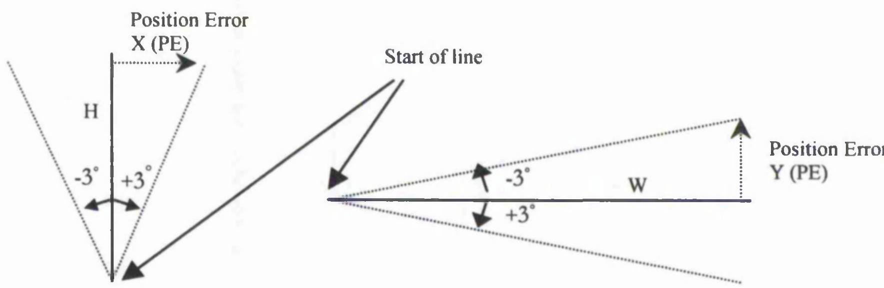


Fig. 3-8. The X & Y position errors (in term of number of pixels) determined by the width (W) and height (H) of the entities and the skew angle.

Therefore, when an entity is subtracted from the image, a margin must be allocated to accommodate these position errors. For example, when removing a pre-printed entity with a width*height of 100*100 pixels and a skew angle of 3°, the subtracted area in the image must be 110*110 pixels (fig. 3-9).

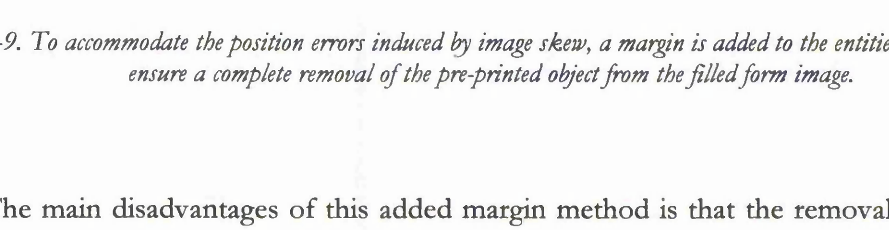


Fig. 3-9. To accommodate the position errors induced by image skew, a margin is added to the entities size in order to ensure a complete removal of the pre-printed object from the filled form image.

The main disadvantages of this added margin method is that the removal process will occasionally remove some portion of the filled-in text, causing degradation in the extracted data quality. This problem mainly occurs in the line removal process as the filled-in data is normally filled in near or even onto the line position. Employing some of the more recently proposed text restoration algorithms (over 96% of accuracy results have been reported) [45,46,51,53,159] can virtually help fully resolve this problem. However, due to the extra processing time these algorithms would incur and the small amount of data that is affected by this factor (less than 5%), these text-repairing algorithms were not implemented. It is believed that by employing a text repairing method in the system there would be a slight increase in recognition results for the black and white extraction system which would reduce the magnitude of the performance gain claimed for the colour extraction system. However, this performance gain is achieved at the expense of extra processing time required and thus the colour system would show a greater improvement in terms of processing speed over the black and white system.

Once the pre-printed entities have been removed from the filled form, a noise removal process is then applied to remove any objects (e.g. noise) that are too small to qualify as text. In this study, the targeted filled-in data is assumed to be handwritten words, hence any connected components that are less than 8 pixels square are considered as noise. This is due to the fact that with typical handwriting words scanned at 200dpi, a handwriting word that is smaller than 8 pixels square is difficult, if not impossible to produce. Unfortunately, this removal process does occasionally remove some portions of some words; for example the dots for 'i' and 'j'. However, as the handwriting recognizer that is used in this work is only looking for word level features such as ascenders, descenders, holes and cups, the removal of small dots is irrelevant and does not affect the recognition performance. The final resultant image after this noise removal process is then saved as a black & white format for the recognition process.

3.3 Experimental Platform

The system has been implemented in C++ programming language under a Windows98 platform on a Pentium Celeron 450Mhz personal computer. Forms were digitized using a Hewlett Packard HP3690C scanner at 200dpi in 24-bit RGB format (except in experiment I where forms were digitized at several different resolutions ranging from 100 to 300 dpi).

The methods developed are designed to work on any type of colour form that is filled-in with unconstrained cursive handwriting data. As there is no commercial CSR package available on the market at the moment, the recognizer that was used in the work is a modified version of a prototype CSR that has been developed within the department of computing at The Nottingham Trent University [17]. This recognizer extracts the vertical bars, loops and cups from the word image and compares these features with a list of words (lexicon) and their pre-defined set of features. Every word in the lexicon is scored according to how well the features match with the target word's set of features. The lexicon words and their scores are then ranked, with the highest ranked word being the likeliest match with the target word. The working resolution for this recognizer is 200x100 dpi, with the capability of accepting off-line (static) word image data as input. It has been designed and optimized to recognize unconstrained, lower case Roman script from any writer, with the assumption that all word images are in bi-level facsimile format. To accommodate the need for this work, the recognizer has been extended to recognize upper, lower and mixed case words.

3.3.1 Experiment I: Colour Reduction and OCR Performance

As the colour content of an image is greatly reduced using the proposed method, a study is needed to access the colour reduction effect on the image quality. One of the parameters that can be used to measure the output image quality is the OCR rate. This experiment aims to evaluate the impact on the OCR performance of using such a colour reduction technique under different resolutions. 15 colour forms of different designs, each containing 2 to 4 colours are used to compute the OCR rate that could be achieved both with and without the quantization process applied. The 15 form samples used for this experiment are shown in Appendix A.

Table 3-3 shows the experimental OCR* results that apply to the proposed colour reduction method output images and the original 24-bit full colour images. In total, there are 7596 machine printed characters for these 15 forms and the OCR rate is computed manually by counting the total number of correctly recognized characters over the total number of characters.

Scanning Resolution	OCR rate (24-bit full colour)	OCR rate (quantized image)	Improvement
100dpi	85.0%	90.4%	+4.6%
150dpi	99.2%	98.8%	-0.4%
200dpi	99.3%	99.3%	0%
250dpi	99.6%	99.6%	0%
300dpi	99.8%	99.8%	0%

Table 3-3. Overall OCR performance tested on 15 forms at various resolutions (total characters=7596).

* The OCR engine used in this work is TextBridge Pro 9.0 from Xerox Imaging System due to its ability to allow OCR at different image resolutions.

The quantized form OCR rate is almost identical to that of the 24-bit full colour image OCR rate above 150dpi resolution with a moderate improvement observed at 100dpi (4.6% increase). These results suggest that although the colour content of the quantized images is reduced to less than 8 colours, there is no significant detrimental effect on the OCR performance at resolutions of 150dpi and above. Indeed, at 100dpi, the OCR performance on the quantized image is greater than that on the original image. This shows that at 100dpi the colour reduction method proposed is better than that employed by the OCR engine. This is believed to be due to the fact that at 100dpi, the image details are smeared and distorted so much so that the original OCR colour handling techniques is not able to recover all of the image details (the engine is probably optimized for 200-300dpi images). Whereas, with the comparison method applied, some of the details can be recovered from the scanning process hence leading to an increase in the OCR rate. The slight decrease in OCR rate at 150dpi is believed to be due to the results of the 'thickening' effect of the proposed colour handling technique. This can be proved by the fact that the increases in miss-recognized characters are mainly with characters such as 'e', 'u' and 'k'. The quantization method will tend to darken the 'hole' for a character e (resulting in an OCR output as a 'c') and join the open end of a character u & k (resulting in an OCR output as o and h respectively). At resolutions higher than 200dpi, the OCR rates become identical and the miss-recognized characters are then mainly due to small font size printed characters which are too small to be recognized by the OCR software.

3.3.2 Experiment II: Extraction Efficiency

Two parameters have been used to measure the efficiency of the extraction system - precision and recall rate. Precision is calculated as the number of correctly extracted objects (connected components) over the total number of objects extracted, whilst recall rate is calculated as the number of objects that are successfully extracted over the total number of expected objects to be extracted from a given form. The recall rate provides a quantitative value in terms of the percentage of expected objects that the system can extract, whilst the precision rate provides a quantitative value on the percentage of the extracted objects that are correct.

Precision= Number of correctly extracted objects / Total number of objects extracted

Recall = Number of correctly extracted objects / Total number of expected objects

For this experiment, another 15 different types of forms were scanned and quantized, each containing a different type of layout design and number of colours. All of these forms were filled-in by a single writer using various types of colour pen. These 15 form samples are shown in Appendix B. To compare the extraction efficiency, a black & white based extraction system was constructed and an intensity-based threshold binarization method was used to convert the 24-bit colour images to black & white images, i.e.

If $I < I_{threshold}$ (then pixel=black)
else if $I >= I_{threshold}$ (pixel=white)
where $I = (R+G+B)/3$

The pixel's intensity is the average value of the pixel's RGB value and the optimum $I_{threshold}$ value for the scanner is 190 (determined in section 3.1). The extraction method adopted on this black & white system is exactly the same as the subtraction technique used in the proposed colour-based extraction system except that instead of working in the filled-in colour domains only, it works with the entire binarized form images.

Form	Total Target objects	Best Case				Worst Case				Conventional			
		Correct	Wrong	Precision	Recall	Correct	Wrong	Precision	Recall	Correct	Wrong	Precision	Recall
1	87	87	0	100.0%	100.0%	85	7	92.4%	97.7%	84	0	100.0%	96.6%
2	73	73	0	100.0%	100.0%	73	6	92.4%	100.0%	71	0	100.0%	97.3%
3	104	104	0	100.0%	100.0%	104	1	99.0%	100.0%	104	2	98.1%	100.0%
4	99	99	0	100.0%	100.0%	99	0	100.0%	100.0%	98	0	100.0%	99.0%
5	209	209	0	100.0%	100.0%	206	0	100.0%	98.6%	194	1	99.5%	92.8%
6	134	134	0	100.0%	100.0%	134	7	95.0%	100.0%	129	0	100.0%	96.3%
7	267	267	0	100.0%	100.0%	267	17	94.0%	100.0%	262	17	93.9%	98.1%
8	229	229	1	99.6%	100.0%	229	0	100.0%	100.0%	229	0	100.0%	100.0%
9	134	134	0	100.0%	100.0%	134	3	97.8%	100.0%	134	2	98.5%	100.0%
10	124	124	0	100.0%	100.0%	123	0	100.0%	99.2%	122	12	91.0%	98.4%
11	158	158	0	100.0%	100.0%	146	0	100.0%	92.4%	118	8	93.7%	74.7%
12	121	121	1	99.2%	100.0%	120	4	96.8%	99.2%	121	1	99.2%	100.0%
13	155	155	0	100.0%	100.0%	155	0	100.0%	100.0%	154	4	97.5%	99.4%
14	130	130	0	100.0%	100.0%	130	40	76.5%	100.0%	130	11	92.2%	100.0%
15	212	212	0	100.0%	100.0%	204	0	100.0%	96.2%	204	0	100.0%	96.2%
Average	149.0	149.1	0.13	99.9%	100.0%	147	5.3	96.6%	98.6%	143.8	4.3	97.2%	96.6%

Table 3-4. Comparison of the proposed colour form-extraction system recall and precision rate under worst and best cases to a black and white form extraction system

Table 3-4 shows the experimental comparison results between the proposed colour extraction method and the black & white extraction method. On average, the colour-based extraction method gives a worst case recall rate of 98.6% and a best case recall rate of 100%. This compares with a 96.6% average recall rate for the black and white extraction method. Thus, there is a clear 2 to 3% recall rate gain for an extraction system that utilizes colour information over a black & white extraction system. However, the precision for the worst case in the colour-based extraction system is lower than that of the black & white system (96.6% compared to 97.2%). This is due to the fact that since the black & white extraction system has a lower recall rate, fewer objects are extracted which results in a lower number of unwanted objects being passed through to the output. When taken together though, we believe the worst case recall and precision rate for a colour based extraction system are similar to those of the black & white extraction system.

	Dimension		Load, Quantise & Extract (Colour-based)		Binarise + extract (sec)
	X	Y	Best Case (sec)	Worst Case (sec)	(Black & White system)
Form01	1320	852	2.14	10.98	8.23
Form02	1532	1095	2.91	5.33	7.58
Form03	1446	732	1.87	7.14	6.65
Form04	1422	736	1.80	6.20	5.50
Form05	1140	811	1.76	9.00	8.07
Form06	1520	1540	3.90	16.73	14.88
Form07	1465	1492	3.73	9.28	12.12
Form08	1422	1525	3.13	10.98	11.70
Form09	1496	1552	3.90	15.92	14.07
Form10	1254	1181	2.69	6.37	5.38
Form11	1614	2280	6.42	26.63	28.73
Form12	1454	2087	5.33	13.19	15.16
Form13	1594	2268	6.49	13.79	15.11
Form14	1492	1130	2.85	6.37	5.06
Form15	1508	1520	3.79	10.12	10.32
Average	1447	1345	3.42	11.01	11.01

Table 3-5. Comparison of processing time required under different filled data colour conditions for the colour-based extraction system and the black and white extraction system

Table 3-5 shows the total processing time required for each of the forms under different colour conditions. As expected, the best case for the colour extraction method is when the filled data colour is of a different colour to that used in the blank form. The worst case is when the filled data colour is of the same colour as the second most dominant colour in the blank form (the most dominant colour for a document usually being the background colour). From the results shown in table 3-5, it can be seen that, in the best case, the proposed colour form extraction method is 3.22 times faster than that of the black & white system. However, when the filled data colour is the same as one of the blank form colours (worst case) then the colour-based system performs with almost identical speed to the black & white extraction system.

3.3.3 Experiment III: Extracted data quality

This experiment aims to assess the extracted data quality by comparing the CSR performance applied to the colour-based and the black & white extracted output images. In this experiment, 3 more different types of form were used each filled by 10 different writers using a colour pen that is a different colour to the colours used in the blank form (blue, black & green). The writers were deliberately instructed to use upper case only as this was found to be the most common and natural style for data entry in form documents. Indeed, when a lower case word restriction was initially imposed, people regularly miss-spelt words when trying to fill-out a form due to the unnatural nature of data entry. The data samples used in this experiment are shown in Appendix C. A holistic cursive word recognizer [32] was then used to assess the CSR performance of the extracted output from the colour-based and black & white extraction methods. This recognizer extracts the word features from the word image and compares those features to the database. It then ranks the words in the database from the highest matched word (with the highest edit distance score) to the lowest matched word.

	Colour extraction method	Black & White extraction method (without text repairing)	Overall Improvement
Form01 (266 words, 132 word lexicon)	58% @top 1	49% @top 1	+9%
Form02 (314 words, 215 word lexicon)	56% @top 1	49% @top 1	+7%
Form03 (256 words, 189 word lexicon)	61% @top 1	50% @top 1	+11%
Overall (836 words)	58% @top 1	49% @top 1	+9%

Table 3-6. The holistic recogniser performance when tested on the extraction output images that are extracted with and without the colour information

Table 3-6 shows the CSR results obtained. As there is no broken text repairing technique implemented in either system and as the recognizer is originally designed for lower case words only, the CSR performance is lower that that reported in [17]. However, the results do demonstrate a clear 9% overall improvement for the colour-based extraction method as compared to the black & white extraction system. This is due to the fact that when colour is not used to extract the data, the line removal process will degrade the text quality, especially when there are a lot of words or characters that are overlapping with the form lines or boxes. The text re-construct methods developed by others would reduce this problem greatly, but at the expense of introducing more computational load into the system. Thus, a black & white extraction system that incorporates text-repairing algorithm could approach the CSR rate of the colour extraction system but at the expense of requiring more processing time.

3.4 Conclusion

A new colour-based form extraction system has been presented, which is shown to be more efficient than a form extraction system that doesn't utilize any colour information. The effect of the colour reduction process on the image quality has been determined. Experimental results suggest that the use of such a colour reduction strategy will improve the form images quality and is shown to work well even with a massive reduction in the total number of colours in an image. The experimental results also demonstrate that at low resolution (100dpi), the colour reduction method is better than that employed by the OCR engine.

The extraction accuracy, recall rate and speed of the colour-based extraction system has also been determined and compared to an extraction system that does not utilize colour information. With the proposed colour reduction method, colour information has been used successfully to reduce the system complexities and computational costs. By adopting colour features in form processing, the extraction speed has been increased up to 3.22 times. The extraction accuracy and recall rate are also shown to improve when using colour information to extract filled data from forms.

The extracted data quality has also been determined by examining the recognition results obtained from a CSR system applied to the output images from both systems. The experimental results suggest that an extraction system that utilizes colour information does produce a better output image quality than that of a black & white system.

This chapter has served to demonstrate the successful usage of colour in an automatic form processing system. The results could probably be further improved if the usage of colour in the forms could be altered or controlled. However, this would limit its applications to an environment where form design is allowed.

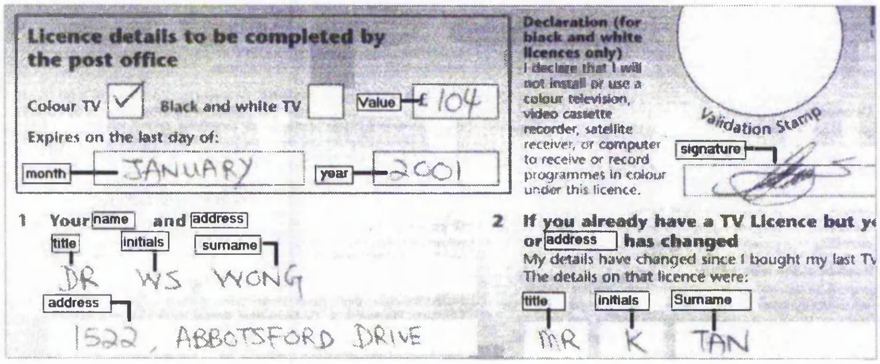
4. CONTEXTUAL FOCUSED RECOGNITION

The advances in OCR technology over the past decades have enabled the development of automatic document-processing systems with OCR accuracies of 99% or greater. Such systems have been used for tasks as diverse as document indexing [160,161], document understanding [162] and Giro and remittance statement processing [163]. Unfortunately, similar advances in CSR have not been forthcoming due, principally, to the vast variability of human handwriting. This chapter investigates a novel method by which the more reliable OCR technology can be used to improve the CSR performance in a form processing application. By taking advantage of the high OCR rate on the machine printed text and the consistency of cue words (instructions or guided text) within a form, the filled data can be automatically 'linked' to the contextual information that these cue words provide. This context can then be used to provide enough information to the CSR system to help produce better recognition results. Section 4.1 describes the idea of using OCR to locate the cue words within a given form. Sections 4.2 and 4.3 then present the character grouping technique to find the filled-in words and a new method for linking these identified words to the contextual cue words. This is followed, in section 4.4, by an experiment demonstrating the effectiveness of the word-finding algorithm that is applied to the extracted data output before the contextual linking method is applied. Sections 4.5, 4.6 and 4.7 then detail three experiments that evaluate the effectiveness and performance gain obtained using the contextual focused recognition technique in terms of cue word retrieval rate, linking efficiency and CSR performance gain.

4.1 Retrieval of Contextual Cue Words From Form Images

Unlike other types of document, a form contains a great deal of useful information in the form of its structural features. A typical filled-in form consists of form frames (including lines and boxes), pre-printed data (logo and machine printed text) and user filled-in data (machine-typed and/or handwritten). These 3 elements are closely related to each other in that the lines and boxes usually signify the filled-in data areas, whilst the machine printed text above or adjacent to the lines and box areas are normally the guiding text that specifies what the filled-in data should be. Thus, if this contextual knowledge can somehow be fed into the data recognition process, it can provide useful information to the recognizers that can then be used to increase the recognition performance.

By knowing the related contextual information of the filled-in data, a recognizer can be directed to reduce its search domain so that only related words are used in recognizing the data. For instance, the contextual cue word 'name' will enable the search domain of a recognizer to be dramatically narrowed by using a 'name' lexicon only. Similarly, different recognizers can be chosen if the nature of the targeted filled-in data is known to the system. For example, a numeric recognizer can be selected for a telephone number data field, a holistic recognizer for any name data fields and a hybrid recognizer for the address or postcode data fields. Fig. 4-1 shows an example of a form with its identified cue words. These are the words that could be used to provide extra information to the subsequent processes in a form processing system.



Licence details to be completed by the post office

Colour TV ✓ Black and white TV Value £ 104

Expires on the last day of:

month JANUARY year 2001

Declaration (for black and white licences only) I declare that I will not install or use a colour television, video cassette recorder, satellite receiver, or computer to receive or record programmes in colour under this licence.

Validation Stamp

signature

1 Your name and address

title initials surname

DR WS WONG

address

1522, ABBOTSFORD DRIVE

2 If you already have a TV Licence but y… or address has changed

My details have changed since I bought my last TV… The details on that licence were:

title initials Surname

MR K TAN

Fig. 4-1. An example of a form with the possible cue words identified that could be used to provide contextual information about the filled data to the subsequent processes in a form processing system.

When a blank colour form has been scanned, quantised and analysed by the form structure identification process discussed in section 3.2.1, the lines, boxes, text and graphic components can be separated from the form. These extracted components can then be passed to an OCR package for recognition. As shown in section 3.3.1, with the current OCR technology, almost 99% of the printed text will be recognised correctly.

A list of all the possible cue words within all the forms in the data set can thus be generated. These cue words are then grouped together according to their context. Table 4-1 shows all the identified cue words and their final groupings for the fifteen different types of form shown in Appendix D. This was done by going through all the forms manually and identifying the 'cue words' that could be used to represent the meaning of the filled-in data. It is important to point out that these identified cue words are not sufficient to represent all of the possible fields that could be found in a form. This is due to the fact that in some cases, there were no possible generic cue words that could be used to represent the meaning for particular fields. This is especially true for special fields where the filled-in data could be any form of words used to convey additional information, i.e. special instructions, answers to specific circumstances or justifications etc. Moreover, there are thousands of types of form that exist in this world, each with their own specific purposes and applications. Trying to generate a list of generic cue words to cope with all the fields for all of these forms is impossible. Fortunately, our experimental evidence suggests that such fields normally account for less than 2% of the total number of fields in a form.

Context Category	Cue word (s)	Characteristics
Name	Name, Company, Employee, Surname, Forenames, Initials, Title, Author, Course, Subject, School, College, University, Requested by, Received by, Who, Organising body,	Characters only
Address (1)	Address, Add, Venue, Destination, Postcode, Post	Characters + Numeric
Address (2)	Country, Town, County, City	Characters only
Number (1)	Tel, Telephone, Phone, Total, $, £, p, Fax, Year, Volume, Page, Pages, Extension, Ext, Edition, Age, Barcode	Numeric only
Number (2)	Time, Day, Month, Date, Number, No., Value, Items	Characters, numeric or both
Number (3)	Amount in words	Characters only
Department	Department, Dept., Faculty, Fac	Characters only
Position	Position	Characters only
Gender	Sex	Characters only
Status	Marital Status, Nationality	Characters only
Occupation	Occupation, Job Title	Characters only
Signature	Signed, Signature	Image
Application specific	Supplier, Model, Make, ID, Username, Serviced used, Colour, Code, from, to, Details, Location, Nature of illness, Reason, Work performed, Branch	Characters, numeric or both

Table 4-1. An example of all the common cue words that could be found in 15 different types of forms to provide the contextual knowledge about the filled-in data to the recognizer.

It is logical to assume that if the OCR engine can recognize all of the words, it will have the corresponding locations for each of these recognized words in the image. However, due to the lack of a Software Development Kit (SDK) for the OCR engine used in this work, the recognized cue words locations in this study were determined manually by marking up the locations in the image. For any of the potential cue words that had one or more miss-recognized characters, the locations of such cue words were discarded from the list. This is to ensure that the manual mark-up process accurately reflects the actual cue word identification rate of the system. This cue word location information was then passed to a simple layout analysis process (described in detail in the next section) so that the filled-in data could be linked to the cue words in order to provide the crucial contextual information necessary for the recognition process. In this process, a mechanism is implemented that attempts to avoid the miss-association of these identified cue words to the wrong field's filled-in words. This is done by examining the location and geometrical characteristics of the cue words (as found by the OCR software) and the filled-in words so that all of the cue words are linked to the appropriate filled-in words.

4.2 Finding The Filled-in Words

The filled data extracted from a form is frequently in the form of chains of characters with some of them being connected together (touching characters). As the recogniser used in this study is a word level recogniser, these characters have to be grouped together to form words before they can be recognized or linked to the contextual cue words. In this work, a bottom-up approach is employed whereby a connected component analysis (contour tracing) method is first applied to the entire extracted data image to determine the individual groups of touching characters. Each of these groups is then merged together to form words depending on the inter-component distance between them. This merging technique is based on the fact that the gap between characters in a word is usually smaller than the gap between words. However, a fixed inter-component threshold for merging the character groups together was found not to be useful as different people have different writing styles and inter-character gap sizes. Thus, a dynamic threshold was considered to be more appropriate.

Ideally, this dynamic threshold value could be determined by averaging all of the filled-in characters widths found in a form. Unfortunately, the connected component widths found might not reflect the true character width due to the fact that the connected components could contain two or more touching characters or an image such as a signature or drawing. Table 4-2 shows the detected connected component distributions for the 10 different writers used in this study. Columns 1 & 2 show the total number of connected components found and the percentages of connected components that are actual single characters in each of the forms. Columns 3 & 4 then show the actual mean and median width values per writer derived from the connected components that are real single characters. Columns 5 & 6 show the computed mean and median width values for each of the writers derived from all the connected components, whilst columns 7 & 8 show the computed mean and median width values based on connected components that have a width of less than 50 pixels.

	Total connected components found	% of CC = single character	Actual mean width for CC= single character	Actual median width for CC= single character	Computed mean width for all CC	Computed median width for all CC	Computed mean width for CC<50 pixels	Computed median width for CC<50 pixels
Writer 1	185	80%	20	21	28	24	23	23
Writer 2	285	90%	17	15	17	18	18	16
Writer 3	275	90%	18	18	21	19	19	20
Writer 4	224	85%	20	22	27	24	23	24
Writer 5	235	89%	21	20	26	24	24	23
Writer 6	184	82%	18	18	22	20	20	20
Writer 7	223	90%	20	20	24	23	23	22
Writer 8	188	90%	19	19	24	21	21	21
Writer 9	231	90%	19	20	24	22	22	22
Writer 10	205	89%	21	21	27	25	24	24
Overall	**2235**	**89%**	**19.8**	**19.4**	**24**	**22**	**21.7**	**21.5**

Table 4-2. The connected components' width distribution for 10 different writers' filled-in data.

The results shown in column 1 indicate that there are around 11% of connected components that are not single characters (i.e. they are a signature, drawing or touching characters etc). These components produce approximately 21% and 13% of the writer specific character width prediction errors calculated using the mean and median values on all the connected components respectively. The results in the last 2 columns show that if the connected components that have a width of more than 50 pixels width are discarded from the calculations (they are unlikely to be a single character) then both the computed mean and median values will produce a fairly accurate character width prediction (less than 10% error rate).

Thus, the dynamic inter-component threshold value can be determined statistically by identifying the mean or median value of the connected components widths that have a width of less than 50 pixels within a given filled-in form. Since both methods produced a reasonable estimation of the actual character width of a writer, either one would be adequate to determine the threshold value. The method chosen for this study is to calculate the mean value of the connected components' width that are less than 50 pixels.

Empirical results (based on these 10 writer samples) show that the character gap between characters that belong to a word is usually less than the mean width of a character whilst the gap between words in the same field is usually more than 1 mean character width. Thus, we deduced that if a horizontal gap between two connected components is less than a mean character width then these two connected component groups should be merged together. Figs. 4-2 and 4-3 show an example of the extracted filled-in data grouped as individual or touching character components and as the final words found.

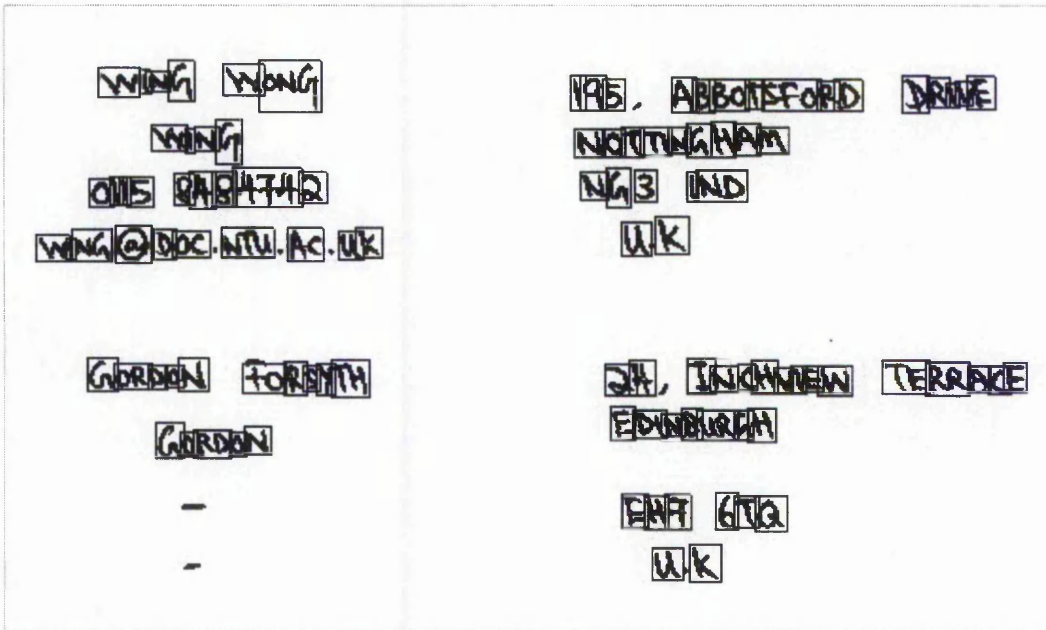


Fig. 4-2. An example of all the detected connected components (characters or groups of touching characters) after the CCA (connected component analysis) process has been applied to an extracted filled data output.

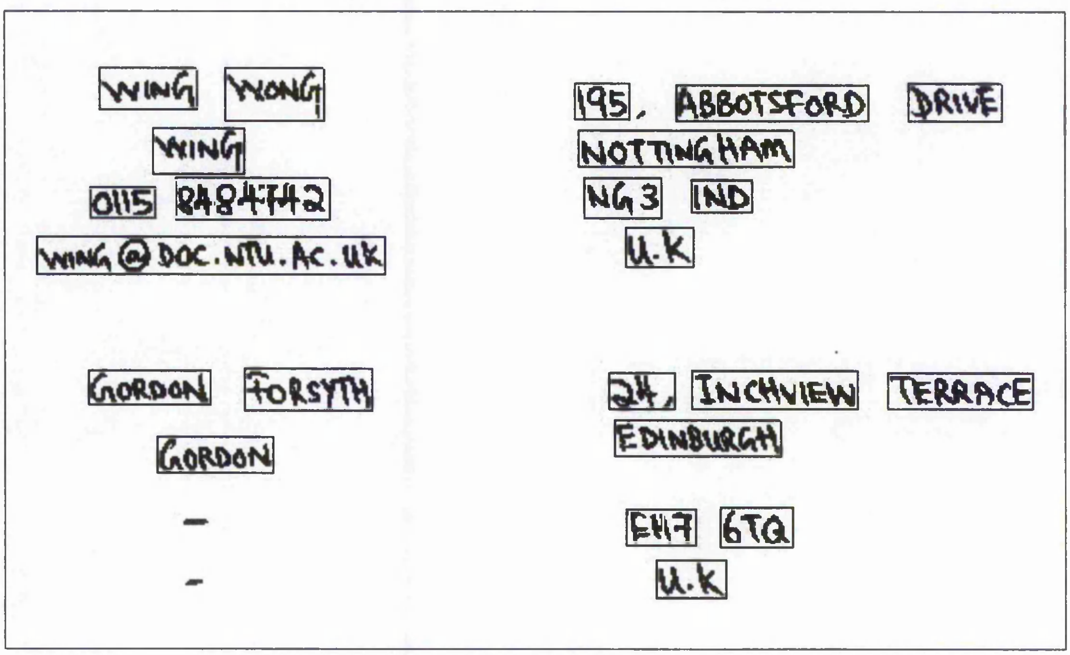


Fig. 4-3. An example of the words found after the merging process for the image shown in fig. 4-2.

For the full experimental results on this character merging technique, please refer to section 4.4 and Appendix E.

4.3 Linking The Cue Words To The Filled-in Words

As mentioned earlier, forms contain cue words that are used to 'guide' the respondent to fill-in the appropriate data or information into the space provided. There are 2 factors that affect the assignment of cue words to the filled-in words – the direction of the cue word in respect to the filled-in word and the distance between them.

Table 4-3 shows the distributions of 236 filled-in words that could be linked to cue words either directly or via other linked filled-in words for the 15 different types of forms shown in Appendix B. From the results, it is clear that cue words are most commonly found at the immediate left hand side of the filled-in space (56.8%). An additional 14% can be linked to a cue word on the left via another linked filled-in word. However, in some cases, cue words can also be located immediately above the filled-in area (15%) or via another linked word. Only in very special cases are the cue words found at the end or bottom of the filled-in space, e.g. signature and date fields.

Contextual Information Source Location	**Percentages**
Cue word exist directly to the left of the filled-in word	56.8%
Contextual information obtained indirectly via a left linked word	14.0%
Cue word exist directly above the filled-in word	14.8%
Contextual information obtained indirectly via an above linked word	12.7%
Other sources (bottom, to the right of the filled-in word)	1.7%

Table 4-3. The distributions of the possible contextual knowledge sources for 236 filled-in words in 15 different kinds of forms.

Simplistically, the correct cue word that should be linked to the filled-in words is the one that has the smallest distance to the filled-in word. However, in real life, the distance between the cue word and the filled-in word can be very large. In some cases, the filled-in data for a specific field may contain several words and the distance between the last one or two filled-in words and the correct cue word is larger than the distance to other cue words (see fig. 4-4). Therefore, a cue word cannot be linked to a filled-in word just by relying on a simple distance calculation.

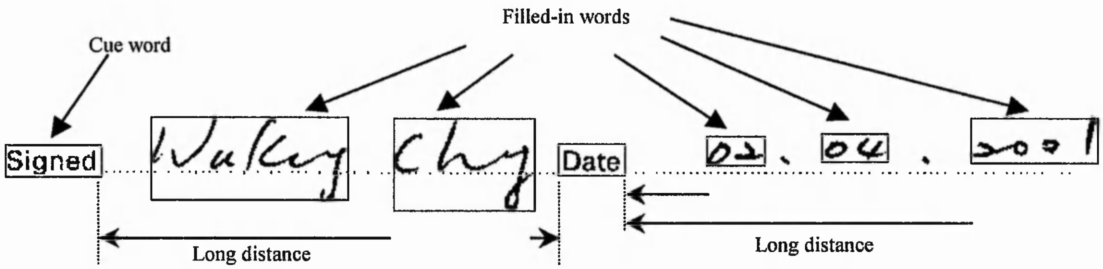


Fig. 4-4. An example of a filled-in word that has a shorter distance to the incorrect cue word than the correct cue word.

In this work, a sequence and set of rules was deduced so that only the most likely cue word is assigned to the filled-in word. According to the observations shown in table 4-2, most of the filled-in words can be linked to the correct contextual cue words using four basic rules in sequence.

These four rules are generalized based on the assumption that all writing is from left to right and the filled-in sequence is from top to bottom. Thus, by starting at the top left position of the form, the rules can be illustrated in detail as follows:

Note:

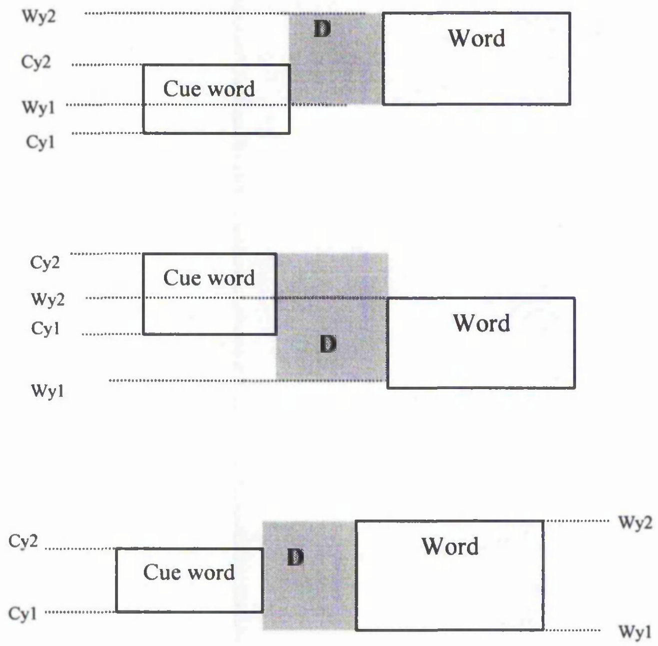
Wx1, Wx2, Wy1, Wy2 denote the bounding box XY coordinates of the word image

Cx1, Cx2, Cy1, Cy2 denote the bounding box XY coordinates of the cue word

Lx1, Lx2, Ly1, Ly2 denote the bounding box XY coordinates of the linked word image

Rule #1:

Search to the left of the word image for the possible presence of a cue word



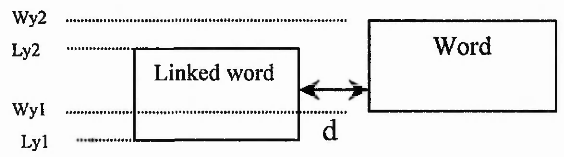
<u>Link the current word to cue word when</u>

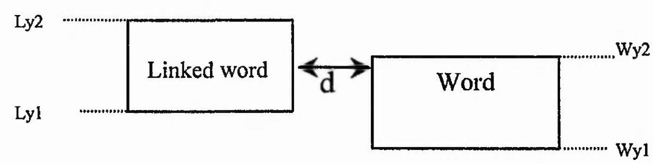
$\{(Cy1 < Wy1 < Cy2)$ or $(Cy1 < Wy2 < Cy2)$ or $(Cy2 <= Wy2\ \&\ Cy1 >= Wy1)\}$ & {No other word exists in between the cue word and the inspected word (area D*)}

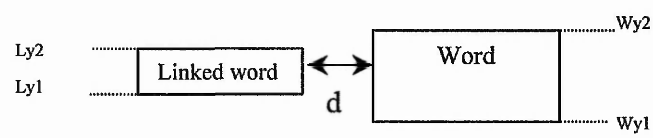
* *There is no restriction to the size of area D and the width of the area is dependent on the distance between the first left hand cue word and the filled-in word*

Rule #2:

If rule # 1 fails, search to the left of the word image for a possible word image that has already been linked to a cue word







<u>Link current word to linked word when</u>

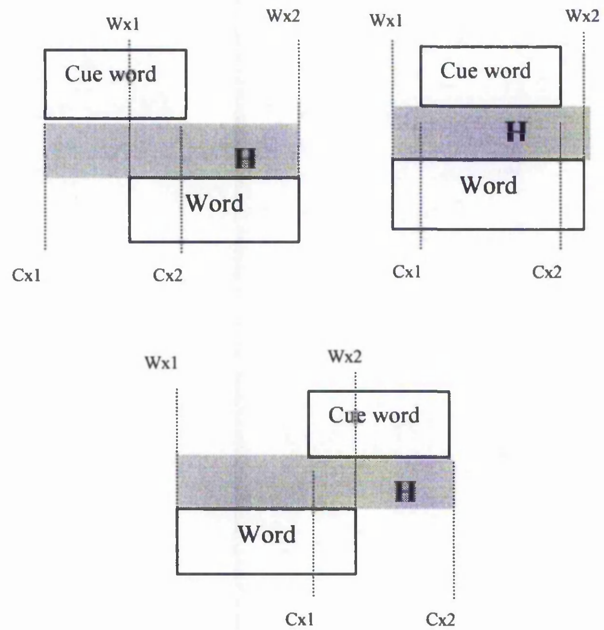
{(Ly1 < Wy1 < Ly2) or (Ly1 < Wy2 < Ly2) or (Ly2 <= Wy2 & Ly1 >= Wy1)}

& **{d<= the average word length found in the document** [#]} & {No other word exists in between them}

\# *This value is chosen based on the observation that the gap between two words in the same field is generally less than the average word length found within that form. Thus, it is believed that by setting a threshold gap of the average word length, the covered searching distance is sufficient to find the potential contextual sources.*

Rule #3:

If rules 1&2 fail, search above the word image for the possible presence of a cue word



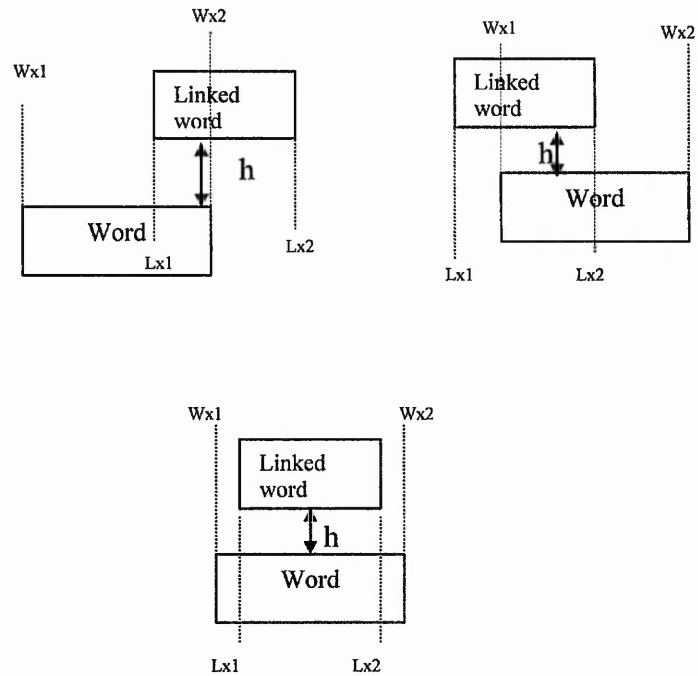
<u>Link current word to cue word when</u>

{(Cx1 < Wx1 < Cx2) or (Cx1 < Wx2 < Cx2) or (Cx2<=Wx2 & Cx1>=Wx1)} & {No other linked word exists in between the cue word and the inspected word (area H[+])}

[+] *There is no restriction to the size of area H and the height of the area is dependent on the distance between the first found cue word and the filled-in word*

Rule #4:

If rules # 1, 2 & 3 fail, search above the word image for a possible word image that has already been linked to a cue word



<u>Link Word to Linked word when</u>

{(Lx1 < Wx1 < Lx2) or (Lx1 < Wx2 < Lx2) or (Lx2 <= Wx2 & Lx1 >= Wx1)} & **{h<=3 times the current word height** [*]} & {No other word exists in between them}

[*] *This h value is chosen based on the observations that the vertical gap between two lines of words in the same field is normally within 3 times the current word height. Thus, it is believed that by setting a threshold gap of 3x the word height, the covered searching distance is sufficient to find the correct contextual sources.*

Unfortunately, there is a fundamental problem for this linking algorithm – the order of the detected words is not the same sequence in which they are written. When the contour tracing method is applied to the image, it starts the search for seeds (black pixels) at the top left corner of the image and steps through the image searching along the x-direction first. Once a line has been completed, the search moves on to the next line of pixels down and continues until the last pixel of the image is reached. Since the height of the handwritten words varies across the field, the order of the detected words are then in the sequence of the searching order (see fig. 4-5).

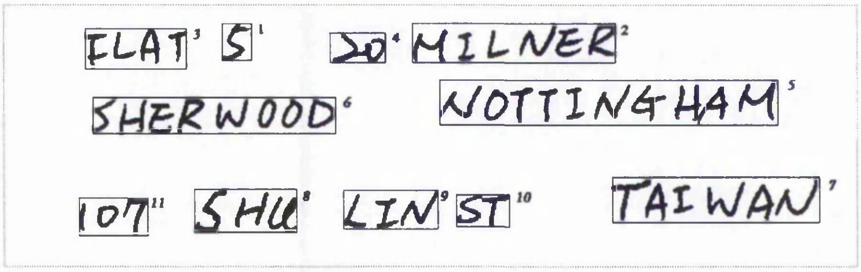


Fig. 4-5. The detected words are found to be in the order of the searching sequence.

To resolve this problem, an alignment method is needed to re-arrange the words, prior to linking, so that they are in the sequence of left to right, top to the bottom. This is done by sub-dividing the image into several regions based on the horizontal histogram profile of the image and re-arranging the detected words of these regions from left to right. Fig. 4-6 shows the horizontal histogram profile for the image shown in fig. 4-5.

Fig. 4-6. The horizontal histogram profile for the image shown in fig 4-5.

Based on this horizontal histogram profile, the words that belong to a sentence or field can thus be identified. The number and size of the regions are determined by the start and end positions of the histogram groups. For example, for the image in fig. 4-5, there will be 3 groups of words after the region has been divided (as shown in fig. 4-7) and the words in each group are then re-arranged in left to right order.

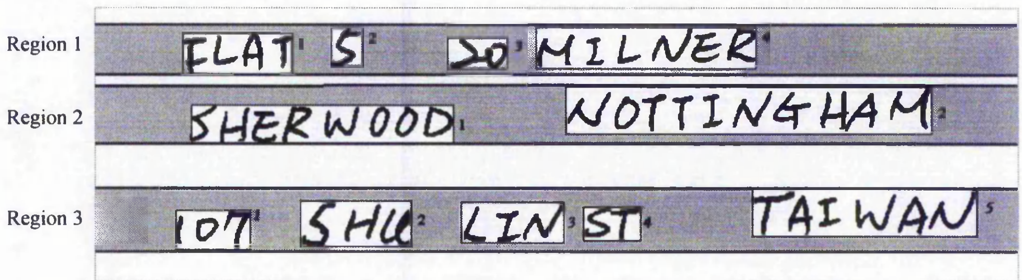


Fig. 4-7. The subdivided regions and the re-arranged words after the horizontal histogram analysis for image shown in fig. 4-5.

Unfortunately, this method is not robust enough to divide the regions correctly every time. For some forms, the histogram projection method will group more than one sentence together in a region. As a result a wrong word sequence is produced leading to a miss linking of filled words to the context. Figs. 4-8 (a) & (b) shows an example of how this can happen.

Fig. 4-8(a). The histogram projection method will occasionally fail to divide the regions correctly; especially for multi-column forms.

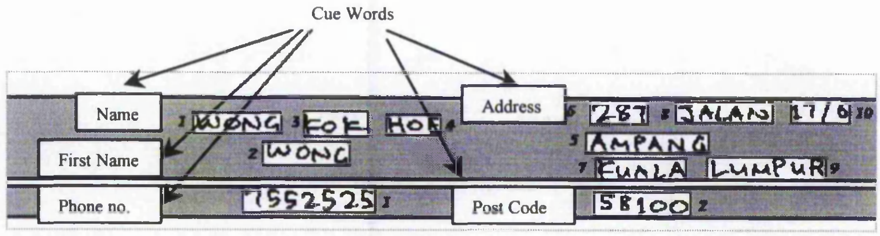


Fig. 4-8(b). The subsequent divided regions and the re-arranged words after the horizontal histogram analysis for image shown in fig. 4-8(a). Note that words no. 5,7 & 9 will fail to find their cue words since word no. 6 is linked to the cue word only after processing word no. 5.

There are two possible methods for re-solving this issue – improving the region dividing technique or modifying the linking method/sequence. Since there is no 100% foolproof method for segmenting the regions correctly that does not incur heavy penalties in terms of processing time and computational needs, a second pass method was employed. It was found that by performing a second linking attempt (using the same set of rules) on the words that have not been linked, this problem could be resolved completely.

There is another potential situation that could cause failure in the linking algorithm – the irregularity of the cue word position. As stated in rules 1 to 4, the contextual information for a word should be found at the left or above cue words or via linked words. However, in some form designs, the position of the cue word is positioned centrally above the filled-in area. This results in some of the filled-in words failing to find their related context. Fig. 4-9 shows an example of this scenario.

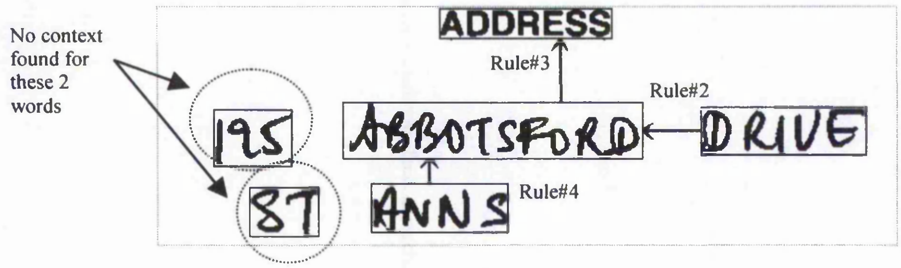
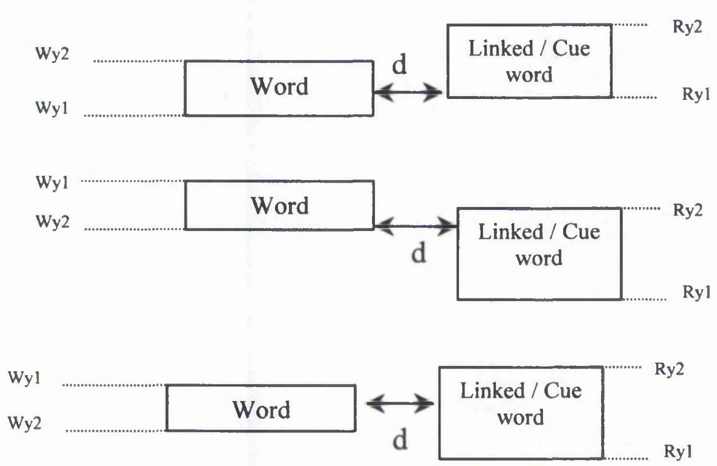


Fig. 4-9. Some of the words fail to locate their related context due to the position of the cue word

To resolve this problem, an extra rule is added to the second linking attempt – i.e. if rules #1-4 still fail to locate the context for an unlinked word in the second pass attempt, the system should search for possible linked/cue words to the right position of the filled word. This is illustrated in detail as follows:

Rule #5:

If rule #1-4 fails during the second pass, search to the right of the word image for a possible cue word or word image that has already been linked to a cue word



Link current word to right cue/linked word when

{(Ly1 < Wy1 < Ly2) or (Ly1 < Wy2 < Ly2) or (Ly2 <= Wy2 & Ly1 >= Wy1) or (Ly2 >= Wy2 & Ly1 <= Wy1)} & {**d<= the average word length found in document**[#]} & {No other word exists in between them}

\# *This value was chosen based on the same observation found for rule no. 2*

Note that this rule is only applied at the second pass attempt to avoid the problem of linking error for some form designs such as the one shown in fig. 4-4. Since there is only less than 1.7% (see table 4-3) of the filled-in words that have their context cue words located to the right position, it is more important to get the 98% of filled words to link correctly. Hence, it is only when all the attempts fail to locate the cue words that this rule is used.

4.4 Experiment I: Filled Word Identification

As the contextual linking algorithm and the CSR engine used in this study required knowledge of the filled-in data in word format, a study is needed to assess the effectiveness of the filled-in data word finding technique. There are 2 sets of data being used in this experiment; the first one is used to assess the effectiveness of the word finding technique when dealing with different kinds of handwriting style, whilst the second data set is used to assess the word finding performance when dealing with extracted data from different kinds of form design. The first set of handwritten words that were used in this experiment was obtained from the extracted data images used in section 3.3.3. These consist of 3 different types of form, each filled-in by 10 different writers. The second data samples were from handwritten words extracted from 15 different types of form, each filled by the same writer. The 30 extracted data images are shown in Appendix E whilst the 15 filled-in form samples are shown in Appendix D. Table 4-4 shows the experimental results for the first data sample. In total, there are 1382 filled-in data words identified correctly out of the 1498 filled-in data word total for these 30 forms. The overall success rate is approximately 92%.

	Form4-1			Form4-2			Form4-3		
	No. of words Expected	No. of word found correctly	Accuracy (%)	No. of words Expected	No. of word found correctly	Accuracy (%)	No. of words Expected	No. of word found correctly	Accuracy (%)
Writer01	60	57	95%	57	57	100%	39	35	90%
Writer02	69	65	94%	56	55	98%	43	43	100%
Writer03	65	59	91%	58	56	97%	42	42	100%
Writer04	51	48	94%	55	52	95%	36	36	100%
Writer05	48	39	81%	44	36	82%	38	32	84%
Writer06	54	51	94%	53	51	96%	40	40	100%
Writer07	55	48	87%	54	53	98%	50	44	88%
Writer08	46	44	96%	49	48	98%	44	42	95%
Writer09	64	60	94%	58	52	90%	40	29	73%
Writer10	45	29	64%	49	44	90%	36	35	97%
Overall	**557**	**500**	**90%**	**533**	**504**	**95%**	**408**	**378**	**93%**

Table 4-4. Overall character-merging results tested on 30 filled-in forms (3 forms x *10 writers).*

The results suggest that the proposed character-merging technique is capable of identifying the filled-in words correctly for more than 90% of the time. The 10% fallout is mainly due to 2 reasons – inconsistent gaps between words or characters and overlapping words due to poor form design (see fig. 4-10).

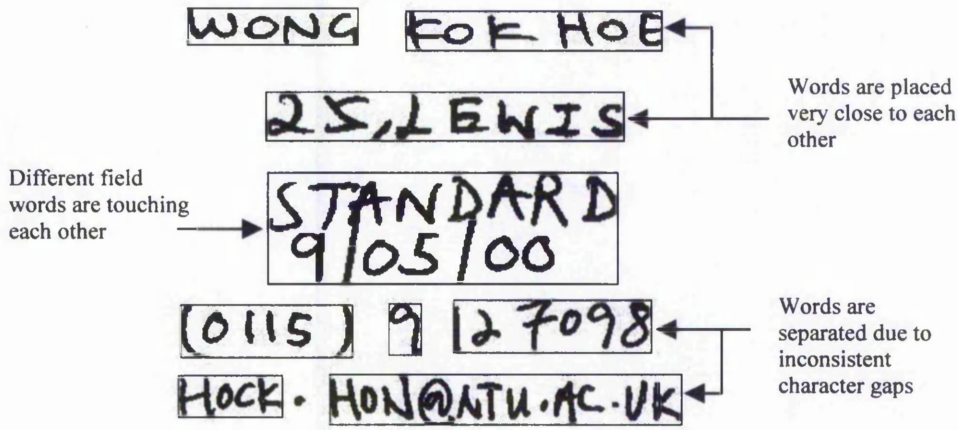


Fig. 4-10. Some of the words are difficult to identify due to inconsistency in the character gap or overlapping characters caused by poor form design.

Table 4-5 shows the experimental results for the second data sample. In total, there are 445 filled-in data words identified correctly out of the 493 filled-in data word total for these 15 forms. The overall success rate is approximately 90%.

These results suggest that form design does influence a person's writing style. A narrow filled-in space will force the characters and words to be packed together while boxed discrete filled-in areas will tend to separate characters apart. Nevertheless, the experimental results have shown that the merging technique was able to merge characters into words correctly 90% of the time both for different writing styles and form designs. Figs. 4-11 and 4-12 show an example of the influence of a poor form design and box discrete fields on writing style

	No. of expected words	No. of word found correctly	Accuracy (%)
Form #1	33	31	94%
Form #2	17	16	94%
Form #3	27	25	93%
Form #4	25	20	80%
Form #5	30	28	93%
Form #6	51	49	96%
Form #7	31	29	94%
Form #8	16	13	81%
Form #9	31	28	90%
Form #10	34	28	82%
Form #11	50	48	96%
Form #12	67	60	90%
Form #13	30	27	90%
Form #14	31	26	84%
Form #15	20	17	85%
Overall	**493**	**445**	**90%**

Table 4-5. Overall characters merging results tested on 15 different forms filled-in by a single writer.

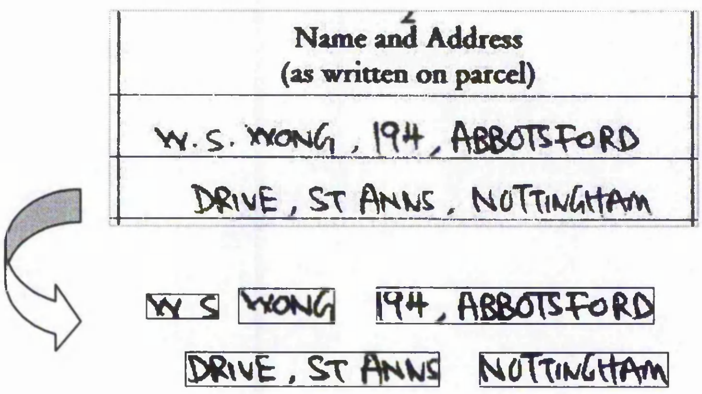


Fig. 4-11. Poor form design such as allocating too little space and asking too many questions at once contributed to some of the word identification failures.

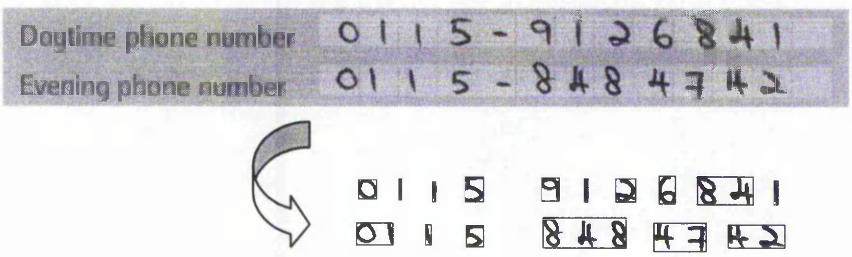


Fig. 4-12. The proposed merging technique fails to form some words correctly where the gap between characters is inconsistence in box discrete field data.

4.5 Experiment II: Contextual Cue Words Retrieval

As the retrieval of contextual cue words relies heavily on the OCR engine, a study is needed to assess the effectiveness of using such a method for retrieving the cue words. This experiment aims to evaluate the cue words retrieval rate for form documents that are digitized with 200dpi resolution. As reported in section 3.3.1, the OCR performance at resolutions of 200dpi and above is identical for form images that are processed with and without the quantization process applied. Therefore, in this experiment, only the original digitized form images were evaluated. The forms that were used in this experiment were the second data set of 15 forms that were used in the previous experiment (section 4.4) – i.e. the form samples shown in Appendix D.

Table 4-6 shows the experimental OCR results obtained from the 15 original 24-bit full colour images. In total, there are 186 cue words for these 15 forms and the retrieval rate is computed manually by counting the total number of correctly recognized cue words in a form over the total number of cue words in that form.

	Total number of cue words	Total number of OCR recognized cue words	Retrieval rate (%)
Form # 1	28	28	100%
Form # 2	20	20	100%
Form # 3	10	9	90%
Form # 4	24	24	100%
Form # 5	15	15	100%
Form # 6	16	16	100%
Form # 7	10	9	90%
Form # 8	14	14	100%
Form # 9	23	22	96%
Form # 10	18	18	100%
Form # 11	42	42	100%
Form # 12	16	15	94%
Form # 13	15	14	93%
Form # 14	19	19	100%
Form # 15	8	8	100%
Overall	**278**	**273**	**98.2%**

Table 4-6. Contextual cue words retrieval rate for 15 form images digitized at 200dpi resolution.

As expected, the cue words retrieval rate is close to 99%. There are only five miss-retrieved cue words, these being due to the very small printed size of the cue words in each particular form. Since cue words are rarely printed in small font size, this miss-retrieval case is considered small and negligible. The experimental results therefore suggest that more than 98% of the cue words can be located successfully with the current OCR software at an image resolution of 200dpi.

4.6 Experiment III: Linking Accuracy

As the contextual cue words are used to guide the CSR search domain to achieve better recognition results, the accuracy of the linking algorithm becomes very important in determining the magnitude of this gain. If the linking accuracy is poor, the CSR will be miss-directed and the improvement in recognition rate will be limited, if not actually reduced. This experiment aims to evaluate the linking accuracy by examining the total number of correct and incorrect cue word to filled-in word links. The accuracy is calculated manually as the number of correctly linked words to the total number of filled-in words in a given form. The same 2 data sets used in experiment I (section 4.4) were used for this experiment – i.e. the form samples shown in Appendix D and E.

Table 4-7 shows the experimental results for the first data sample. In total, there are only 12 words linked incorrectly out of the 1509 linked word total for these 30 forms. The overall accuracy rate is approximately 99%. The snapshots of the linking results are shown in Appendix F.

	Form4-1			Form4-2			Form4-3		
	No. of correct linked words	No. of wrong linked word	Accuracy (%)	No. of correct linked words	No. of wrong linked word	Accuracy (%)	No. of correct linked words	No. of wrong linked word	Accuracy (%)
Writer01	70	1	98.6%	66	0	100%	49	0	100%
Writer02	66	3	95.7%	60	0	100%	42	0	100%
Writer03	60	1	98.4%	62	0	100%	38	0	100%
Writer04	56	0	100%	58	0	100%	38	0	100%
Writer05	41	3	93.2%	47	1	98.0%	37	0	100%
Writer06	50	1	98.0%	53	0	100%	38	0	100%
Writer07	41	0	100%	59	0	100%	45	0	100%
Writer08	45	0	100%	49	0	100%	41	0	100%
Writer09	59	0	100%	58	0	100%	39	0	100%
Writer10	36	2	94.7%	56	0	100%	38	0	100%
Overall	**524**	**11**	**97.9%**	**568**	**1**	**99.8%**	**405**	**0**	**100%**

Table 4-7. Overall linking results tested on 30 filled-in forms with 10 different writers.

The linking errors are mainly due to form design, where some of the fields are placed very close to each other (fig. 4-13).

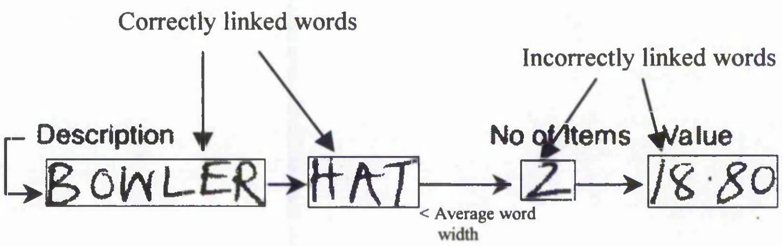


Fig. 4-13. When several fields are placed very close together (less than the average width of the filled words), then the linking method will fail to link the correct context to the filled word.

It is important to point out that in this experiment, not all of the words found in a form were linked. There are 76 extracted words (4.8% of 1585 total extracted words) that were not linked to any context. The main reason for this is due to the fact that about 4% of the extracted words are formed by two or more words from different fields touching with each other. Thus, when the linking method applied, there will be more than one cue word that meets the linking criteria for these touching words (see fig. 4-14). Such 'words' will then be excluded from the link list and classified as unlinked by the linking engine. However, the linking of a 'word' to more than one cue word could be used as an indicator to the system of possible touching words within the identified area. Additional techniques could then be invoked to attempt to separate them.

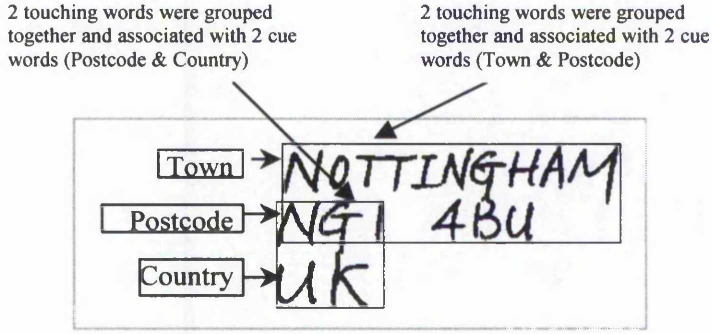


Fig. 4-14. Some of the touching characters will group 2 or more words together from different fields. The linking of more than one cue word could thus be used to detect such cases.

There is another 0.8% of the extracted words that fail to link to any context due to the absence of a cue word. This is a case where no suitable cue word is available for the field as mentioned in section 4.1.

	Total number of correct linked words	Total number of wrong linked words	Linking accuracy (%)
Form # 1	34	0	100%
Form # 2	16	0	100%
Form # 3	27	0	100%
Form # 4	32	0	100%
Form # 5	32	3	91.4%
Form # 6	45	5	90.0%
Form # 7	29	3	90.6%
Form # 8	7	3	70%
Form # 9	28	1	96.6%
Form # 10	59	4	93.7%
Form # 11	72	8	90.0%
Form # 12	62	5	92.5%
Form # 13	50	3	94.3%
Form # 14	40	0	100%
Form # 15	22	0	100%
Overall	**555**	**37**	**93.8%**

Table 4-8. Overall linking results tested on 15 different types of form filled-in by one writer.

Table 4-8 shows the experimental results for the second data sample. In total, there are 37 words linked incorrectly out of the 592 linked words total for these 15 forms. As expected, since the writing style is affected by the form design, different forms produce different linking results. However, the overall accuracy rate is generally over 90% for all but form number 8. The poor linking rate in this form is due to poor form design and lack of data for this particular form (see fig. 4-15).

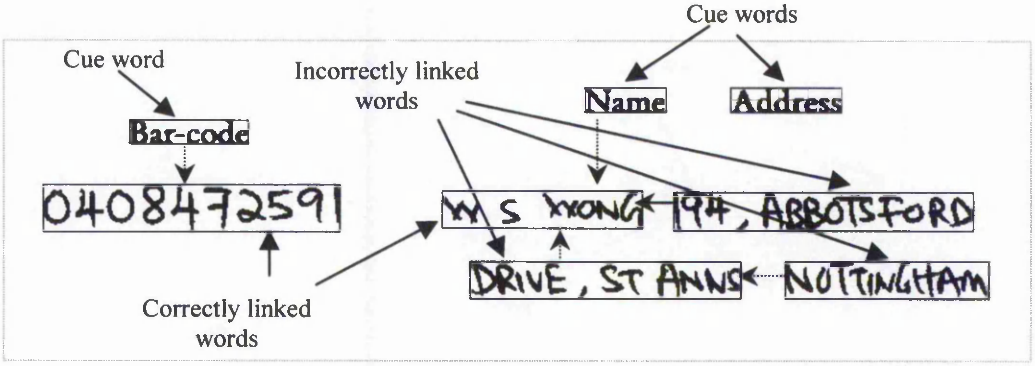


Fig. 4-15. Sample form#8, which has a very poor layout design that allocates too little space and asks more than a single question per field.

There are 19 extracted words (3.4% of the total extracted words) that were not linked to any context. Most of these words (13 out of 19) failed to find their context due to the inconsistency of gap distance between the words, which, again, is mainly caused by the form design.

The results from this experiment suggest that the linking method is capable of linking the contextual cue words to the filled-in words with a very high accuracy rate (close to 100%) for certain types of form. The overall linking accuracy tested on 15 different types of forms yields an average accuracy rate of 93.8%. The experimental results also shows that handwriting styles have little effect on the linking accuracy; in fact the linking accuracy is only really dependent on the form design. This experiment also demonstrates that the linking method can be potentially used to detect touching words from different fields.

4.7 Experiment IV: Contextual Focused CSR

Research has shown that the CSR recognition results can be improved by reducing the size of the lexicons used [32,164,165]. By using the contextual knowledge acquired from the linked cue words, it is possible to pre-select an appropriate lexicon prior to any recognition. Thus, the lexicon used to recognise a word could be reduced to a specific lexicon such as Name, Address or Postcode.

This experiment aims to quantify the overall CSR performance that could be gained by using contextual focussed recognition rather than the conventional CSR method. In this experiment, only the words (obtained from the 3 different types of form that were filled by 10 writers in section 4.5) that are suitable for a holistic word recognizer were used. Thus, the data fields considered in this experiment were 'name', 'month', 'town', 'country', 'address' and some specific fields such as 'service used' and 'item description'. The total number of unique words in the CSR database was 937, which could be divided into several sub-dictionaries by the contextual information as shown in table 4-9. By using a different lexicon for each of the words (based on their linked context), the search domain for the recogniser could be effectively reduced by approximately 56%-99%.

Contextual Cue word	Number of words
Country	230 words
Address	524 words
Name	167 words
Town	163 words
Month	20 words
Item Description	21 words
Service used	5 words
Total unique words	**937**

Table 4-9. The words distribution for each of the data fields and the total number of unique words used by the CSR.

Note that effort has been made to include all of the possible words for each of these fields (except the name, address and item description fields). Since it is impossible to include all of the names, items or addresses that exist in this world into the dictionary, the lexicons that were used by the CSR for these fields contained only the name, item and address data that appeared in the test set forms. However, the address data do include all the major UK town and country names. For the full list of lexicons that were used in this experiment, please refer to appendix G.

Table 4-10 and Figure 4-16 show the results of the CSR performance on the extracted data both with and without the use of contextual information. There were a total of 807 words from these 30 filled forms, in which 23 of them are unlinked words and only 2 of them are incorrectly linked words. The low number of incorrectly linked words is due to the fact that most of the incorrectly linked words that are reported in experiment III in section 4.6 were numeric words. These were not considered in this experiment as the holistic recognition engine used could not deal with such data. For the 23 unlinked words, there can be no improvement in the CSR rates since the full dictionary must be used when performing the recognition task.

		Top 1	Top 5	Top 10
Form 4-1	No Context Aid	41.3%	66.9%	74.8%
(242 words tested)	With Contextual Focus	59.1%	77.3%	80.6%
	Improvement	**+17.8%**	**+10.4%**	**+5.8%**
Form 4-2	No Context Aid	38.7%	64.5%	73.2%
(313 words tested)	With Contextual Focus	47.9%	74.1%	81.2%
	Improvement	**+9.2%**	**+9.6%**	**+8.0%**
Form 4-3	No Context Aid	49.6%	75.8%	82.1%
(252 words tested)	With Contextual Focus	59.9%	82.5%	88.1%
	Improvement	**+10.3%**	**+6.7%**	**+6.0%**
Overall	No Context Aid	42.9%	68.8%	76.5%
(807 words tested)	With Contextual Focus	55.0%	77.7%	83.1%
	Improvement	**+12.1%**	**+8.9%**	**+6.6%**

Table 4-10. A comparison of recognition results and overall performance gain using either a 937 word lexicon (without contextual focus) or separate sub-directory sized lexicons (with contextual focus).

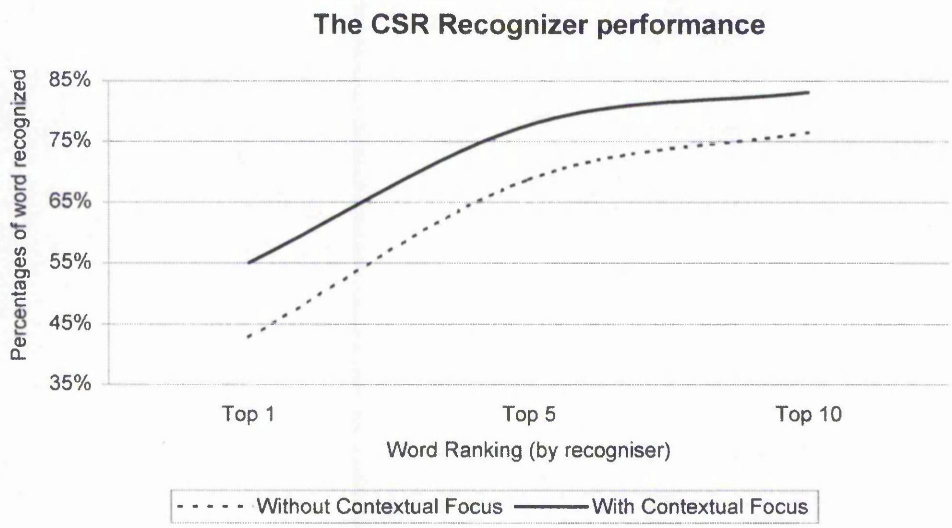


Fig. 4-16. A graphical representation of results shown in table 4-7.

Overall, the experimental results show that a recognition system that utilises contextual knowledge does have a significantly higher recognition rate (9-17%) than a similar recognition system that does not employ any contextual information. The different improvements in recognition rate seen in the 3 cases are due to the fact that the specifics of the words tested in each case are different. For example, form 4-2 contained much more address data than form 4-1, thus the lexicon size reduction in form 4-2 is less than in form 4-1 which therefore allows less scope for improvement. Thus, it is clear that when different types of form are used, the performance gains achieved will be different. In fact, if a poorly designed form was used, the recognition results could actually be reduced. Fortunately, as demonstrated in experiment III in section 4.6, this will rarely happen and in most of the cases, a linking accuracy of higher than 90% can be achieved.

It is important to point out that if the top 5 and top 10 results are considered, the overall improvement is decreased. This is due to the fact that it is always more difficult to improve a system that has a higher recognition rate than a poorly performing recognition system. Nevertheless, the contextual aided recognition system does consistently outperform a recognition system that does not employ any contextual knowledge and produces a net 6-12% improvement (top 10 and top 1 results), even when taking into account the miss-linked words.

4.8 Conclusion

A novel method has been presented that uses high OCR rates to help improve the CSR performance in a form processing application. A new character-merging technique was developed that can group related characters together to form words for the CSR task. By using the high OCR rate on the pre-printed text on a form, contextual cue words were retrieved and automatically linked to the filled-in word. These linked words, along with their contextual knowledge, are then presented to a CSR engine for recognition.

Several experiments have been conducted to evaluate the performances for each of the steps involved in the process. The experimental results show that the proposed character-merging technique is able to merge the related characters into words correctly 90% of the time for different writing styles and form designs. Using commercially available OCR software on 15 different types of form, scanned at 200dpi, an average of 98% of the cue words were successfully retrieved. With the proposed linking algorithm, these retrieved cue words were then linked to the filled-in data words with an accuracy rate of over 90% on most of the forms. This linking accuracy is shown to be sensitive to form design but not to writer style. Consequently, an average of about 3-5% of the words were not linked to any context.

The significance of this contextual knowledge linking has been determined by comparing the recognition rates for the contextual focused and conventional CSR methods. On average, the contextual information provided a 12% CSR rate improvement for top 1 word classification over the non-context aided CSR system.

5. MODEL-LESS FORM PROCESSING

The research studies conducted in chapters 3 and 4 were based on the assumption that the blank form is always present. However, under certain circumstances, the blank form will not be available to the system. For example, in a company where thousands of different kinds of invoices are being processed daily, there will be no specific design of invoice to expect and a form processing system will need to be able to process these forms (invoices) without using the blank form knowledge. The same problem may also occur in a market research company where there are lots of different kinds of forms being used in different research projects. In this case, it could be quite difficult to process all of the forms that have been accumulated over the years. All sorts of awkward conditions may arise such as the blank form version of some forms being missing, different kinds of forms being mixed together etc.

This chapter investigates the possibility of using the techniques and algorithms developed in chapters 3 and 4 to process a filled form without using the blank form. Section 5.1 first describes the idea of using OCR and the line detection algorithm presented in chapter 3 to locate and remove the pre-printed text and lines from a form. Section 5.2 then presents the idea of using the word finding and contextual linking algorithms described in chapter 4 to locate and extract the filled-in data. This is followed by two experiments to evaluate the system performance, using such an approach, in terms of the extraction recall and accuracy rate (section 5.3) and the filled word retrieval and accuracy rate (section 5.4).

5.1 Pre-printed Entities Removal

In order to reliably remove the pre-printed entities from a form without using a reference model, one needs to develop a method that is based on the pre-printed entities common characteristics. As stated in section 1.1.2, the characteristics of form pre-printed entities are as follows:

- Lines are commonly oriented in horizontal and vertical directions
- Most of the pre-printed text can be recognized using current OCR software packages

Based on these two characteristics, form pre-printed text entities can be easily located by applying the OCR software to the raw image. However, unlike the previous cases, data that has been filled-in using a typewriter or machine printed text will not be as easily distinguished from the pre-printed text as in a form processing system that has the blank form image. It is only the data that has been filled-in by hand that can be distinguished from the pre-printed text using OCR software. Experiments carried out on the handwriting samples in all of the forms used in this work has shown that OCR software is unable to recognize any handwritten words at 200dpi resolution. Hence, for a hand-filled form, by removing* all of the recognizable text from the form image, the remaining elements in the image will then be just the pre-printed lines, logos and the filled-in data. To further ensure that all of the pre-printed text can be successfully removed, OCR is applied to each of the colour domains separately. Since pre-printed text entities normally occupy only one of the colour domains in a form, by processing each of these colour domains separately the OCR will have a higher probability of recognizing the text. This is

* In this work, this was done manually by removing all the fully recognized pre-printed text found in the image using commercial available OCR software package. Note that a fully recognized word is defined as being any completely recognized word that is in the pre-defined OCR dictionary.

especially so for cases where the filled-in words are of a different colour to the pre-printed text and are not overlapping with the pre-printed text.

In addition, as the targeted filled-in text is assumed to contain handwritten words only, any connected components that are less than 8 pixels square can also be removed. As explained in section 3.2.4, this is due to the fact that a handwriting character that is smaller than 8 pixels square is difficult, if not impossible, to produce in a form document that has been scanned at 200dpi resolution. This process is done by applying the contour tracing method, used in section 3.2.2, to the image that has already had the larger OCR recognized pre-printed text removed. Any connected pixel group that is less than 8 pixels square can then be automatically removed. This action effectively removes most of the small size pre-printed text that was unrecognizable to the OCR. After the pre-printed text and noise removal process has been performed, the vertical and horizontal lines are then removed using the line detection algorithm developed in chapter 3. However, unlike our earlier work, it was not found necessary to apply RLSA to the image before line detection since the dotted lines will have already been removed* by the above connected component removal process. A smaller vertical and horizontal run length threshold was thus used in this system to ensure that short vertical lines (those used to form boxes or frames) were also removed. To avoid sections of handwritten data and other non-line components being accidentally removed, a more reliable line verification process is also employed. As handwriting stroke widths are generally thicker than the pre-printed lines in a form, pre-printed lines can be ascertained by comparing the width of the suspected lines to a given threshold. Since long and prominent lines can easily be detected with the algorithm

* During the connected component identification process, each of the line dots will be less than 8 pixels square and will thus be removed by the system

developed in chapter 3, only short run length lines that are shorter than the long line threshold are subject to this verification process. In this study, the threshold values that were used for the vertical and horizontal run length were 50 pixels (short) and 100 pixels (long) whilst the line verification threshold value was set to 3 pixels width. This is based on the empirical results obtained from 15 of the form samples, in which it was found that no 'line-like' handwritten words (dash, '1' or 'l' character) had a width of less than 3 pixels. Because pre-printed lines are generally thinner than the handwriting stroke width, we can selectively remove any detected short run length lines that have a width of less than 3 pixels.

After the removal process, the remaining objects in each of the colour domains are then merged together to form a black and white image ready for CSR process. Fig. 5-1 shows the system diagram for this pre-printed text and line removal process.

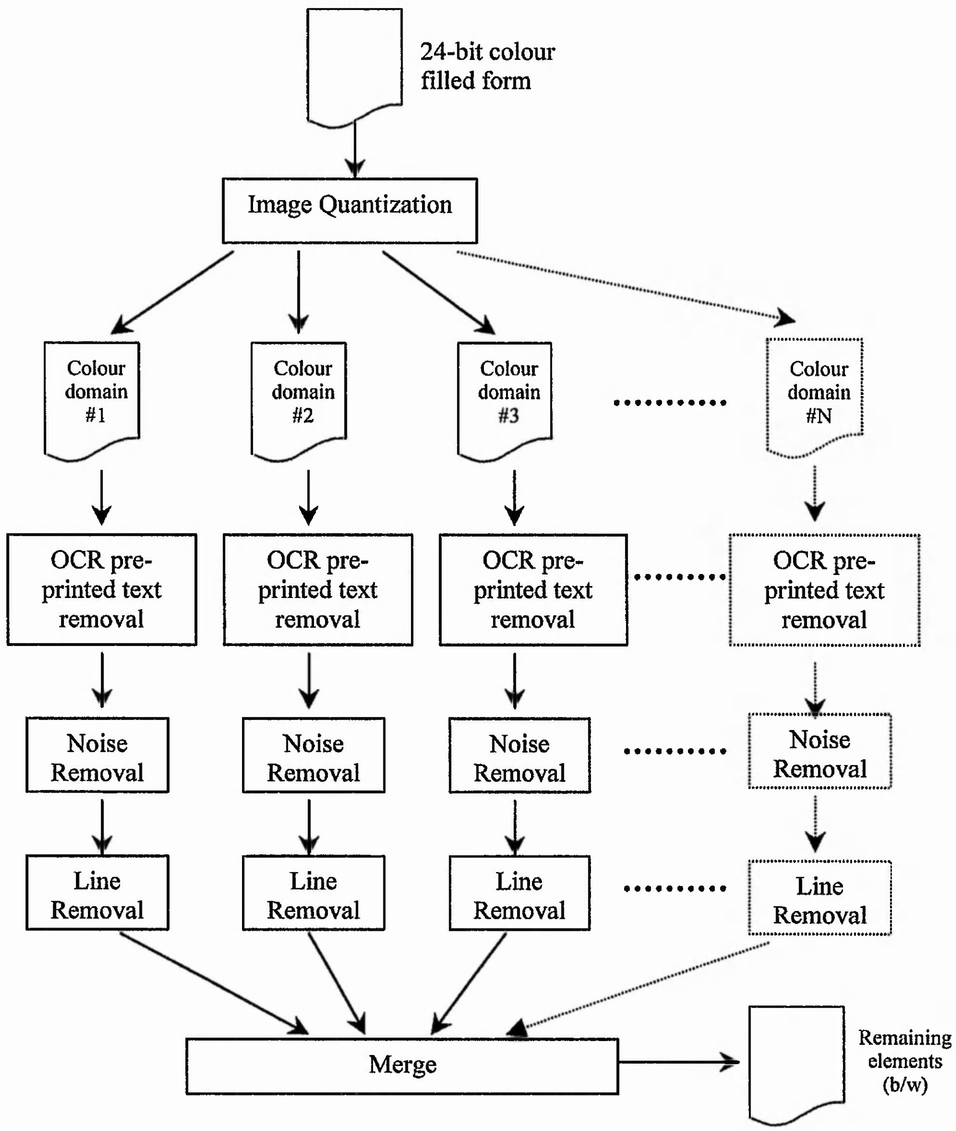


Fig. 5-1. System diagram for the pre-printed entities removal process that does not utilize any blank form sample.

5.2 Filled-In Word Extraction

Although there are a number of unsuccessfully removed pre-printed components (logos and unrecognised pre-printed text etc) left in the image after pre-printed entity removal, it was hypothesised that the cue word linking method could be used to locate the filled-in words and hence filter these unrecognised pre-printed components. Generally, filled-in data areas are unlikely to be located near to these components. This is due to the fact that a company logo is usually located at the top or bottom corner of a form and any un-recognized pre-printed text is normally the small print information that is located far from the filled-in areas. Therefore, when the 2-pass linking rule method developed in chapter 4, is applied, many of the pre-printed components that have not been removed by previous process can be discarded. Fig. 5-2 shows an example of how this can happen. In this particular example, the inability of the OCR to recognize the inverse printed text is believed to be due to the noise that was generated by the colour segmentation process. However, these words are successfully discarded using the cue word linking method.

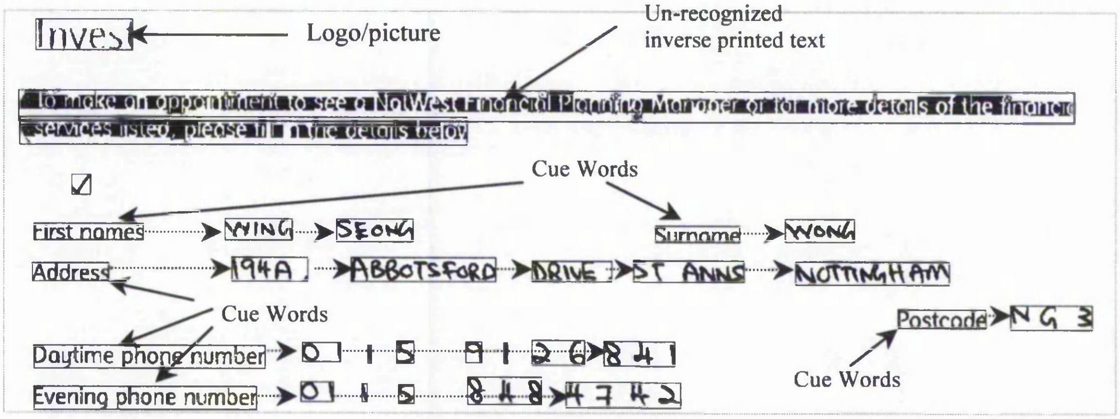


Fig. 5-2. By using the cue words and linking algorithm that was developed in chapter 4, filled-in words can be distinguished from the unsuccessfully removed pre-printed entities.

By using the cue words definitions defined in table 4-1 section 4.1, all the recognized cue word locations can be recorded. Unfortunately, like our previous OCR experiment, as there was no OCR SDK available that could be used to provide this location information automatically, the cue word locations were manually obtained. However, we expect that this information would be readily available if an SDK was used.

The word finding algorithm described in section 4.2 was then applied to the extracted output obtained in section 5.2 to locate the filled-in words. By using the linking rules developed in section 4.3, words that met the linking conditions were then extracted and fed into the CSR engine for recognition.

5.3 Experiment I: Extraction Efficiency

To assess the effectiveness of this model-less form extraction system, the same two parameters that were used in experiment II in section 3.3.2 were employed, i.e. precision and recall rate. As explained in section 3.3.2, the precision rate reflects the percentage of **correctly extracted** components over the total number of **extracted** components, whilst the recall rate provides the percentage of components that the system **can extract** over the total number of **expected** components from the specific form.

The same 2 data sets used in experiment I in section 4.4 were chosen for this study so that the final results could be compared. The first data set consists of 15 different types of form, each filled-in by one single writer and the second data set comprises of 3 different types of form, each filled by 10 different writers. The first data set is used to test the system extraction efficiency when handling different form types, whilst the second data set is used to check if the extraction performance is affected by different handwriting styles within the same type of form.

Table 5-1 shows a summary of the extraction recall and precision rate for each of the forms in the first data set. As expected, without using the blank form images, the extraction precision rate is dramatically reduced when compared to the model-based system. The precision rate of 96-99% in experiment II of section 3.3.2 compares to an extraction precision rate of approximately 71% for the model-less system. This decrease is closely related to the recall rate, which has increased from 96-100% in the system that utilized blank form images to 141% for this model-less extraction system. A recall rate of over 100% indicates that the extraction system has extracted more components than expected. These extra components are the pre-printed entities that the system failed to remove.

	Total number of expected components	Total number of extracted components	**Recall (%)**	Total number of correct components	**Precision (%)**
Form#1	126	126	**100%**	126	**100%**
Form#2	69	72	**104%**	69	**95.8%**
Form#3	152	169	**111%**	152	**89.9%**
Form#4	137	150	**109%**	137	**91.3%**
Form#5	124	125	**101%**	124	**99.2%**
Form#6	117	119	**102%**	117	**98.3%**
Form#7	97	351	**362%**	97	**27.6%**
Form#8	69	106	**154%**	69	**65.1%**
Form#9	84	126	**150%**	84	**66.7%**
Form#10	110	351	**319%**	110	**31.3%**
Form#11	211	236	**112%**	211	**89.4%**
Form#12	138	179	**130%**	138	**77.1%**
Form#13	125	154	**123%**	125	**81.2%**
Form#14	129	134	**104%**	129	**96.3%**
Form#15	71	83	**117%**	71	**85.5%**
Overall	1759	2481	**141%**	1759	**70.9%**

Table 5-1. Summary of the extraction recall and precision rate tested on 15 different types of form without using the blank form image as a reference.

The total number of pre-printed components that the system failed to remove is seen to be directly related to the form design. For example, in the case of form#10, there are a lot of small printed characters that were unrecognizable by the OCR but which were bigger than 8 pixels square size and, hence, close to some of the small handwriting character sizes. Consequently these components were not removed (see fig. 5-3 and fig. 5-4).

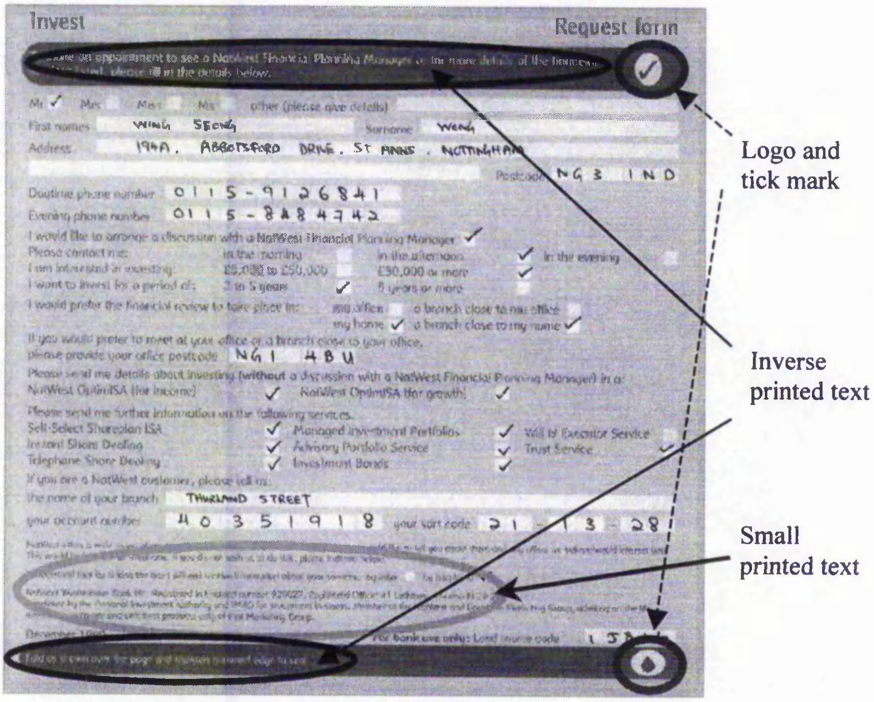


Fig. 5-3. Sample form#10, which contains many small machine printed text, inversely printed text and logo areas.

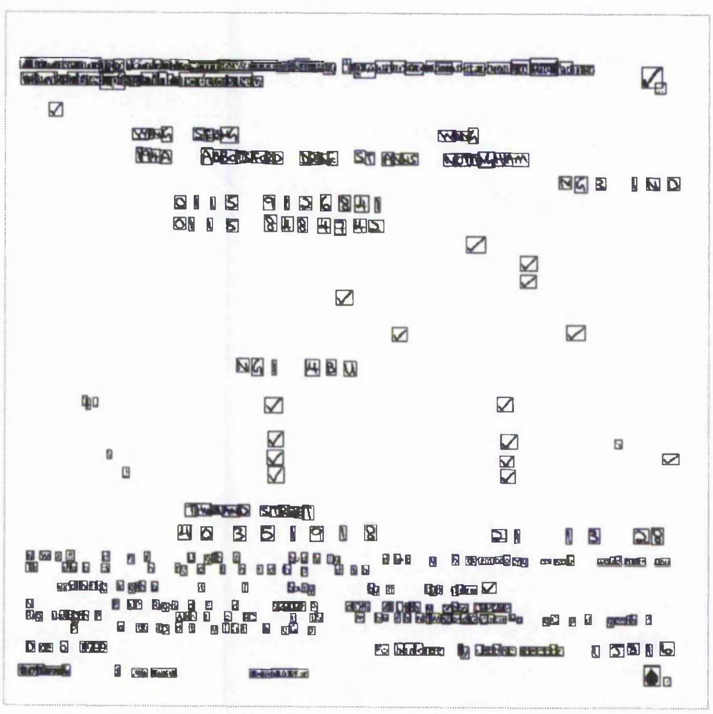


Fig. 5-4. The extracted output components for sample form#10 shown in fig. 5-3.

Similarly, in the case of form#7, a lot of the pre-printed text was inversely printed and small in size. The OCR failed to recognize these components leading to a very high recall rate and, therefore, a low precision rate.

Although the 70% precision rate seems low, this model-less extraction system can be considered a success if the recall rate is taken into account. As the recall rate is consistently over 100% for all of the cases, there will be no filled-in data loss after the removal process. In fact, the extra miss-extracted logo and small pre-printed text components can be further filtered by relying on some post-processing method such as cue word linking (discussed in the next section) and recognizer's confidence analysis (discussed in the future work section of chapter 6).

	Form 1		**Form 2**		**Form 3**	
	Precision (%)	**Recall** (%)	**Precision** (%)	**Recall** (%)	**Precision** (%)	**Recall** (%)
Writer01	93.4	107	96.5	104	99.4	100.6
Writer02	96.2	104	97.0	103	99.5	100.5
Writer03	95.7	104	98.5	102	100	100
Writer04	95.2	105	98.3	102	100	100
Writer05	92.0	109	96.5	104	100	100
Writer06	92.9	108	97.3	103	100	100
Writer07	94.0	106	98.4	102	100	100
Writer08	94.0	106	97.1	103	100	100
Writer09	93.5	107	97.7	102	100	100
Writer10	93.1	107	97.9	102	100	100
Overall	**94.1**	**106**	**97.5**	**103**	**99.9**	**100.1**

Table 5-2. Summary of the extraction recall and precision rate tested on 3 different types of form, each filled by 10 different writers, without using the blank form image as a reference.

Table 5-2 shows the extraction recall and precision rate tested on the second data set. Overall, the extraction precision rate is consistent across the different writers in a given type of form. In all the cases, the recall rate is more than 100%, which again means the system extracted more components than expected. Logos and miss-detected lines contributed almost 5% of the extra components in form 1. The experimental results confirm that the system extraction and recall rate is not affected by different handwriting styles, but is affected by the form design. This is inline with the conclusion made in [3] and previous chapters 3 & 4, where form design has been shown to change handwriting styles and affect the capturing accuracy.

5.4 Experiment II: Filled Word Retrieval Rate

This experiment aims to assess the effectiveness of retrieving the filled-in words using the cue words and linking algorithm in a model-less form-processing environment. Two new parameters were used to measure the system performance – retrieval accuracy and retrieval rate. The retrieval accuracy is defined as the percentage of **words that are retrieved correctly** over the total number of **words retrieved** by the system, whilst the retrieval rate is defined as the percentage of **words that are retrieved correctly** over the total number of **expected words** to be retrieved. In section 4.6, an experiment was conducted to measure the linking accuracy of the model-based extraction system. We can calculate the filled-in word retrieval and accuracy rate for the results obtained in section 4.6 by considering the number of correctly linked words as the number of correctly retrieved words and the total number of linked words plus the unlinked words as the total number of retrieved words expected.

Table 5-3 shows a comparison between the calculated filled-word retrieval accuracy and accuracy rate results for the model-based extraction system and the results obtained using the model-less extraction system on the first data set.

	Model-Based Extraction System					Model-less Extraction System				
	Retrieved words (correct)	Retrieved words (total)	Expected words	Accuracy Rate (%)	Retrieval Rate (%)	Retrieved words (correct)	Retrieved words (total)	Expected words	Accuracy Rate (%)	Retrieval Rate (%)
Form #1	34	34	34	**100**	**100**	34	34	34	**100**	**100**
Form #2	16	16	16	**100**	**100**	16	16	16	**100**	**100**
Form #3	27	27	27	**100**	**100**	27	27	27	**100**	**100**
Form #4	32	32	32	**100**	**100**	32	34	32	**94.0**	**100**
Form #5	32	35	35	**91.4**	**91.4**	32	35	35	**91.4**	**91.4**
Form #6	45	50	51	**90.0**	**88.2**	45	51	51	**88.2**	**88.2**
Form #7	29	32	36	**90.6**	**80.6**	29	52	36	**55.8**	**90.6**
Form #8	7	10	15	**70.0**	**46.7**	7	12	15	**58.3**	**46.7**
Form #9	28	29	31	**96.6**	**75.7**	28	34	31	**82.4**	**75.7**
Form #10	59	63	66	**93.7**	**89.4**	59	78	66	**75.6**	**89.4**
Form #11	72	80	81	**90.0**	**88.9**	72	83	81	**86.7**	**88.9**
Form #12	62	67	67	**92.5**	**92.5**	51	76	67	**66.6**	**76.1**
Form #13	50	53	56	**94.3**	**89.3**	50	58	56	**86.2**	**89.3**
Form #14	40	40	40	**100**	**100**	40	40	40	**100**	**100**
Form #15	22	22	22	**100**	**100**	22	25	22	**88.0**	**100**
Total	555	590	609	**94.1**	**91.1**	544	658	609	**82.7**	**89.3**

Table 5-3. Summary and comparison of the filled-in word retrieval accuracy rates using the cue words locating and linking methods in the model-based and modeless extraction system on the first test set data.

The experimental results show that when using the blank form image to perform the extraction and the cue word linking method to locate the filled-in words, 91.1% of the expected filled-in words will be retrieved successfully. The retrieval accuracy rate for the model-based system is also seen to be 94.1%, which means that on average there will be extra 6 words that are wrongly retrieved for every 100 retrieved words. These wrongly retrieved words being those filled-in words that were wrongly linked to incorrect cue words by the linking algorithm.

When the same data set was tested on the model-less extraction system, the retrieval rate drops from 91.1% to 89.3%. This is equivalent to another 1.8% of the expected filled-in words being lost. This extra 11 word loss is caused by the merging of unrecognized pre-printed components to filled-in words, causing them to be miss-treated as a single word during the character merging process (see fig. 5-5). The retrieval accuracy for this model-less system also drops from 94.1% to 82.7%. Thus, for every 100 words that the system retrieved, 17 of them are incorrect. As before, 6% of these errors are due to the linking algorithm itself whilst an additional 11% of errors were found to be due to the failure of the removal process where small machine printed text was unable to be recognized and removed by the OCR.

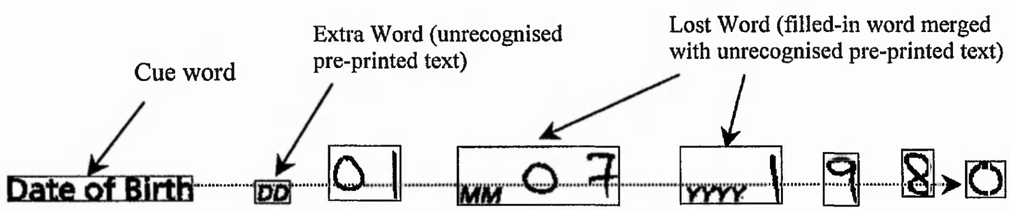


Fig. 5-5. An example of how unrecognized pre-printed text could be treated as a valid word or part of a filled-in word in a model-less extraction system.

It is worth pointing out that although the linking method failed to reject all of the non-filled-in components (logo and unrecognized pre-printed text), approximately 75% of these components were successfully filtered by the system. However, the magnitude of the retrieval and accuracy rate is seen to be dependent on the form designs. Form images that contain many small printed text and logos near to the filled-in areas (forms #7, #8, #12, etc), have a lower filled word retrieval rate. For form images that keep such components away from the filled-in area (form #10), then many of the miss-retrieved pre-printed entities are filtered by the linking rules.

Table 5-4 shows the comparison results between the calculated filled-word retrieval accuracy rate for the model-based extraction system and the results obtained using this model-less extraction system with the second data set.

	Form 1		Form 2		Form 3	
	Model-based Retrieval Accuracy (%)	Model-less Retrieval Accuracy (%)	Model-based Retrieval Accuracy (%)	Model-less Retrieval Accuracy (%)	Model-based Retrieval Accuracy (%)	Model-less Retrieval Accuracy (%)
Writer 1	98.6	97.2	100	94.3	100	100
Writer 2	95.7	95.7	100	93.8	100	100
Writer 3	98.4	95.2	100	93.9	100	100
Writer 4	100	100	100	93.5	100	100
Writer 5	93.2	93.2	98.0	90.0	100	100
Writer 6	98.0	90.9	100	93.0	100	100
Writer 7	100	100	100	93.7	100	100
Writer 8	100	97.8	100	90.7	100	100
Writer 9	100	95.2	100	93.5	100	100
Writer 10	94.7	94.7	100	93.3	100	100
Overall	97.9	96.0	99.8	93.1	100	100

Table 5-4. Summary and comparison of the filled-in word retrieval accuracy rates using the cue words locating and linking method in the model-based and model-less extraction system on the second test set data.

The experimental results further confirm that the filled-word retrieval accuracy rate is consistent over the 10 different writers samples for a given form. Note that the model-less retrieval accuracy for form 2 is lower than for form 1 despite the fact that form 2 has a lower number of extra components than form 1 (3% in form 2 compared to 6% in form 1). This is due to the fact that although there is higher number of extra components found in form 1 these components are further from the filled-in areas and the cue words than in form 2. Thus, when using the linking algorithm to locate the filled-in words, many of the extra components found in form 1 are discarded. In addition, there are 2 filled-in words lost in form 2 for writer 5 due to the merging of extra components into the filled-in word.

In total, there were 65 words retrieved incorrectly out of the 1509 word total retrieved from these 30 forms. Out of these 65 words, 12 of them are caused by linking errors, with the remaining being the unsuccessfully removed pre-printed components, of which 43 of them were miss-linked to the cue word number, 4 of them to date, 4 of them to price and 2 of them to town. It is obvious that if the same CSR test carried out in section 4.6 was performed again for this model-less system, then there would be virtually no difference at all in terms of CSR performance. This is due to the fact that only words that are suitable for the CSR process are selected for the test, meaning that, in this case, there will only be 2 extra words added to the test (i.e. the pre-printed components that were miss-linked to the cue word town). However, if we assume that all of these miss-linked words are fed to a recognizer and are wrongly recognized as a valid word/number string, then an additional 6.2% (53 wrongly retrieved words/807+53 total words tested) reduction in CSR performance would be introduced into the system. With the demonstrated average CSR improvement of 12.1% at top 1 position found in section 4.6 when using contextual aided recognition method, there will still be a clear 6% average improvement for using contextual focused recognition in the model-less system even when all these errors are taken into account.

	Form 1		Form 2		Form 3	
	Model-based Retrieval Rate (%)	Model-less Retrieval Rate (%)	Model-based Retrieval Rate (%)	Model-less Retrieval Rate (%)	Model-based Retrieval Rate (%)	Model-less Retrieval Rate (%)
Writer 1	90.9	90.9	100	100	96.1	96.1
Writer 2	81.1	81.1	100	100	95.5	95.5
Writer 3	89.6	89.6	100	100	95.0	95.0
Writer 4	96.6	96.6	100	100	95.0	95.0
Writer 5	80.4	80.4	98.0	98.0	94.9	94.9
Writer 6	87.7	87.7	100	100	95.0	95.0
Writer 7	78.8	78.8	100	100	95.7	95.7
Writer 8	91.8	91.8	100	100	95.3	95.3
Writer 9	93.7	93.7	100	100	95.1	95.1
Writer 10	73.5	73.5	100	100	95.0	95.0
Overall	87.2	87.2	99.8	99.8	97.6	97.6

Table 5-5. Summary and comparison of the filled-in word retrieval rates using the cue words locating and linking method in the model-based and model-less extraction system on the second test set data.

Table 5-5 shows the comparison results between the calculated filled-word retrieval rate for the model-based extraction system and the results obtained using this model-less extraction system with the second data set. The results are exactly the same for both cases. This is due to the fact that none of the extra components are wrongly merged to the filled-in words, thus there were no further filled words lost when using the model-less system rather than the model-based system. The retrieval rate is again seen to be form dependent and is fairly consistent within a same type of form. Overall, there are 88 filled-in words that were not retrieved using the cue word locating and linking method out of the 1585 total words expected. This is equivalent to an overall retrieval rate of 94.4% for both systems when using the cue word linking retrieval method.

5.5 Conclusion

This chapter has demonstrated how the methods developed in a model-based form processing system could be applied to a model-less form-processing environment. OCR software was used to remove the pre-printed text whilst vertical and horizontal lines were removed using an improved line detection algorithm. The remaining connected components were then merged together to form words using the character-merging algorithm developed in section 4.2. This was followed by utilizing the cue words and linking methods developed in section 4.3 to retrieve only the related filled-in words.

The first experimental results show that without using the blank form, the system extraction accuracy drops to around 70%. The extra errors are mainly caused by the unsuccessful removal of components such as small pre-printed text and logos that the OCR failed to recognize. On average, there are about 40% of extra components included in the extracted output. The number of extra components being extracted is directly related to the form design. Generally, the more small printed text and logos/pictures in a form, the lower the extraction accuracy and the higher the recall rate will be. The second experimental results show that by using the cue words and linking algorithm, 75% of these unwanted components can be successfully discarded. However, as there are still around 25% of the extra components that remain, the average retrieval accuracy drops from 94.1% in a model-based system to 82.7% for a model-less system. That said, the average retrieval rate of the model-less system is almost identical to the model-based system, with an overall decrease of only 1.8% (down from 91.1% to 89.3%). This decrease is due to the merging of the extra components to the filled-in words during the character merging process.

The later experimental results for the second data set show that the retrieval accuracy is independent of writer style. The average retrieval accuracy is 95.7% for the 30 forms, in which 0.8% is due to linking errors and 3.5% is the extra unrecognized pre-printed components that are miss-linked to telephone number, date, price and town. If we assume a worst-case scenario where all of these miss-linked words are wrongly recognized by the CSR as a valid word, then this amounts to an additional 6% of errors, which effectively halves the 12% CSR gain reported in section 4.7 when using the contextual focus recognition method. However, these errors could be further reduced by incorporating some post-processing methods such as recognizer confidence analysis to help eliminate the unwanted non-filled-in word components. This is discussed further in the future work section of chapter 6.

6. CONCLUSIONS & FUTURE WORK

This work is concerned with the extraction and recognition of filled-in data from colour form documents. The initial stage of the research investigated a method for using colour information to facilitate the filled data extraction process. The second stage of the research investigated techniques for improving the CSR performance through using the contextual knowledge contained within the form. The final stage of the research then investigated the feasibility of processing a filled form without using its blank form equivalent.

A novel colour handling technique has been presented that can reduce the colour content of an image to 8 colours or less. It was found that by reducing the colour content of an image to just a few colours, it is possible to extract the filled-in data directly using a software based colour dropout process. It has been shown that an extraction system that utilizes colour information will have a better extraction performance than a black and white extraction system, with an extraction speed of up to 3.22 times faster. In terms of extraction recall rate, the colour-based system produced an improvement of around 3% (up from 96.6% to ~99%), whilst the extraction accuracy improved from 97.2% to a maximum value of 99.9%. The colour based extraction system also outperformed the black and white system in terms of extracted text quality, in that the CSR performance was shown to improve from 49% to 58%. Overall, the results of this work have proven that colour can be successfully used to facilitate the extraction process and help improve the performance of a form processing system.

The novel concept of using OCR to aid CSR has also been demonstrated. It has been found that forms contain certain cue words that can be reliably located using OCR software and that the information so gained can be used to guide the CSR system. Several methods have been investigated to assist the association of the cue words to the filled-in words. A character merging technique was developed to group the related characters into words and a novel linking algorithm was proposed to identify the correct filled-in words for each of the cue words found in the form. The experimental results show that the proposed character merging technique was able to find an average of 90% of the handwritten words correctly across different writing styles and form design. 98% of the cue words were successfully located using the OCR software and an average of 94% of the linked words were associated correctly. However, it was found that these results are highly dependent on form designs, consequently, the linking accuracy can range from 70% to 100% and the character merging accuracy can range from 80% to 96%. The significance of this contextual knowledge linking was demonstrated by comparing the CSR rate of a contextual focused CSR method to that of a conventional CSR method. On average, the contextual information provided a 12% CSR rate improvement for top 1 word classification over a non-context aided CSR system.

A further study has been conducted to investigate the feasibility of using the developed methods to process a filled form without using a blank form image. The experimental results suggest that this can be done but then the extraction rate and CSR performance will be reduced; the magnitude of the reduction being highly dependent on the form design. Without the use of a blank form equivalent, the experimental results show that the average extraction accuracy will drop from 96% to 70% and that the recall rate will increase from 99% to 141%. As a result, the average filled-in word retrieval accuracy rate using the linking method, tested on 15 different types of form, dropped from 94.1% to 82.7% when compared to the model-based extraction system. However, most of this decrease is the result of miss-retrieved OCR text rather than non-retrieved handwritten data. Indeed, the actual CSR rates would only drop by around 1.8% due to additional loss of 11 filled-in words. That said, these results do suggest that further work needs to be done in order to improve the system reliability if a fully automated system is to be produced.

The methods presented here have already been published in the proceedings of 2 international conferences of high standing, denoting that they represent a significant contribution to the knowledge of the scientific community in the area of document analysis and recognition. The results of this work contain several key contributions. Firstly, the work has demonstrated that the inclusion of colour into a document processing system need not necessarily be considered as a burden to the system. It shows that when colour is handled effectively, the inclusion of colour can actually improve the overall system performance. Secondly, the work shows that it is possible to use a reliable recognition engine to help improve a relatively less reliable recognition process. With the ability of current OCR software, it is possible to reliably retrieve the contextual knowledge from a form document and use this information to help recognize the handwritten filled-in data. Lastly, the presented methods have also been shown to be effective even when the blank form image is not present to the system.

This concludes the PhD work into the issues surrounding form extraction and handwriting recognition using colour and context. The remaining sections provide suggestions that emanate from the experience gained whilst conducting this research and the findings that these works have evolved. The suggestions serve to link this PhD work with other potential projects in the future and bring a variety of insights/key ideas to other areas of concern. The central discussion will concentrate on how to expand the developed methods to improve on partially working and non-working cases, and how the findings may be applicable to other areas of application. Section 6.1 first discusses the major achievements and weaknesses of the developed techniques and their possible solutions. Section 6.2 then presents the possible areas of further investigation that could improve the system performance in order to realize a fully automated form processing system.

6.1 Possible Improvements

6.1.1 Colour Reduction

In section 3.1, a novel colour reduction method was introduced to reduce the colour domains of an image to 8 colours or less. This method has investigated the feasibility of using colour to extract the filled-in data from a form and has proved to be more effective than an equivalent black and white system in terms of speed, extraction accuracy, recall rate and extracted text quality. However, there is a weakness in this colour handling method – only the 8 pre-defined colours are allowed to be used to represent the image. Whilst it is true that a colour form can be represented by using just a few colours, the use of static colour domains is destined to failure under certain situations. For example, when the filled-in data has been filled-in with pencil (i.e. gray or silver colour). Although this only rarely happens (normally there is an instruction to ask the respondent to fill-in the form using a BLACK or BLUE ink ball point pen), it can happen. Similarly, the proposed colour reduction method will also fail to handle other colours such as: orange, purple, silver or gold etc. Most of the time, when a form containing one or more of such colours is presented to the system, noise will be generated. This was exactly the case for form#10 in

Experiment I section 5.3 (fig. 5-2) where a golden colour form was segmented across both the yellow and green colour domains, causing OCR errors on the machine printed text.

One of the possible methods to resolve this problem is to use dynamic colour reduction. Since we know that a form can be represented using just a few colour domains (as stated in section 3.1), we can dynamically determine the colours to be used to quantize a colour form after analyzing the original image colour distributions. The colour reduction technique proposed in section 3.1 was based on a comparison between the image pixels' RGB values and a set of threshold values that were fine-tuned to the scanner colour and brightness characteristics. Each of the pixels was then provisionally labeled to one of the 8 pre-defined colours. To convert these pre-defined colours to dynamic colours, a colour distribution analysis is suggested. Instead of labeling the pixels into one of the pre-defined colours, a colour distribution histogram could be generated. From this histogram, a maximum of 8 distinct colours peaks could be chosen as the possible pre-defined colours, depending on the percentage threshold. Based on the 'distance' between these 8 colours, two or more of these colours could be merged together (similar to the clustering method described in [8, 65]) and the remaining number of colours could then be used to quantize the image. Experimental work is needed to determine the optimum distance threshold values for different colours. Based on our previous experience, these distance threshold values will be different for each colour and the colour response characteristic of the scanner must be taken into consideration.

6.1.2 Word Identification

A character merging technique was developed in section 4.2 that can group the related characters together to form words for the holistic CSR system. Experimental results show that the method is capable of identifying the words correctly for more than 90% of the time. The 10% fallout is mainly due to the inconsistent gaps between words or characters and the overlapping of words from different fields. As shown in [166], high-level knowledge such as writing style, syntactic and semantic information could be used to help improve the segmentation system performance. Thus, it is believed that in the same context, the contextual information obtained could be used to help improve the segmentation performance. As demonstrated in experiment III section 4.6, touching words from different fields will tend to have more than one cue word linked to them. Thus, we can make use of this information to separate the words accordingly. Similarly, the cue word can also be used to help the merging process group the related characters into different groups of word in order to facilitate the recognition process (see fig 6-1).

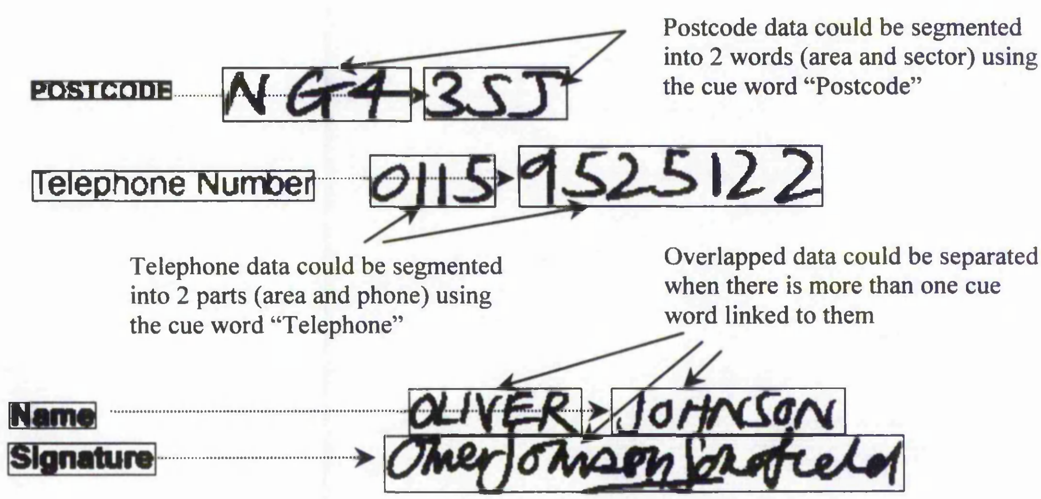


Fig. 6-1. Cue words could be used to merge characters to form words or segment touching words to facilitate the recognition process.

The contextual information can also be used as a guide to segment words into characters for character-based recognizers. For instance, segmenting a word into separate digits or numerals for the telephone number or post-code fields. This is especially useful if a discrete box design was used in a form. In this case, characters or numerals are usually well separated from each other. By counting the number of connected components for this field, the number of numerals or characters could be ascertained and any touching characters could be identified.

Besides using the cue words to identify the potential errors, lines can also be used to identify and segment joined words. If we assume that lines are used to define the filled-in areas, an extracted word that has a straight line at the middle of the word gives a strong indication of overlapping words. It is believed that by incorporating lines and contextual features into the merging and segmentation methods, the system will be more robust and a performance closer to 100% could be achieved.

6.1.3 Contextual Focused CSR

In section 4.7, an experiment has been reported which aims to evaluate the effect of the presence of contextual information on the recognition system. The recognizer that was chosen for the study is a holistic (word-based) recognizer. Only those fields containing word data that was suitable for the recognizer were chosen as the target data. Experimental results have shown that there is an average 12% increase for top 1 classification. However, it is clear that the improvement shown merely represents the performance gain for word data only thus, the experiment is somewhat restricted in its scope. The limitations of the holistic recognizer caused inflexibility in generating the field lexicons, resulting in a very limited (i.e. non-real life) application. For example, whilst it was possible to include all the country names into a lexicon, it was definitely not possible to include all person names. To investigate the full effect of the presence of contextual information on the recognition performance, the use of a character-based alphanumeric recognizer is suggested. Such a recognizer would not only provide flexibility in lexicon creation, it would also allow postcode or numeric recognition to be performed, thus allowing a complete picture of the overall performance gain, when using the contextual information, to be obtained. A detailed study is also needed to investigate whether different strategies are needed for different fields in order to utilize the contextual information to help improve the recognizer performance.

It is also believed that an intelligent document reading system could be employed to utilize the contextual information at a higher level to further improve the recognition system performance. For instance, by recognizing the postcode data, the search domain for the street and town name in the address field could be focused further. Similarly, by knowing the telephone number and the postcode, the address could be ascertained. Thus, by combining several cue words together, recognition confidence could be enhanced and better results could be obtained.

6.2 Future Work

This section presents the possible areas of further investigation that could link this work with other potential projects in the future. The suggestion given in the previous section (section 6.1) obviously also form possible future extension of this work, but have not been reiterated here to avoid unnecessary repetition.

6.2.1 Model-Based Form Processing

Fig. 6-2 shows the block diagram for the model-based system that has been developed in this work. There are several sub-processes that have yet to be fully investigated and implemented, namely: form recognition, form definition and form reconstruction and data recognition.

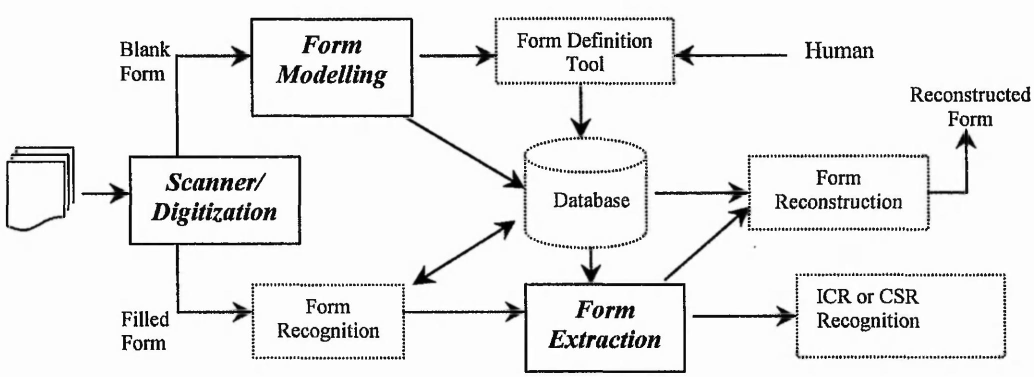


Fig 6-2. The model-based form processing system diagram that has been partially developed in this work. Solid boxes represent processes that have been fully developed in this work whilst dotted boxes represent processes yet to be fully investigated or implemented.

6.2.1.1 Form Recognition

As discussed in chapter 2, many techniques had been proposed to recognize a form document based on its line features. Potentially, colour features could also be used to facilitate the form recognition process. It is believed that by adding colour information to the process, form documents could possibly be recognized faster and more accurately. Colour features such as dominant colour and number of colours will provide useful information for discriminating between similarly structured forms that cannot be distinguished by the conventional methods.

6.2.1.2 Form Retrieval and Reconstruction

A form reconstruction process is the process of reconstructing the form from the extracted data for visualization purposes. This process can also be developed to be a form retrieval system. There is a possibility of using the methods developed in this work to facilitate the form retrieval process. Instead of using manual key word indexing, a form image could be automatically indexed by its colour, layout and data. This could be very useful especially when one is trying to locate a filled-in form according to its layout, features or even the filled-in data. For example, with such a system, it would be possible to retrieve accurately all forms that are filled-in by "John" (name) or by someone who lives in "Nottingham" (address).

6.2.1.3 Data Recognition

In this work, a contextual focused recognition method was introduced that can narrow down the search space and improve the recognition performance. However, besides constraining the search space of a recognizer, there are other ways of improving the recognition performance using the contextual knowledge. One of the possible methods is by using the contextual information to choose the best recognizer for each of the fields. Obviously, using a word-based recognizer to perform a numeric recognition is inappropriate and destined to failure. Therefore, it is strongly believed that by using the contextual information to choose the recognizer, the recognition performance can be further improved. Another possible method for utilizing the contextual information to improve the recognition performance is by employing a contextual weighted multi-expert system. In a multi-recognizer system, several recognizers are combined and a voting system is employed to decide which recognizers' results give the most likely answer. It is believed that since each recognizer has different strengths for different data, incorporating contextual information into the voting system will allow the recognizers to be weighted accordingly and hence the overall performance of the system improved.

6.2.2 Model-less Form Processing

Chapter 5 introduced the concept of processing a filled form without using any form template or blank form image. This novel idea could potentially eliminate the need for template creation; hence reducing the need for human intervention in the form processing system. However, in order to realize a fully automated form processing system, the model-less system performance must be improved. The experimental results in section 5.4 showed that approximately 90% of the filled-in words were successfully retrieved for a model-less system. Out of the 10% of miss-retrieved words, 6% of them were due to linking errors, whilst the remaining errors were due to the unavailability of suitable/recognized cue words. There is also a problem associated with the residual pre-printed text left behind after the OCR text removal process.

One of the possible ways to improve the retrieval rate is to employ a text type separation technique [167,168]. By distinguishing the handwriting data from the printed text, the filled-in data can be extracted. However, an investigation is needed to ensure the technique is robust to noise and font size. This is due to the fact that when the OCR software fails to recognize the printed text, there is a possibility that the printed text has lost its printed text characteristics. Thus, a careful study is needed to ensure that the text type separation method is performing better than the OCR software.

Alternatively, the filled-in data could be retrieved using the form line features. As lines are commonly used in a form to frame the layout of the document, there is a very high probability that the filled-in data is positioned right above the lines. Thus, by using the line information obtained in the line removal process, filled-in data can be further ascertained and retrieved. It is believed that by incorporating these line features into the cue word retrieval method, the retrieval rate could be further improved. It is also possible that this line information could help improve the cue word linking method.

Chapter 5 also demonstrated that the contextual aided recognition system would have a lower improvement gain in a model-less system due to the presence of noise (from the un-removed pre-printed text component). This can be resolved by employing some simple confidence thresholding. It was noticed that all of the noise that were mistakenly treated as valid words would have a very low scoring confidence. Thus, by using a CSR confidence threshold measure, all this noise could be effectively rejected. The CSR rates in the model-less system would then approach those in the model-based system.

6.2.3 Other Applications

Although the methods developed in this work were mainly focused on a form processing application, the developed methods could also be potentially used in other areas. Some of the possible areas are:

1. **Automated Assessment**

   By using the form extraction techniques developed in this work, cursively written question responses could be extracted. The cue words linking method could then be used to locate the answer (filled-in data) and by applying lexicon constraint to the extracted answer, obtain a scoring confidence value for that recognized response. Based on this scoring confidence, the recognized response could then be marked as 'correct', 'incorrect' or 'unsure'.

2. **Text Extraction from Web page and video images**

Due to the limitation of current Internet bandwidth, web page images often contain only limited colours. Similarly, for video images, text colour usually resides in one specific colour domain. This characteristic coincides with the colour form image characteristics, i.e. a form usually contains only a few colours and the text is often confined to a single colour domain. By using the colour reduction technique developed in this work, web page and video images could be segmented into a limited number of colour domains and the textual information could thus be retrieved. The proposed colour handling method could potentially outperform the methods proposed in [169,170], in which web/video image based textual data was usually segmented into several colour domains.

3. **Multi-Modal Form Processing**

With the rapid development in other recognition methods, a multi-modal form processing system that integrates speech, OCR, CSR and other recognition methods has become feasible. This work has demonstrated that OCR could be used to help improve the CSR performance. It is believed that by integrating other reliable recognition methods, such as speech recognition, it would be possible to improve the usability and reliability of a form processing system.

REFERENCES

1. Y.Y. Tang and J Liu, "Information Acquisition and Storage of Forms in Document Processing", Proceeding International Conference on Document Analysis and Recognition, pp. 170-174, 1997
2. Richard Casey, David Ferguson, K. Mohiuddin and Eugene Walach, "Intelligent Forms Processing System", Machine Vision and Applications, Vol. 5, pp. 143-155, 1992
3. Michael D Garris and Darrin L Dimmick, "Form Design for High Accuracy Optical Character Recognition", IEEE transactions on Pattern Analysis and Machine Intelligence, Vol 18, No. 6, pp. 653-656, June 1996.
4. Y. Zhong, K. Karu and A. K. Jain, "Locating Text In Complex Color Images", Pattern Recognition, Vol. 28, No. 10, pp. 1523-1535, 1995
5. W.Y. Chen and S.Y. Chen, "Adaptive Page Segmentation For Color Technical Journals' Cover Images", Image and Vision Computing, Vol. 16, pp. 855-877, 1998
6. A.K. Jain and B. Yu, "Automatic Text Location In Images and Video Frames", Pattern Recognition, Vol. 31, No. 12, pp. 2055-2076, 1998
7. H. Hase, T. Shinokawa, M. Yoneda, M. Sakai and H. Maruyama, "Character String Extraction from A Color Document", Proceeding 5th International Conference on Document Analysis and Recognition, pp. 75-78, 1999
8. H. Kasuga, M. Okamoto and H. Yamamoto, "Extraction of Characters From Color Documents", IS&T/SPIE 12th International Symposium On Electronic Imaging, pp278-285, 2000
9. H. Hase, T. Shinokawa, M. Yoneda, C.Y. Suen, "Character String Extraction From Color Documents", Pattern Recognition, Vol. 34, pp. 1349-1365, 2001
10. C. D. Yan, Y. Y. Tang and C. Y. Suen, "Form Understanding System Based on Form Description Language", Proceeding 1st International Conference on Document Analysis and Recognition, pp. 283-293, 1991
11. Y.Y. Tang, M. Cheriet, J. Liu, J.N. Said and C.Y. Suen, "Document Analysis and Recognition By Computers", Handbook of Pattern Recognition and Computer Vision, 2nd edition, chapter 8, World Scientific Publishing Company, 1998
12. Y.Y. Tang, C.D. Yan and C.Y. Suen, "Form Description Language and Its Mapping onto Form Structure", Technical Report, Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, 1990
13. P. Duygulu and Volkan Atalay, Ebru Dincel, "Logical Structure Representation of Form Documents based on Line Information", Technical Report TR-97-6, Department of Computer Engineering, Middle East Technical University, 1997

14. C. Cracknell, A.C. Downton and L. Du, "An Object-Oriented Form Description language and Approach to Handwritten Form Processing", Proceeding International Conference on Document Analysis and Recognition, pp. 180-184, 1997

15. Y.Y. Tang, C.D. Yan, M. Cheriet and C.Y. Suen, "Automatic Analysis and Understanding of Documents", Handbook of Pattern Recognition and Computer Vision, pp. 625-654, World Science Publishing Company, 1993

16. Y.Y. Tang, C.D. Yan, C.Y. Suen, "Document Processing For Automatic Knowledge Acquisition", IEEE Transaction on Knowledge and Data Engineering, Vol. 6, No. 1, pp. 3-21, 1994

17. P. Duygulu and Volkan Atalay, "A Hierarchical Representation of Form Documents for Identification and Retrieval", Document Recognition and Retrieval, SPIE Electronic Imaging, 2000

18. A. Ting and Maylor K.H. Leung, "Form Recognition Using Linear Structure", Pattern Recognition 32, pp. 645-656, 1999

19. L. Y. Tseng and R.C. Chen, "The Recognition of Form Documents Based on Three Type of Line Segments", Proceeding International Conference Document Analysis and Recognition, pp. 71-75, 1997

20. S. Shimotsuji and M. Asano, "Form Identification Based on Cell Structure", Proceedings of International Conference Pattern Recognition, pp. 793-797, 1996

21. Jon A. Pastor and Suzanne L. Taylor, "Recognizing Structured Forms Using Neural Networks", Proceedings of IJCNN, Seattle WA, July 1991.

22. R. Safari, N. Narasimhamurthi, M. Sshridhar and M. Ahmadi, "Form Registration : A Computer Vision Approach", Proceeding International Conference Document Analysis and Recognition, pp. 758-761, 1997

23. S. L. Taylor, Richard Fritzson and Jon A. Pastor, "Extraction of Data from Preprinted Forms", Machine Vision and Applications, Vol 5, No. 3, pp. 211-222, 1992

24. David Gans, "Transformations and Geometries", Appleton Century Crofts, New York, 1969

25. Y. Lamdan and H. J. Wolfson, "Geometric Hashing : A General and Efficient Model-Based Recognition Scheme", Proceedings of ICCV, pp. 238-249, 1988

26. Graham Leedham, and David Monger, "Evaluation of an Interactive Tool for Handwritten Form Description", Proceeding International Conference Document Analysis and Recognition, pp. 1185-1188, 1995

27. David Monger, Graham Leedham, and Andy Downton, "An Interactive Document Image Description for OCR of Handwritten Forms", Proceeding International Conference Document Analysis and Recognition, pp. 524-527, 1993

28. G. Dimauro, S. Impedovo, and G. Pirlo, "Document Analysis Systems: Problems and Perspectives, A Real Case Study", IEE publication, pp. 18/1-18/8, 1994.
29. J. Woods, "Recognition Technology for Data Entry – A Guide and Directory", Imagecapture Solutions Limited, Cimtech Ltd, ISBN 0-900458-68-2, pp. 33, 1995
30. C.C. Tappert, "Adaptive on-line Handwriting Recognition", IEEE 7th International Conference on Pattern Recognition, pp. 1004-1007, 1984
31. M.A. Just and P.A. Carpenter, "The Psychology of Reading and Language Comprehension", Allyn & Bacon Inc., Boston, 1987
32. N. Sherkat and T.J. Allen, "Whole Word Recognition in Facsimile Images", Proceedings International Conference on Document Analysis and Recognition, pp. 547-550, 1999
33. Wing Seong Wong, Nasser Sherkat and Tony Allen, "Use of Colour in Form Layout Analysis", Proceeding International Conference Document Analysis and Recognition, 2001
34. Wing Seong Wong, Nasser Sherkat and Tony Allen, "Contextual Focus for Improved Recognition of Hand-Filled Forms", Proceeding International Conference Document Analysis and Recognition, 2001
35. Wing Seong Wong, Nasser Sherkat and Tony Allen, "Form Extraction by Colour Information", 4th IAPR International Workshop on Document Analysis Systems, pp. 109-120, 2000.
36. L. Simoncini and Z.M. Kovacs-V, "A System for Reading USA Census'90 Hand-Written Fields", Proceedings International Conference on Document Analysis and Recognition", pp. 86-91, 1995
37. J. Mao, M. Abayan and K. Mohiuddin, "A Model-Based Form Processing System", Proceeding 13th International Conference on Pattern Applications, 5 (3), pp. 143-155, 1992
38. D. S. Doermann and A. Rosenfeld, "The processing of Form Documents", Proceedings International Conference on Document Analysis and Recognition", pp. 497-501, 1993
39. A. Ting, M.K.H. Leung, S. C. Hui and K. Y. Chan, "A Syntatic Business Form Classifier", Proceeding International Conference on Document Analysis and Recognition, pp. 301-304, 1995
40. A. Ting, M.K.H Leung, "Business Form Classification Using Strings", Proceeding 13th International Conference Pattern Recognition, Vol. 2, pp. 690-694, 1996
41. T. Watanabe, Q Luo and N. Sugie, "Layout Recognition of Multi-Kinds of Table-Form Documents", IEEE Transactions Pattern Analysis and Machine Intelligence", 17 (4), pp. 432-445, 1995
42. H.S. Baird, "The Skew Angle of Printed Documents", Proceeding of Society of Photographic Scientists and Engineers, Vol. 40, pp. 21-27, 1987

43. S.T. Hinds, J.L. Fisher and D.P. d'Amato, "A Document Skew Detection method Using Run-Length Encoding and Hough Transform", International Conference on Pattern Recognition, Vol. 1, pp. 464-468, 1990

44. A. Henig, G. Raza, N. Sherkat and R.J. Whitrow, "Detecting a Document's Skew: A Simple Stochastic Approach", Vision Interface'97, the 11th Canadian Conference on Computer Vision, Signal and Image Processing and Pattern Recognition, pp. 97-102, 1997

45. D. Wang and S.N. Srihari, "Analysis of Form Images", International Journal of Pattern Recognition and Artificial Intelligence, 8(5), 1031-1052, 1994

46. X. Y. Ye, M. Cheriet and C.Y. Suen, "A Generic System to Extract and Clean Handwritten Data from Business Forms", Proceedings 7th International Workshop on Frontiers in Handwriting Recognition, pp. 63-72, 2000

47. J. K. Guo and M. Y. Mo, "Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models", Proceedings International Conference on Document Analysis and Recognition", pp. 439-443, 2001

48. S. W. Lam, L. Javanbakht and S.N. Srihari, "Anatomy of a Form Reader", Proceedings International Conference on Document Analysis and Recognition", pp. 506-508, 1993

49. M. Okada and M. Shridhar, "A Morphological Subtraction Scheme for Form Analysis", Proceedings of International Conference on Pattern Recognition, pp. 190-194, 1996

50. J. Yuan, Y.Y. Tang and C.Y. Suen, "Four Directional Adjacency Graphs (FDAG) and their Applications in Locating Fields in Forms", Proceedings of International Conference on Pattern Recognition, pp. 752-755, 1995

51. B. Yu and A.K. Jain, "A Generic System For Form Dropout", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 18, No. 11, pp. 1127-1134, 1996

52. B. Yu and A.K. Jain, "A Form Dropout System", Proceedings of International Conference on Pattern Recognition, pp. 701-705, 1996

53. S. H. Kim, S. H. Jeong and H.K. Kwag, "Line Removal and Character Restoration Using BAG Representation of Form Images", Proceedings 7th International Workshop on Frontiers in Handwriting Recognition, pp. 43-52, 2000

54. L. Xingyuan, D. Doermann, W.G. Oh and W. Gao, "A Robust Method for Unknown Form Analysis", International Conference on Document Analysis and Recognition, pp. 531-533, 1999

55. L.Y. Tseng and R.C. Chen, "Recognition and Data Extraction of Form Documents Based on Three Types of Line Segments", Pattern Recognition, 31(10), pp. 1525-1540, 1998

56. J.L. Chen and H.J. Lee, "An Efficient Algorithm for Form Structure Extraction Using Strip Projection", Pattern Recognition, 31 (9), pp. 1353-1368, 1998

57. J.L. Chen and H.J. Lee, "A Novel Form Structure Extraction Method Using Strip Projection", Proceedings of International Conference on Pattern Recognition, pp. 823-827, 1996

58. K. Ueda, "Extraction of Signature and Seal Imprint from bankchecks by Using Color Information", Proceedings International Conference on Document Analysis and Recognition", pp. 665-668, 1995

59. J.P. Braquelaire and L. Brun, "Comparison and Optimisation of Methods of Color Image Quantization", IEEE Transactions on Image Processing, Vol. 6, No. 7, pp. 1048-1052, 1997

60. M.S. Shyu and J.J. Leou, "A Genetic Algorithm Approach to Color Image Enhancement", Pattern Recognition 31, No.7, pp. 871-880, 1998

61. C.Y. Yang and J.C. Lin, "Color Quantization by RWM-cut", Proceeding International Conference Document Analysis and Recognition, pp. 669-672, 1995

62. D.C. Tseng, Y.F. Li and C.T. Tung, "Circular Histogram Thresholding For Color Image Segmentation", Proceeding International Conference Document Analysis and Recognition, pp.673-676, 1995

63. D. Zugaj and V. Lattuati, "A New Approach Of Color Images Segmentation based On Fusing Region and Edge Segmentations Outputs", Pattern Recognition, Vol. 31, No. 2, pp. 105-113, 1998

64. N. Sherkat, Dhiraj Mighlani and R.J. Whithrow, "A Descriptive Retrieval Engine For Image Database", Proceedings of IEE colloquium on Intelligent Image databases, ISSN 0963-3308, pp. 11/1 – 11/6, 1997

65. M. Worring and L. Todoran, "Segmentation of Color Documents By Line oriented clustering using Spatial information", Proceeding International Conference Document Analysis and Recongition, pp. 67-70, 1999

66. C. Connolly and T. Fliess, "A Study of Efficiency and Accuracy in the transformation from RGB to CIELAB Color Space", IEEE Transactions On Image Processing, Vol. 6, no. 7, pp.1046-1048, 1997

67. M. Celenk, "A Colour Clustering Technique for Image Segmentation", Computer Vision, Graph, Image Processing, Vol. 52, pp. 145-170, 1990

68. Y.W. Lim and S.U. Lee, "On the Colour Image Segmentation Algorithm Based on the Thresholding and the Fuzzy C-Means Techniques", Pattern Recognition, 23 (9), pp. 935-952, 1990

69. C. Kotropoulos, E. Auge and I. Pitas, "Two-Layer Learning Vector Quantizer for Colour Image Quantization", Signal Processing IV: Theories and Applications, Elsevier, pp. 1177-1180, 1992

70. T. Kohonen, "Self-Organizing Maps", Springer Series in Information Sciences, Springer-Verlog, 30, 1995

71. John Woods, "Recognition Technology for Data Entry – A Guide and Directory", Cimtech ltd, ISBN 0-900458-68-2, pp. 54, 1995

72. K. Sayre, "Machine Recognition of Handwritten Words: A Project Report", Pattern Recognition, 5 (3), pp. 213-228, 1973

73. G. Casey and E. Lecolinet, "Strategies in Character Segmentation: A Survey, Proceeding 3rd International Conference on Document Analysis and Recognition, Vol. 1, pp. 1028-1033, 1995

74. Alessandro Vinciarelli, "A Survey On Off-Line Cursive Script Recognition", IDIAP Research Report, Switzerland, 2000

75. P.D. Gader, M. Mohamed and J.H. Chiang, "Handwritten Word Recognition With Character And Inter-Character Neural Networks", IEEE Transactions on System, Man and Cybernetics – Part B: Cybernetics, 27, pp. 158-164, 1997

76. G. Kim and V. Govindaraju, "Handwritten Word Recognition For Real-Time Applications", Proceedings International Conference on Document Analysis and Recognition, pp. 24-27, 1995

77. G. Kim and V. Govindaraju, "A Lexicon Driven Approach to Handwritten Word Recognition For Real-Time Application", IEEE Transaction on Pattern Analysis and machine Intelligence, 19 (4), pp. 366-379, 1997

78. M. Shridhar, G. Houle and F. Kimura, "Handwritten Word Recognition Using Lexicon Free and Lexicon Directed Word Recognition Algorithms", Proceedings International Conference Document Analysis and Recognition, pp. 861-865, 1997

79. J. T. Favata, "General Word Recognition using Approximate Segment-String Matching", Proceedings International Conference Document Analysis and Recognition, pp. 92-96, 1997

80. J.T. Favata and S. N. Srihari, "Recognition of handwritten Words Using a Hypothesis Generation and Reduction Methodology", Proceedings of U.S. Postal Service 5th Advanced Technology Conference, pp. 237-251, 1992

81. M.Y. Chen, A. Kundu and J. Zhou, "Off-Line handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network", IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (5), pp. 481-496, 1994

82. M.Y. Chen, A. Kundu and S. N. Srihari, "Variable Duration Hidden Markov Model and Morphological Segmentation for Handwritten Word Recognition", Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 600-601, 1993

83. Y. He, M.Y. Chen and A. Kundu, "Off-Line handwritten Word Recognition Using HMM with Adaptive Length Viterbi Algorithm", Proceedings International Conference on Pattern Recognition, pp. 460-462, 1994

84. A.W. Senior and A.J. Robinson, "An Off-Line Cursive Handwriting Recognition System", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (3), pp. 309-321, 1998

85. A. E. Yacoubi, J. M. Bertille and M. Gilloux, "Conjoined Location and Recognition of Street Names Within A Postal Address Delivery Line", Proceedings International Conference Document Analysis and Recognition, pp. 1024-1027, 1995

86. S. Madhvanath and V. Karpasundar, "Pruning Large Lexicons Using Generalized Word Shape Descriptors", Proceedings International Conference Document Analysis and Recognition, pp. 552-555, 1997

87. C. Parisse, "Global Word Shape Processing In Off-Line Recognition Of Handwriting", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18 (4), pp. 460-464, 1996

88. T. Paquet and Y. Lecourtier, "Handwriting Recognition: Application On Bank Cheques", Proceeding International Conference Document Analysis and Recognition, pp. 749-757, 1991

89. Andreas Hennig, "Recognizing a Page of Unconstrained Cursive Handwriting", PhD Thesis, The Nottingham Trent University, 1998

90. W.P. de Waard, "An Optimised Minimal Edit Distance For Handwriting Word Recognition", Pattern Recognition Letters, 16 (10), pp. 1091-1096, 1995

91. M. Leroux, J. C. Salome and J. Badard, "Recognition Of Cursive Script Words In A Small Lexicons", Proceedings International Conference Document Analysis and Recognition, pp. 774-782, 1991

92. H. Bunke, M. Roth and E. G. Schukat-Talamazzini, "Off-Line Recognition Of Cursive Script Produced By A Cooperative Writer", Proceedings International Conference on Pattern Recognition, pp. 383-386, 1994

93. H. Bunke, M. Roth and E. G. Schukat-Talamazzini, "Off-Line Handwriting Recognition Using Hidden Markov Models", Pattern Recognition, 28 (9), pp. 1399-1413, 1995

94. W. Cho, S.W. Lee and J.H. Kim, "Modelling and Recognition of Cursive Words With Hidden Markov Models", Pattern Recognition, 28 (12), pp. 1941-1953, 1995

95. D. Guillevic and C.Y. Suen, "HMM Word Recognition Engine", Proceedings International Conference Document Analysis and Recognition, pp. 544-547, 1997

96. M. Gilloux, M. Leroux and J.M. Bertille, "Strategies For Handwriting Words Recognition Using Hidden Markov Models", Proceedings International Conference Document Analysis and Recognition, pp. 299-304, 1993

97. M. Gilloux, M. Leroux and J.M. Bertille, "Recognition Of Handwriting Words In A Limited Dynamic Vocabulary", Proceedings International Workshop On Frontiers In Handwriting Recognition, pp. 417-422, 1993

98. I. Taylor and M. M. Taylor, "The Psychology of Reading", New York Academic, 1983

99. S. Edelman, T. Flash and S. Ullman, "Reading Cursive Handwriting By Alignment Of Letter Prototypes", International Journal Of Computer Vision, 5(3), pp. 303-331, 1990

100. M. Cote, E. Lecolinet, M. Cheriet and C.Y. Suen, "Building A Perception Based Model For Reading Cursive Script", Proceeding International Conference Document Analysis and Recognition, pp. 898-901, 1995

101. M. Cote, E. Lecolinet, M. Cheriet and C.Y. Suen, "Automatic Reading of Cursive Scripts Using Human Knowledge", Proceeding International Conference Document Analysis and Recognition, pp. 107-111, 1997

102. C. Faure, P. Keuss, G. Lorette, A. Vinter, (Eds), "Cursive Script Recognition by Backward Matching", Advances in Handwriting and Drawing: A Multidisciplinary Approach, pp. 117-135, 1994

103. R. K. Powalka, N. Sherkat and R. J. Whitrow, "The Use of Word Shape Information for Cursive Script Recognition", 4th International Workshop on Frontiers of Handwriting Recognition, 1994

104. K. Paap, S. Newsome, J. McDonald and R. Schvaneveldt, "An Activation-Verification Model For Letter And Word Recognition: The Word Superiority Effect", Psychological Review 89, pp. 573-594, 1982

105. M.D. Garris, J.L. Blue, G.T. Candela and D.L. Dimmick, "Off-Line Handwriting Recognition from Forms", International Conference on System, Man and Cybernetics: Intelligence System for the 21st Century, pp. 2783-2788, 1995

106. A. C. Downton, S. J. Hanlon and A. Amiri, "A Visual Programming Toolkit Demonstrator for Offline Handwritten Forms Recognition", Proceeding International Conference Document Analysis and Recognition, pp. 707-710, 1995

107. B. Verma, M. Blumenstein and S. Kulkarni, "Recent Achievements in Off-Line Handwriting Recognition Systems", International Conference on Computational Intelligence and Multimedia Application, Australia, pp. 27-33, 1998

108. M. L. Simner, C. G. Leedham, A. J. M. Twomassen (Eds), "Cursive Script Recognition: A Survey", Hanwriting and Drawing Research: Basic and Applied Issues, pp. 267-284, 1996

109. N. Gorski, V. Anisimov, E. Augustin, O. Baret, D. Price and J. Simon, "A2iA Check Reader: A Family Of Bank Check Recognition Systems", Proceeding International Conference Document Analysis and Recognition, pp. 523-526, 1999

110. E. Letherlier, M. Leroux and M. Gilloux, "An automatic Reading System for Handwritten Numeral Amounts on French Checks", Proceeding International Conference Document Analysis and Recognition, pp. 15-18, 1995

111. Wang Song, M. Feng and X. Shaowei, "A Chinese Bank Check Recognition System Based on the Fault Tolerant Technique", Proceedings International Conference Document Analysis and Recognition, pp. 1038-1042, 1997

112. P. Pereira, L. Heutte, O. Bougeois, J. V. Moreau, (Eds), "Numeral Amount Recognition on Multi-Bank Checks", IEEE Proceedings of International Conference on Pattern Recognition, pp. 165-169, 1996

113. L. S. Oliveira, R. Sabourin, F. Bortolozzi and C. Y. Suen, "A Modular System to Recognize Numerical Amounts on Brazilian Bank Cheques", Proceedings International Conference on Document Analysis and Recognition", pp. 389-394, 2001

114. G. Dimauro, M. R. Grattagliano, S. Impedovo and G. Pirlo, "A System for Bankchecks Processing", Proceedings International Conference Document Analysis and Recognition, pp. 454-459, 1993

115. S. Kelland and S. Wesolkowski, "A Comparison of Research and Production Architectures for Check Reading Systems", Proceedings International Conference Document Analysis and Recognition, pp. 99-102, 1999

116. M. Morita, A. El Yacoubi, R. Sabourin, F. Bortolozzi and C. Y. Suen, "Handwritten Month Word Recognition on Brazilian Bank Cheques", Proceedings International Conference on Document Analysis and Recognition", pp. 972-976, 2001

117. J.P. Dodel and R. Shinghal, "Symbolic/Neural Reconition of Cursive Amounts on Bank Cheques", Proceedings International Conference Document Analysis and Recognition, pp. 15-18, 1995

118. D. Guillevic and C. Y. Suen, "Cursive Script Recognition Applied To The Processing Of Bank Cheques", Proceedings International Conference Document Analysis and Recognition, pp. 11-14, 1995

119. A. Kornai, K.M. Mohiuddin and S.D. Connell, "Recognition of Cursive Writing on Personal Checks", 5th International Workshop on the Frontiers of Handwriting Recognition, pp. 617-622, 1996

120. V. Di Leece, G. Dimauro, A. Guerriero, S, Impedovo, G. Pirlo and A. Salzo, "A New Hybrid Approach For Legal Amount Recognition", 7th International Workshop on Frontiers in Handwriting Recognition", pp. 199-208, 2000

121. G. Saon and A. Belaid, "High Performance Unconstrained Word Recognition System Combining HMMs and Markov Random Field", World Science Publication, Singapore, pp. 309-326, 1997

122. U. Mahadevan and S. N. Srihari, "Parsing and Recognition of City, State and Zip Codes in Handwritten Addresses", Proceeding International Conference Document Analysis and Recognition, pp. 325-328, 1999

123. S.H. Kim, S. Jeong and C.Y. Suen, "A Lexicon Driven Approach for Off-Line Recognition of Unconstrained Handwritten Korea Words", International Conference on Pattern Recognition, pp. 507-510, 1999

124. C.K. Lee and G. Leedham, "Empirical Design of a Holistic Verifier for Automatic Sorting of Handwriten Singapore Postal Address", Proceeding International Conference Document Analysis and Recognition, pp. 733-736, 1999

125. N.W. Strathy AMDG and C.Y. Suen, "A New System for Reading Handwritten Zip Codes", Proceeding International Conference Document Analysis and Recognition, pp. 74-77, 1995

126. E. Lecolinet and J.V. Moreau, "Off-Line Recognition of Handwritten Cursive Script for the Automatic Reading of City Names on Real Mail", International Conference on Pattern Recognition, pp. 674-676, 1990

127. S. N. Srihari, Y.C. Shin, V. Ramanaprasad and D.S. Lee, "Name and Address Reader System for Tax Form Processing", Proceeding International Conference Document Analysis and Recognition, pp. 5-10, 1995

128. T.A. Bayer and H.U. M-Schneider, "A Generic System For Processing Invoices", Proceeding International Conference Document Analysis and Recognition, pp. 740-744, 1997

129. D.A. Kosiba and R. Kasturi, "Automatic Invoice Interpretation: Invoice Structure Analysis", Proceedings International Conference on Pattern Recognition, pp. 721-725, 1996

130. J. Mao, R. Lorie and K. Mohiuddin, "A System for Automatic Reading IATA Flight Coupons", Proceeding International Conference Document Analysis and Recognition, pp. 153-157, 1997

131. J. Paik, S. Jung and Y. Lee, "Multiple Combined Recognition System for Automatic Processing of Credit Card Slip Applications", Proceeding International Conference Document Analysis and Recognition, pp. 520-523, 1993

132. F. Grandidier, R. Sabourin, A. E. Yacoubi, M. Gilloux and C.Y. Suen, "Influence of Word Length on Handwriting Recognition", Proceeding International Conference Document Analysis and Recognition, pp. 777-780, 1999

133. F. Kimura, M. Shridhar and Z. Chen, "Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words", Proceeding International Conference Document Analysis and Recognition, pp. 18-22, 1993

134. A. P. Whichello and H. Yan, "Linking Broken Character Borders With Variable Sized Masks to Improve Recognition", Pattern Recognition, 29 (8), 1996

135. J. Chai and Z. Q. Liu, "Off-Line Unconstrained Handwritten Word Recognition", International Journal of Pattern Recognition and Artificial Intelligence, 14 (3), pp. 259-280, 2000

136. Z. M. Kovacs-V, "A Novel Architecture For High Quality Hand-printed Character Recognition", Pattern Recognition, 28(11), pp. 1685-1692, 1995

137. A. Vinciarelli and J. Luettin, "Off-Line Cursive Script Recognition Based on Continuous Density HMM", Proceedings 7th International Workshop on Frontiers in Handwriting Recognition, pp. 493-498, 2000

138. M. C. Fairhurst and A.F.R. Rahman, "Generalised Approach to the Recognition of Structurally Similar Handwritten Characters Using Multiple Expert Classifiers", IEE Proceedings Vision Image Signal Process, Vol. 144, No. 1, pp. 15-22, 1997

139. A. Kosmala, J. Rottland and G. Rigoll, "Improved On-Line Handwriting Recognition Using Context Dependent Hidden Markov Models", Proceeding International Conference Document Analysis and Recognition, pp. 641-644, 1997

140. A. Kosmala, J. Rottland and G. Rigoll, "Large Vocabulary On-Line Handwriting Recognition With Context Dependent Hidden Markov Models", Advances in Handwriting Recognition, chapter 2, World Scientific, 1999

141. G.J. Bellaby, L.J. Evett and R.K. Powalka, "Coping with Ambiguity and Error in Cursive Script Recognition", Proceedings International Workshop on Frontiers in Handwriting Recognition, 1996

142. T. Caesar, J.M. Gloger and E. Mandler, "Using Lexical Knowledge for the Recognition of Poorly Written Words", Proceeding International Conference Document Analysis and Recognition, pp. 915-918, 1995

143. S. Madhvanath, V. Krpasundar and V. Govindaraju, "Syntactic Methodology of Pruning Large Lexicons in Cursive Script Recognition", Patter Recognition 34, pp. 37-46, 2001

144. G. Seni, R.K. Srihari and N. Nasrabadi, "Large Vocabulary Recognition of On-Line Handwritten Cursive Words", IEEE Transaction on Pattern Analysis and Machine Intelligence, 18(7), pp. 757-762, 1996

145. T.G. Rose and L.J. Evett, "Semantic Analysis for Large Vocabulary Cursive Script Recognition", Proceeding International Conference Document Analysis and Recognition, pp. 236-239, 1993

146. D. Bouchaffra, E. Koontz, V. Krpasundar and R.K. Srihari, "Incorporating Diverse Information Sources in Handwriting Recognition Post-processing", International Journal of Imaging Systems and Technology, Vol. 7, Issue 4, 1996

147. Steven W. Smith, "The Scientist and Engineer's Guide to Digital Signal Processing", California Technical Publishing, chapter 23, ISBN 0-9660176-3-3, 1997

148. Michael D. Garris, "Correlated Run Length Algorithm (CURL) for Detecting Form Structure Within Digitized Documents", 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 415-424, UNLV, April 1994.

149. Jaekyu Ha, Ihsin T. Phillips, and Robert M. Haralick, "Document Page Decomposition Using Bounding Boxes of Connected Components of Black Pixels", SPIE Vol. 2422, pp. 140-151, 1995.

150. Juan F. Arias, Atul Chhabra and Vishal Misra, "Finding Straight Lines in Drawing", IEEE Proceedings International Conference Document Analysis and Recognition, pp. 788-791, 1997

151. Opas Chutatape and Linfeng Guo, "A Modified Hough Transform For Line Detection and Its Performance", Pattern Recognition 32, pp. 181-192, 1999

152. Bin Kong, Su Chen and Robert M. Haralick, "Automatic Line Detection in Document Images Using Recursive Morphological Transforms", SPIE Vol. 2422, pp.163-174, 1995

153. Yi Kai Chen and Jhing Fa Wang, "Skew Detection and Reconstruction Based on Maximization of Variance of Transition-counts", Pattern Recognition 33, pp. 195-208, 2000

154. Yonki Min, Sung Bae Cho and Yillbyung Lee, "A Data Reduction for Efficient Document Skew Estimation Based on Hough Transformation", Proceedings of ICPR, pp. 732-736, 1996

155. Nadine Rondel and Gilles Burel, "Cooperation of Multi-Layer Perceptions for Estimation of Skew Angle in Text Document Images", 3rd Proceeding International Conference on Document Analysis and Recognition, pp. 1141-1144, 1995

156. A. Hennig, "Recognising a Page of Unconstrained Cursive Handwriting", PhD thesis, Department of Computing, The Nottingham Trent University, 1999

157. Huei Fen Jiang, Chin Chuan Han and Kuo Chin Fan, "A Fast Approach to Detect and Correct Skew Documents", Proceedings of ICPR, pp. 742-746, 1996

158. Avanindra and Subhasis Chaudhuri, "Robust Detection of Skew in Document Images", IEEE Transactions on Image Processing, Vol. 6, No. 2, pp. 344-349, 1997

159. Y. Chung, K. Lee, J. Paik and Y. Lee, "Extraction and Restoration of Digits Touching or Overlapping Lines", IEEE Proceedings of ICPR, pp. 155-159, 1996

160. C. Lagoze, E. Shaw, J.R. Davis and D.B. Krafft, "Implementation Reference Manual", Technical Report, Cornell University, 1995

161. K. Tsuda, S. Senda, M. Minoh and K. Ikeda, "Clustering OCR-ed Texts for Document Image Database", Proceeding International Conference Document Analysis and Recognition", pp. 171-174, 1995

162. F. Esposito, D. Malerba and G. Semeraro, "Automated Acquisition of Rules for Document Understanding", IEEE Proceedings, pp. 650-654, 1993

163. John Woods, "Recognition Technology for Data Entry", ISBN 0-900458-68-2, Cimtech Ltd, pp. 22, 1995

164. R. Hoch and T. Kieninger, "On Virtual Partitioning of Large Dictionaries for Contextual Post-Processing to Improve Character Recognition", Proceeding International Conference Document Analysis and Recognition, pp. 226-231, 1993

165. S. Madhvanath and S.N. Srihari, "Effective Reduction of Large Lexicons for Recognition of Offline Cursive Script", 5th Int. Workshop on Frontiers in Handwriting Recognition, Essex, England, pp. 445-452, 1996

166. Q. Xu, L. Lam and C. Y. Suen, "A Knowledge-based Segmentation System for Handwritten Dates on Bank Cheques", Proceedings International Conference on Document Analysis and Recognition", pp. 384-388, 2001

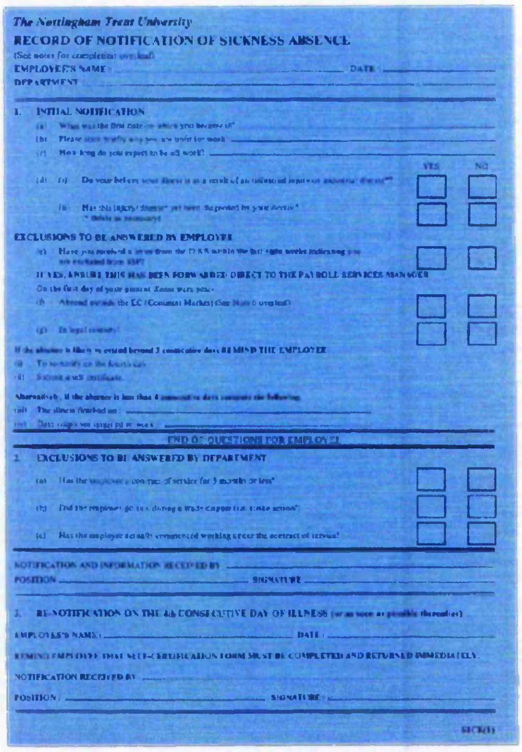
167. U. Pal and B.B. Chaudhuri, "Automatic Separation of Machine-Printed and Hand-Written Text Lines", Proceeding International Conference Document Analysis and Recognition, pp. 645-648, 1999

168. J. Franke and M. Oberlander, "Writing Style Detection by Statistical Combination of Classifiers in Form Reader Applications", Proceeding International Conference Document Analysis and Recognition, pp. 581-584, 1993

169. S. Antani, D. Crandall and A. Narasimhamurthy, V. Y. Mariano, R. Kasturi, "Evaluation of Methods for Detection and Localization of Text in Video", 4th IAPR International Workshop on Document Analysis Systems, pp. 507-514, 2000

170. A. Antonacopoulos and D. Karatzas, "An Anthropocentric Approach to Text Extraction from WWW Images", 4th IAPR International Workshop on Document Analysis Systems, pp. 515-526, 2000

APPENDIX A – 15 Blank Forms for OCR performance test



The Nottingham Trent University

RECORD OF NOTIFICATION OF SICKNESS ABSENCE

EMPLOYEE'S NAME : DATE :

DEPARTMENT

1. INITIAL NOTIFICATION

EXCLUSIONS TO BE ANSWERED BY EMPLOYEE

END OF QUESTIONS FOR EMPLOYEE

2. EXCLUSIONS TO BE ANSWERED BY DEPARTMENT

NOTIFICATION AND INFORMATION RECEIVED BY

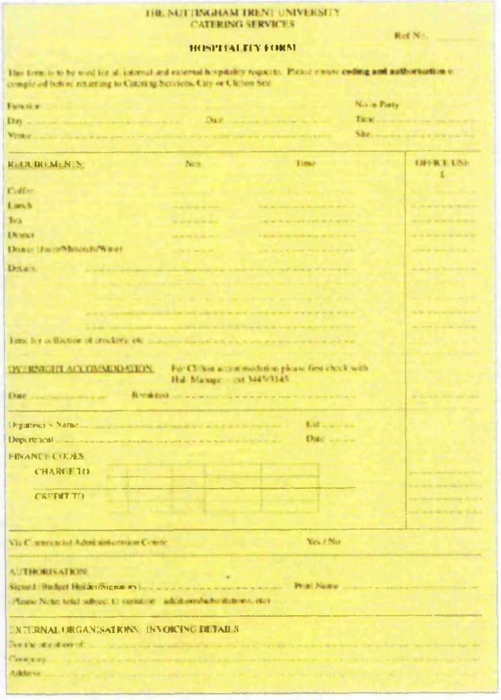
POSITION SIGNATURE

3. RE-NOTIFICATION ON THE 4th CONSECUTIVE DAY OF ILLNESS

EMPLOYEE'S NAME : DATE :

NOTIFICATION RECEIVED BY

POSITION : SIGNATURE :



THE NOTTINGHAM TRENT UNIVERSITY
CATERING SERVICES

HOSPITALITY FORM

REQUIREMENTS

OVERNIGHT ACCOMMODATION

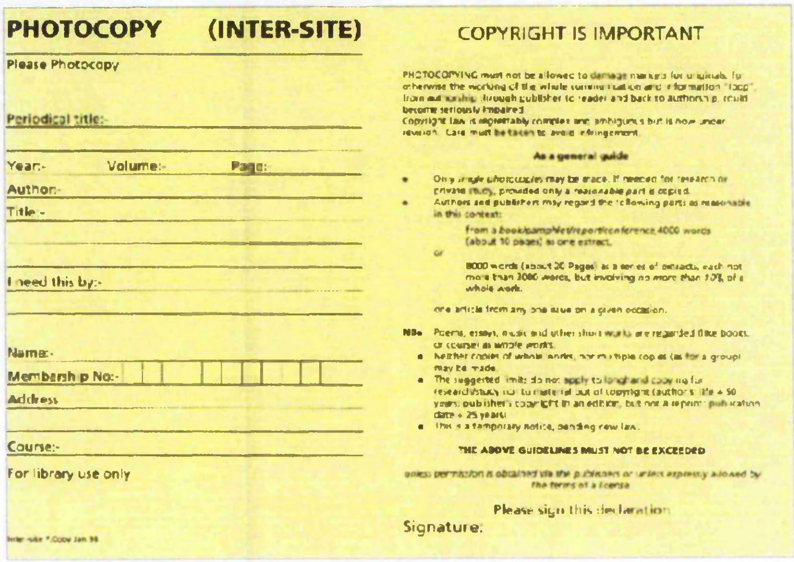
FINANCE CODES

CHARGE TO

CREDIT TO

AUTHORISATION

EXTERNAL ORGANISATIONS, INVOICING DETAILS



PHOTOCOPY (INTER-SITE)

Please Photocopy

Periodical title:-

Year:- Volume:- Page:

Author:-

Title -

I need this by:-

Name:-

Membership No:-

Address

Course:-

For library use only

COPYRIGHT IS IMPORTANT

As a general guide

THE ABOVE GUIDELINES MUST NOT BE EXCEEDED

Please sign this declaration

Signature:

ORDER FORM **Your details in BLOCK CAPITALS please**

Name Student Teacher

Address for delivery

Post code

Telephone number

Email address

Degree Course

School/College/University

Please supply **Cost**

Microsoft Windows 95/NT	£50	£	
Microsoft Office	£129	£	
Microsoft Visual Studio	£129	£	
Microsoft Publisher	£59	£	
Microsoft Works	£39	£	
✔ Post and Packing	FREE	£	FREE
All prices shown include VAT	Total cost	£	

Method of payment

Cheque payable to On-Line Computers please attach to this order

Credit/debit card

Card type e.g. VISA, MasterCard, Switch

Card number Expiry date

Switch issue number (where applicable)

Name on card

Credit card address if different from delivery address

Cardholder's signature

ON-LINE COMPUTERS

FREEPOST · LANCASTER LA1 1ZZ

TEL 01534 33366 · FAX 39666

Email sales@olc.co.uk

ITEM SEARCH AND PROGRESS REPORT

Author ....

Title ....

.... Edition ....

Shelf No. .... Barcode No.

Name ....

Borrower No.

Latest date of use ....

The Nottingham Trent University Library & Information Services

Item Search & Prog Jan 98

East London Assessment Group Trial Spelling Test
Test C (Version 2b)

Name

School

Date of birth

1		2	
3		4	
5		6	
7		8	
9		10	
11		12	
13		14	
15		16	
17		18	

S1

SECTION C Yes No

12. Have you received a "linking" letter from the D.S.S. within the last eight weeks indicating you are excluded from SSP?

IF YES, ENSURE THIS HAS BEEN FORWARDED DIRECT TO THE PAYROLL SERVICES MANAGER

13. (i) Were you abroad outside the EC (Common Market) on the first day of illness?

AND

(ii) If you have been abroad outside the EC during your present illness please give the date on which you first entered a non-EC country ______ 19____

14. (i) Were you in legal custody on the first day of illness?

AND

(ii) If you were taken into legal custody during your present illness please give the date. ______ 19____

SECTION D (see paragraph (n) over)

EITHER

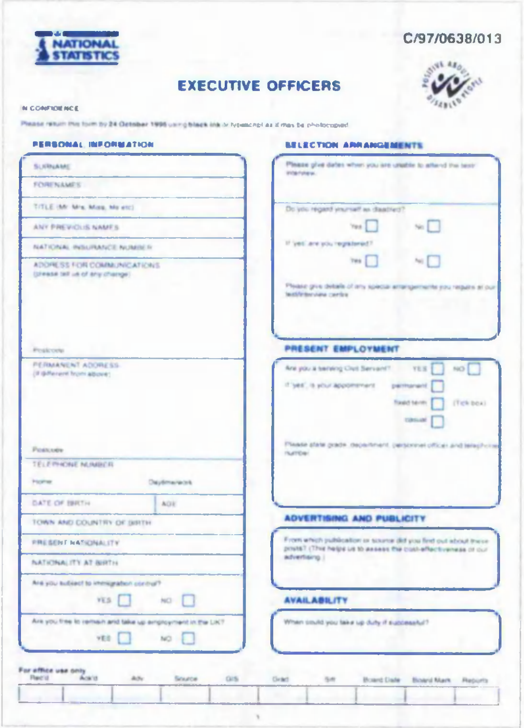
(i) I declare that I have not worked during the period of sickness stated and to the best of my knowledge the information given is correct.

OR

(ii) On behalf of the employee named above, I declare that he/she has not worked during the period of sickness stated and that to the best of my knowledge the information given is correct.

Signed ______ Date ______

Tick the box if you have signed on behalf of the employee named above.


NATIONAL STATISTICS

C/97/0638/013

EXECUTIVE OFFICERS

IN CONFIDENCE

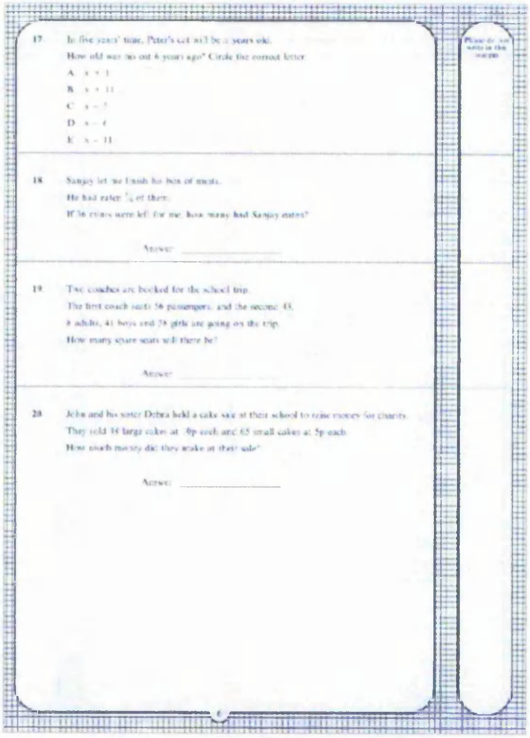
PERSONAL INFORMATION

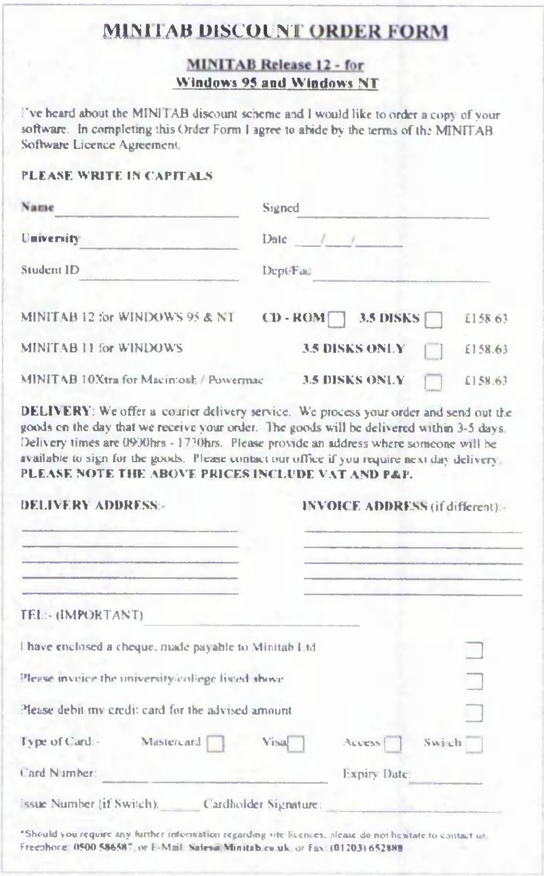
SELECTION ARRANGEMENTS

PRESENT EMPLOYMENT

ADVERTISING AND PUBLICITY

AVAILABILITY




MINITAB DISCOUNT ORDER FORM

MINITAB Release 12 - for Windows 95 and Windows NT

I've heard about the MINITAB discount scheme and I would like to order a copy of your software. In completing this Order Form I agree to abide by the terms of the MINITAB Software Licence Agreement.

PLEASE WRITE IN CAPITALS

Name ________ Signed ________

University ________ Date __/__/____

Student ID ________ Dept/Fac ________

MINITAB 12 for WINDOWS 95 & NT CD - ROM ☐ 3.5 DISKS ☐ £158.63

MINITAB 11 for WINDOWS 3.5 DISKS ONLY ☐ £158.63

MINITAB 10Xtra for Macintosh / Powermac 3.5 DISKS ONLY ☐ £158.63

DELIVERY: We offer a courier delivery service. We process your order and send out the goods on the day that we receive your order. The goods will be delivered within 3-5 days. Delivery times are 0900hrs - 1730hrs. Please provide an address where someone will be available to sign for the goods. Please contact our office if you require next day delivery. **PLEASE NOTE THE ABOVE PRICES INCLUDE VAT AND P&P.**

DELIVERY ADDRESS:- INVOICE ADDRESS (if different):-

TEL:- (IMPORTANT) ________

I have enclosed a cheque, made payable to Minitab Ltd ☐

Please invoice the university/college listed above ☐

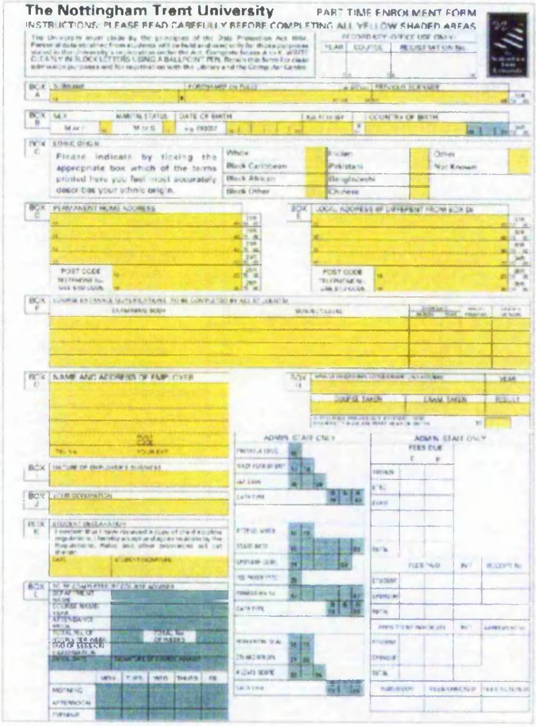
Please debit my credit card for the advised amount ☐

Type of Card:- Mastercard ☐ Visa ☐ Access ☐ Switch ☐

Card Number: ________ Expiry Date: ________

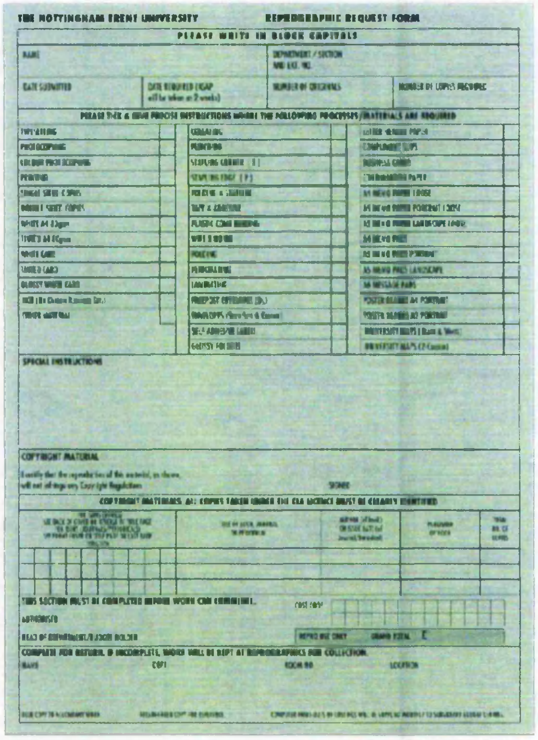
Issue Number (if Switch): ______ Cardholder Signature: ________

*Should you require any further information regarding site licences, please do not hesitate to contact us. Freephone: 0500 586587, or E-Mail: Sales@Minitab.co.uk, or Fax: (01203) 652888


The Nottingham Trent University

PART TIME ENROLMENT FORM

INSTRUCTIONS: PLEASE READ CAREFULLY BEFORE COMPLETING ALL YELLOW SHADED AREAS



THE NOTTINGHAM TRENT UNIVERSITY — REPROGRAPHIC REQUEST FORM

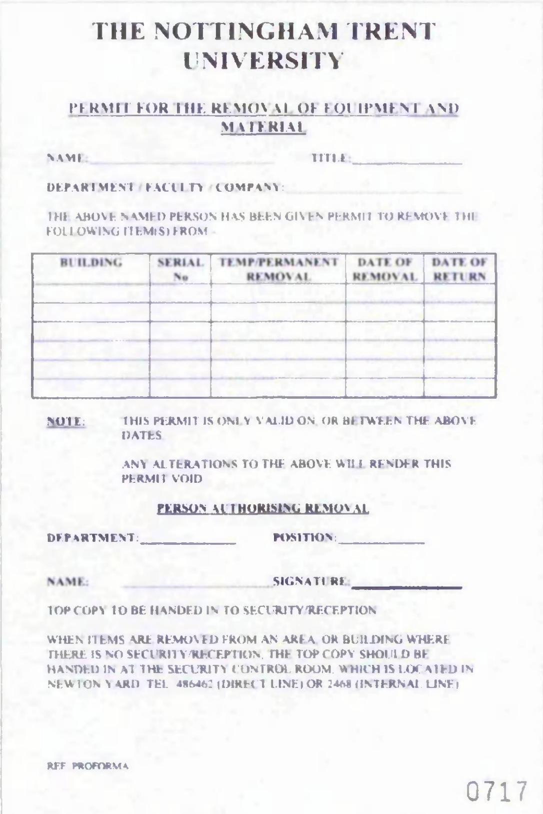
PLEASE WRITE IN BLOCK CAPITALS

NAME — DEPARTMENT / SECTION AND EXT. NO.

DATE SUBMITTED — DATE REQUIRED (ASAP will be taken as 2 weeks) — NUMBER OF ORIGINALS — NUMBER OF COPIES REQUIRED

SPECIAL INSTRUCTIONS

COPYRIGHT MATERIAL



THE NOTTINGHAM TRENT UNIVERSITY

PERMIT FOR THE REMOVAL OF EQUIPMENT AND MATERIAL

NAME: ______________________ TITLE: __________

DEPARTMENT / FACULTY / COMPANY: ______________

THE ABOVE NAMED PERSON HAS BEEN GIVEN PERMIT TO REMOVE THE FOLLOWING ITEM(S) FROM -

BUILDING	SERIAL No	TEMP/PERMANENT REMOVAL	DATE OF REMOVAL	DATE OF RETURN

NOTE: THIS PERMIT IS ONLY VALID ON, OR BETWEEN THE ABOVE DATES.

ANY ALTERATIONS TO THE ABOVE WILL RENDER THIS PERMIT VOID

PERSON AUTHORISING REMOVAL

DEPARTMENT: __________ POSITION: __________

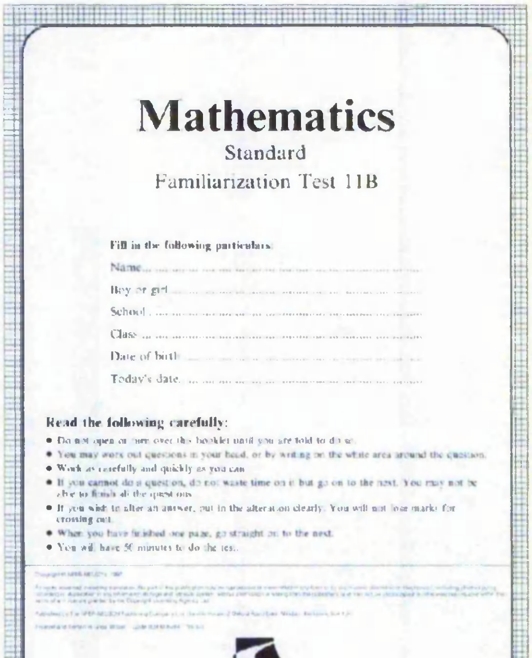
NAME: __________ SIGNATURE: __________

TOP COPY TO BE HANDED IN TO SECURITY/RECEPTION

WHEN ITEMS ARE REMOVED FROM AN AREA, OR BUILDING WHERE THERE IS NO SECURITY/RECEPTION, THE TOP COPY SHOULD BE HANDED IN AT THE SECURITY CONTROL ROOM, WHICH IS LOCATED IN NEWTON YARD. TEL. 486462 (DIRECT LINE) OR 2468 (INTERNAL LINE)

REF. PROFORMA

0717



Mathematics

Standard

Familiarization Test 11B

Fill in the following particulars:

Name ....

Boy or girl ....

School ....

Class ....

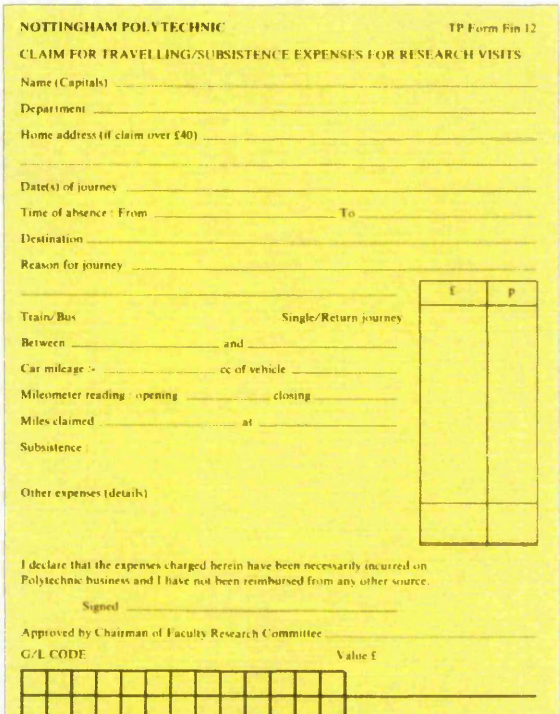
Date of birth ....

Today's date ....

Read the following carefully:

- Do not open or turn over this booklet until you are told to do so.
- You may work out questions in your head, or by writing on the white area around the question.
- Work as carefully and quickly as you can
- If you cannot do a question, do not waste time on it but go on to the next. You may not be able to finish all the questions.
- If you wish to alter an answer, put in the alteration clearly. You will not lose marks for crossing out.
- When you have finished one page, go straight on to the next.
- You will have 50 minutes to do the test.

NFER-NELSON



NOTTINGHAM POLYTECHNIC — TP Form Fin 12

CLAIM FOR TRAVELLING/SUBSISTENCE EXPENSES FOR RESEARCH VISITS

Name (Capitals) __________

Department __________

Home address (if claim over £40) __________

Date(s) of journey __________

Time of absence : From __________ To __________

Destination __________

Reason for journey __________

	£	p
Train/Bus — Single/Return journey		
Between __________ and __________		
Car mileage :- __________ cc of vehicle __________		
Mileometer reading : opening __________ closing __________		
Miles claimed __________ at __________		
Subsistence :		
Other expenses (details)		

I declare that the expenses charged herein have been necessarily incurred on Polytechnic business and I have not been reimbursed from any other source.

Signed __________

Approved by Chairman of Faculty Research Committee __________

G/L CODE — Value £

Collected by __________ Date __________

APPENDIX B – 15 Filled Forms (Single Writer)


Television Licence

Licence details to be completed by the post office

Colour TV ✓ Black and white TV Value £104

Expires on the last day of month JANUARY year 2001

1 Your name and address

MR W.S. WONG

1522 ABBOTSFORD DRIVE

ST ANNS

NOTTINGHAM

Post code NG3 1NB

2 If you already have a TV Licence but your name or address has changed

MR K TAN

ST8BC10101

Post code NG1 4BU

What this licence allows


Despatch Certificate (PFU 75)

PARCELFORCE WORLDWIDE Standard

Number of parcels received 1

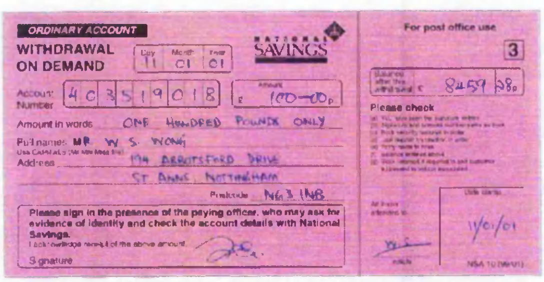
1 Barcode | Name and Address | Postcode

0208472591 | W.S. WONG, 194 ABBOTSFORD DRIVE, ST ANNS, NOTTINGHAM | NG3 1ND

11/1/00

PLEASE KEEP THIS CERTIFICATE SAFE. IT MUST BE PRODUCED IN THE EVENT OF A CLAIM

11/1/00


ORDINARY ACCOUNT

WITHDRAWAL ON DEMAND

NATIONAL SAVINGS

Day 11 Month 01 Year 01

Account Number 4C351901 8

Amount £ 100—00p

Amount in words ONE HUNDRED POUNDS ONLY

Full names MR W S WONG

Address 194 ABBOTSFORD DRIVE ST ANNS, NOTTINGHAM

Postcode NG3 1NB

Please sign in the presence of the paying officer, who may ask for evidence of identity and check the account details with National Savings.

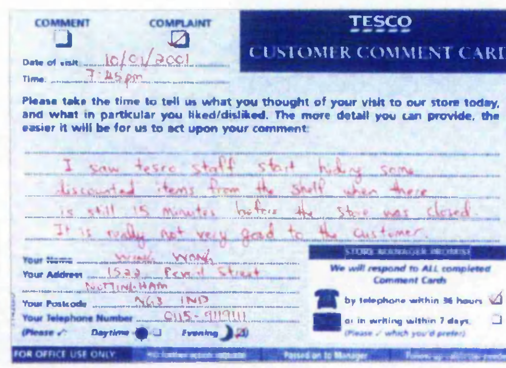
Signature

For post office use 3

8459 28p

Please check

11/01/01


COMMENT COMPLAINT

TESCO

CUSTOMER COMMENT CARD

Date of visit 10/01/2001

Time 7:45pm

Please take the time to tell us what you thought of your visit to our store today, and what in particular you liked/disliked. The more detail you can provide, the easier it will be for us to act upon your comment.

I saw tesco staff start hiding some discounted items from the shelf when there is still 15 minutes before the store was closed. It is really not very good to the customer.

Your Name WING WONG

Your Address 1522 Pencil Street NOTTINGHAM

Your Postcode NG3 1ND

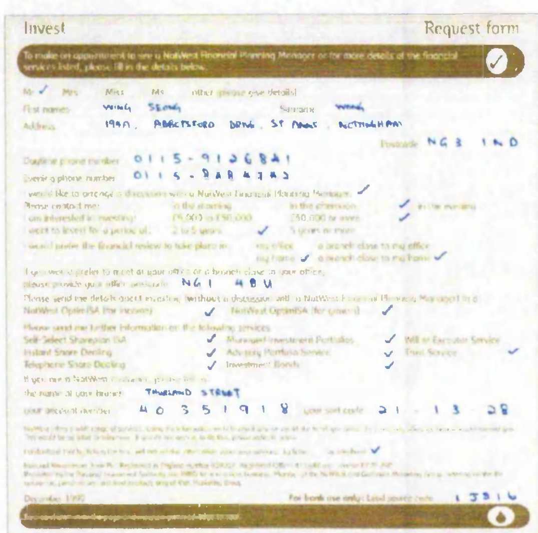
Your Telephone Number 0115-9119111

We will respond to ALL completed Comment Cards

by telephone within 36 hours

or in writing within 7 days

FOR OFFICE USE ONLY


Invest Request form

First names WING SEONG Surname WONG

Address 194A ABBOTSFORD DRIVE, ST ANNS, NOTTINGHAM

Postcode NG3 1ND

Daytime phone number 0115-9126841

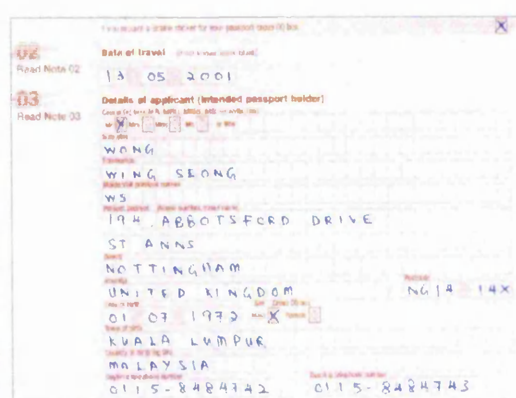
Evening phone number 0115-9484742

NG1 4BU

THURLAND STREET

40351918 21-13-28

13516


02 Date of travel 13 05 2001

03 Details of applicant (intended passport holder)

WONG

WING SEONG

WS

194 ABBOTSFORD DRIVE

ST ANNS

NOTTINGHAM

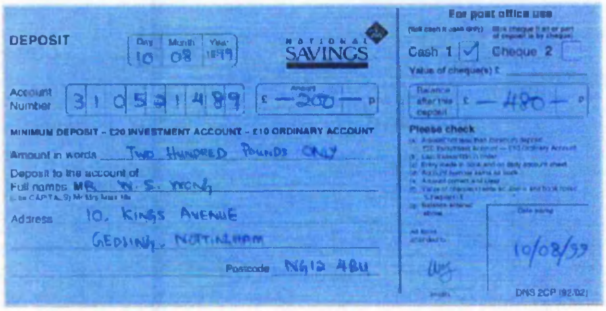
UNITED KINGDOM NG1A 1AX

01 07 1972

KUALA LUMPUR

MALAYSIA

0115-8484742 0115-8484743



DEPOSIT

Day 10 Month 08 Year 1999

NATIONAL SAVINGS

Account Number 3 1 0 5 2 1 4 8 9 £ —200— p

MINIMUM DEPOSIT – £20 INVESTMENT ACCOUNT – £10 ORDINARY ACCOUNT

Amount in words TWO HUNDRED POUNDS ONLY

Deposit to the account of
Full name MR W. S. WONG

Address 10. KINGS AVENUE
GEDLING, NOTTINGHAM

Postcode NG12 4BU

For post office use

Cash 1 ✓ Cheque 2

Value of cheque(s) £

Balance after this deposit £ — 480 — p

Please check

10/08/99

DNS 2CP (02/02)

Personal Loan Application Form

Please complete both pages. Then either:

Call our Loans Hotline on 0800 24 34 86

About you

First applicant's details only

Residential status

If you are a home owner please state

Marital status

Your employment

Occupation STUDENT

Apply for your Tesco credit card today

TESCO Personal Finance

4.9% APR

Savings Request form

Dr

First names WING SEONG Surname WONG

Address 194, ABBOTSFORD DRIVE, ST ANNS
NOTTINGHAM Postcode NG3 1ND

Daytime phone number 0115-348 4742 Evening 0115-9 26 8 40

Please send me more information on the following NatWest services.

I would like to discuss my needs with a Financial Planning Manager

Amount I have available for investment £ 100,000

I want to invest for: 2 to 5 years ✓ 5 years or more

Please contact me: in the morning in the afternoon ✓ in the evening

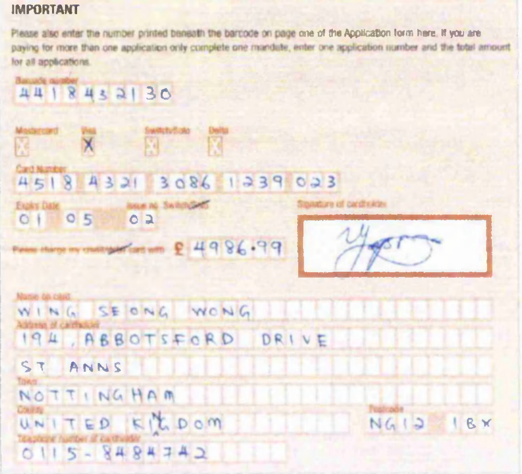
I would prefer the financial review to take place at: my office ✓ my home a branch close to my office ✓ a branch close to my home

If you would prefer to meet at your office, or a branch close to your office, please provide your office postcode NG1 4BU

If you are a NatWest customer, please tell us: The name of your branch VICTORIA STREET

Your account number 40351802 Your sort code 40-35-18

December 1999 For Bank use only: Lead source code



IMPORTANT

Please also enter the number printed beneath the barcode on page one of the Application form here. If you are paying for more than one application only complete one mandate, enter one application number and the total amount for all applications.

Barcode number 4418432130

Mastercard Visa X Switch/Solo Delta

Card Number 4518 4321 3086 1239 023

Expiry Date 01 05 Issue no. Switch 02

Signature of cardholder

Please charge my card with £ 4986.99

Name on card WING SEONG WONG

Address of cardholder 194 ABBOTSFORD DRIVE

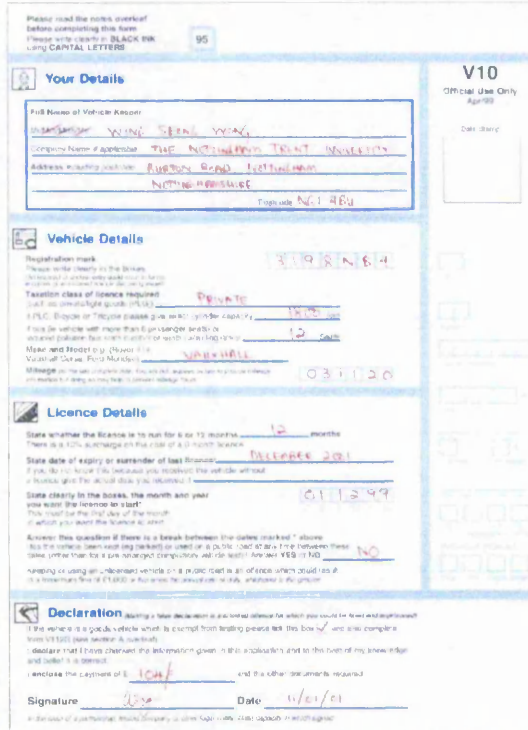
ST ANNS

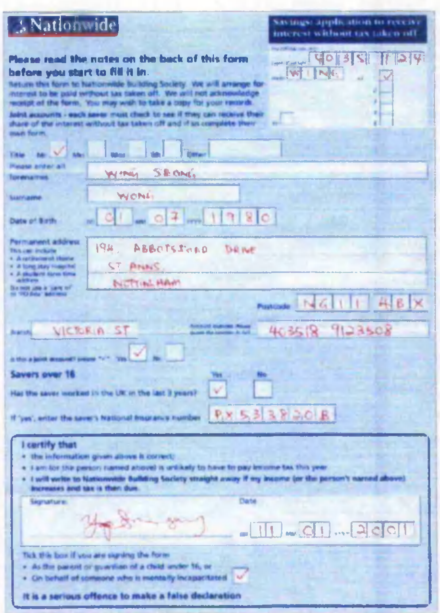
Town NOTTINGHAM

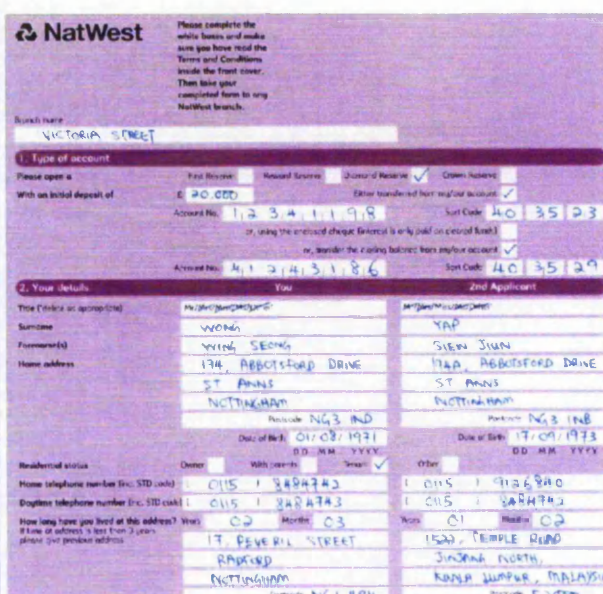
Country UNITED KINGDOM Postcode NG12 1BX

Telephone number of cardholder 0115-8484742


APPLICATION FORM


Your Details
V10
Vehicle Details
Licence Details
Declaration
Signature
Date


Nationwide
Please read the notes on the back of this form before you start to fill it in.


NatWest
1. Type of account
2. Your details
You
2nd Applicant

APPENDIX C – 3 Different Forms (each filled by 10 writers)

This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF88CT.

RETAIL CUSTOMER CLAIM FORM

Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ☐ ADDRESSEE ☑ THIRD PARTY ☐

ALL SECTIONS MUST BE COMPLETED

Your reference

2. SENDER NAME
Name/Company: HON HOCK WOON
Contact Name: HON HOCK WOON
Telephone Number: (0115) 9127093
Fax / E-Mail Number: HOCK.HON@NTU.AC.UK

3. SENDER ADDRESS
1C LORNE WALK
ST. ANNS NOTTINGHAM
Postcode: NG3 4FX
Country: UNITED KINGDOM

4. ADDRESSEE NAME
Name/Company: CHAN CHEE CHONG
Contact Name: CHAN CHEE CHONG
Telephone Number: (0115) 9119110

5. ADDRESSEE ADDRESS
363 HAMPDEN STREET
Postcode: NG1 3BW
Country: U.K.

If there are any further details regarding the collection/delivery of the item please attach on a separate sheet.

6. NATURE OF CLAIM
LOSS ☐ DELAY ☐ DAMAGE* ☑ MISSING CONTENTS* ☐
* Please indicate where parcel can be inspected SENDER/ADDRESSEE (Delete as appropriate)

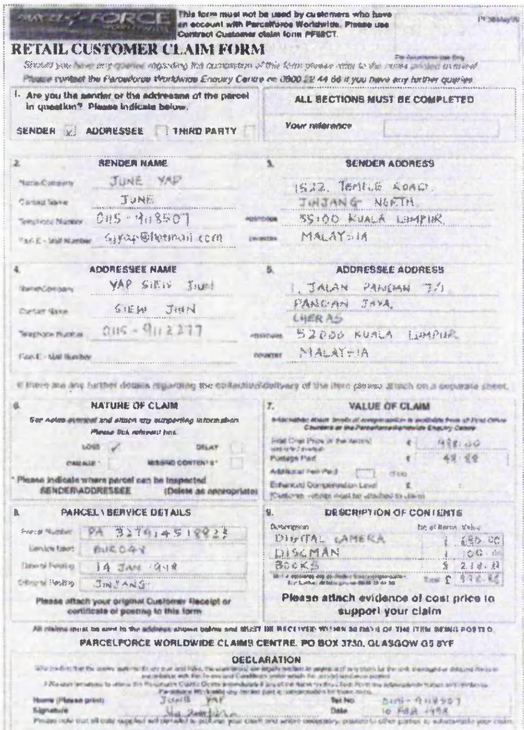
7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 25.05
Postage Paid: £ 6.98

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA 66335554791
Service Used: 9, 10 and 12
Date of Posting: 9.5.00
Office of Posting: NOTTINGHAM
Please attach your original Customer Receipt or certificate of posting to this form

9. DESCRIPTION OF CONTENTS
SOUND CARD | 1 | 25.05
Total £ 25.05
Please attach evidence of cost price to support your claim

PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION
Name (Please print): HON HOCK WOON
Tel No: 0115 9127093
Date: 9.5.00



RETAIL CUSTOMER CLAIM FORM

1. SENDER ☑ ADDRESSEE ☐ THIRD PARTY ☐

2. SENDER NAME
Name/Company: JUNE YAP
Contact Name: JUNE
Telephone Number: 0115-9118507
Fax / E-Mail Number: sjyap@hotmail.com

3. SENDER ADDRESS
1522, TEMPLE ROAD,
JINJANG NORTH,
Postcode: 55100 KUALA LUMPUR
Country: MALAYSIA

4. ADDRESSEE NAME
Name/Company: YAP SIEW JUN
Contact Name: SIEW JUN
Telephone Number: 0115-9112277

5. ADDRESSEE ADDRESS
1, JALAN PANDAN 3/1
PANDAN JAYA,
CHERAS
Postcode: 52000 KUALA LUMPUR
Country: MALAYSIA

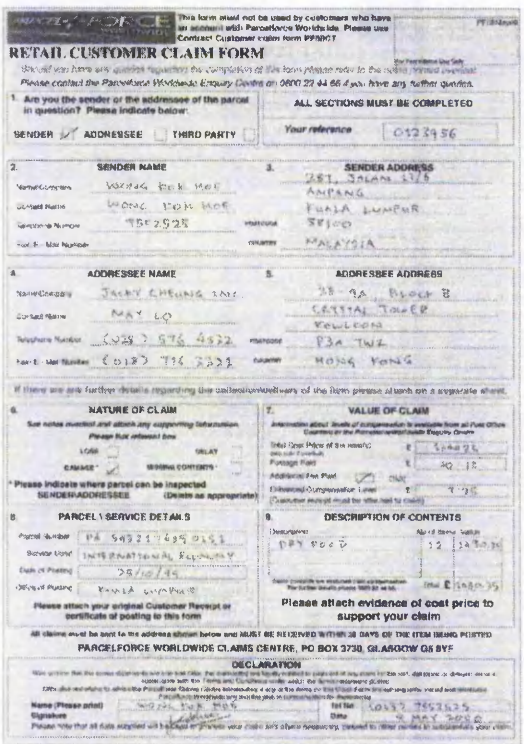
6. NATURE OF CLAIM
LOSS ☑ DELAY ☐ DAMAGE* ☐ MISSING CONTENTS* ☐

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 918.00
Postage Paid: £ 48.88

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA 32791451892
Service Used: AIR 24
Date of Posting: 14 JAN 1998
Office of Posting: JINJANG

9. DESCRIPTION OF CONTENTS
DIGITAL CAMERA
DISCMAN
BOOKS

DECLARATION
Name (Please print): JUNE YAP



RETAIL CUSTOMER CLAIM FORM

1. SENDER ☑ ADDRESSEE ☐ THIRD PARTY ☐
Your reference: 0123456

2. SENDER NAME
Name/Company: WONG FOK HOE
Contact Name: WONG FOK HOE
Telephone Number: 7552525

3. SENDER ADDRESS
281, JALAN 11/6
AMPANG
KUALA LUMPUR
Postcode: 58100
Country: MALAYSIA

4. ADDRESSEE NAME
Name/Company: JACKY CHEUNG SAU
Contact Name: MAY LO
Telephone Number: (038) 576 4532
Fax / E-Mail Number: (038) 776 3321

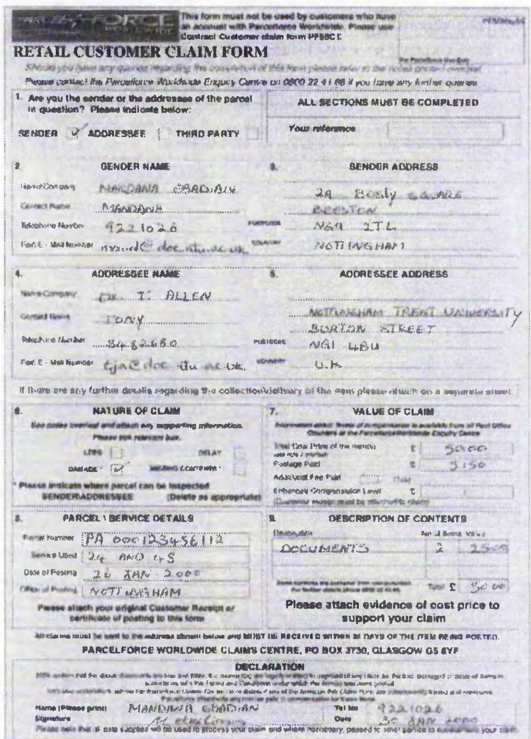
5. ADDRESSEE ADDRESS
38-9A, BLOCK B
CRYSTAL TOWER
KOWLOON
Postcode: P3A 7W2
Country: HONG KONG

6. NATURE OF CLAIM
DAMAGE* ☑

8. PARCEL \ SERVICE DETAILS
Office of Posting: KUALA LUMPUR

9. DESCRIPTION OF CONTENTS
DRY FOOD | 12

DECLARATION
Name (Please print): WONG FOK HOE



RETAIL CUSTOMER CLAIM FORM

1. SENDER ☑ ADDRESSEE ☐ THIRD PARTY ☐

2. SENDER NAME
Name/Company: MANDANA EBADIAN
Contact Name: MANDANA
Telephone Number: 9221026

3. SENDER ADDRESS
2A BOSLY SQUARE
BEESTON
Postcode: NG9 2TL
Country: NOTTINGHAM

4. ADDRESSEE NAME
Name/Company: DR. T. ALLEN
Contact Name: TONY
Telephone Number: 8486680
Fax / E-Mail Number: tja@doc.ntu.ac.uk

5. ADDRESSEE ADDRESS
NOTTINGHAM TRENT UNIVERSITY
BURTON STREET
Postcode: NG1 4BU
Country: U.K.

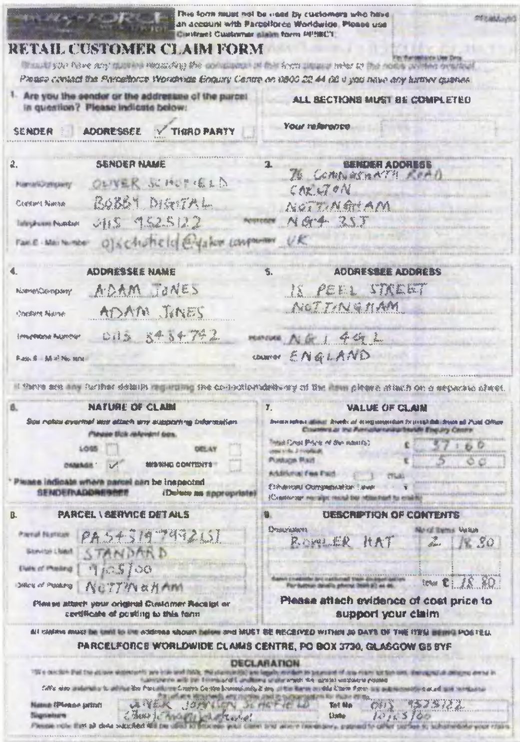
6. NATURE OF CLAIM
LOSS ☐ DELAY ☐ DAMAGE* ☑ MISSING CONTENTS* ☐

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 50.00
Postage Paid: £ 5.50

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA 000123456112
Service Used: 24 AND 48
Date of Posting: 26 JAN 2000
Office of Posting: NOTTINGHAM

9. DESCRIPTION OF CONTENTS
DOCUMENTS | 2 | 25.00
Total £ 50.00

DECLARATION
Name (Please print): MANDANA EBADIAN
Tel No: 9221026



This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF58CT.

RETAIL CUSTOMER CLAIM FORM

Should you have any queries regarding the completion of this form please refer to the notes printed overleaf.
Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ☐ ADDRESSEE ☑ THIRD PARTY ☐

ALL SECTIONS MUST BE COMPLETED
Your reference

2. SENDER NAME
Name/Company: OLIVER SCHOFIELD
Contact Name: BOBBY DIGITAL
Telephone Number: 0115 9525122
Fax/E-Mail Number: ojschofield@yahoo.com

3. SENDER ADDRESS
76 CANNINGHAM ROAD
CARLTON
NOTTINGHAM
Postcode: NG4 3ST
Country: UK

4. ADDRESSEE NAME
Name/Company: ADAM JONES
Contact Name: ADAM JONES
Telephone Number: 0115 8484742

5. ADDRESSEE ADDRESS
15 PEEL STREET
NOTTINGHAM
Postcode: NG1 4GL
Country: ENGLAND

If there are any further details regarding the collection/delivery of the item please attach on a separate sheet.

6. NATURE OF CLAIM
LOSS ☐ DELAY ☐ DAMAGE* ☑ MISSING CONTENTS* ☐
* Please indicate where parcel can be inspected SENDER/ADDRESSEE (Delete as appropriate)

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 37.60
Postage Paid: £ 5.00

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA54319799215T
Service Used: STANDARD
Date of Posting: 9/05/00
Office of Posting: NOTTINGHAM
Please attach your original Customer Receipt or certificate of posting to this form

9. DESCRIPTION OF CONTENTS

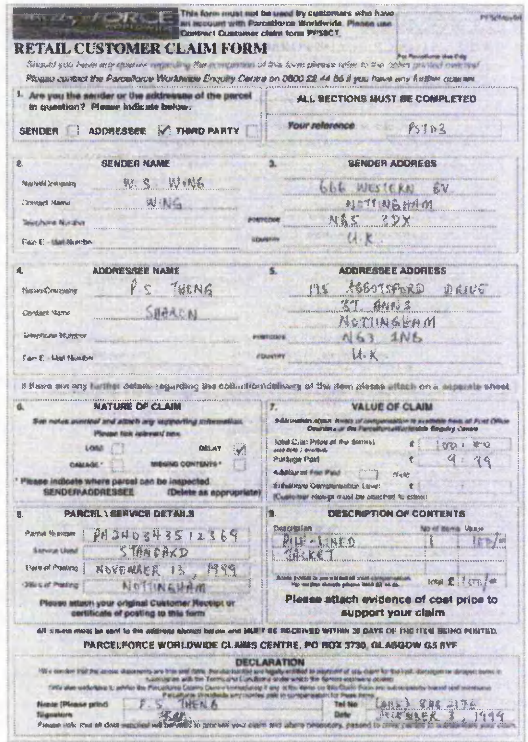
Description	No of Items	Value
BOWLER HAT	2	18.80

Total £ 18.80

Please attach evidence of cost price to support your claim

All claims must be sent to the address shown below and MUST BE RECEIVED WITHIN 30 DAYS OF THE ITEM BEING POSTED.
PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION
Name (Please print): OLIVER JOHNSON SCHOFIELD
Tel No: 0115 9525122
Date: 10/05/00



This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF58CT.

RETAIL CUSTOMER CLAIM FORM

Should you have any queries regarding the completion of this form please refer to the notes printed overleaf.
Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ☐ ADDRESSEE ☑ THIRD PARTY ☐

ALL SECTIONS MUST BE COMPLETED
Your reference: PST03

2. SENDER NAME
Name/Company: W. S. WING
Contact Name: WING

3. SENDER ADDRESS
666 WESTERN BV
NOTTINGHAM
Postcode: NG5 2DX
Country: U.K.

4. ADDRESSEE NAME
Name/Company: P S THENG
Contact Name: SHARON

5. ADDRESSEE ADDRESS
195 ABBOTSFORD DRIVE
ST ANNS
NOTTINGHAM
Postcode: NG3 1NB
Country: U.K

If there are any further details regarding the collection/delivery of the item please attach on a separate sheet.

6. NATURE OF CLAIM
LOSS ☐ DELAY ☑ DAMAGE* ☐ MISSING CONTENTS* ☐
* Please indicate where parcel can be inspected SENDER/ADDRESSEE (Delete as appropriate)

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 100.00
Postage Paid: £ 9.99

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA2403435 12369
Service Used: STANDARD
Date of Posting: NOVEMBER 13, 1999
Office of Posting: NOTTINGHAM
Please attach your original Customer Receipt or certificate of posting to this form

9. DESCRIPTION OF CONTENTS

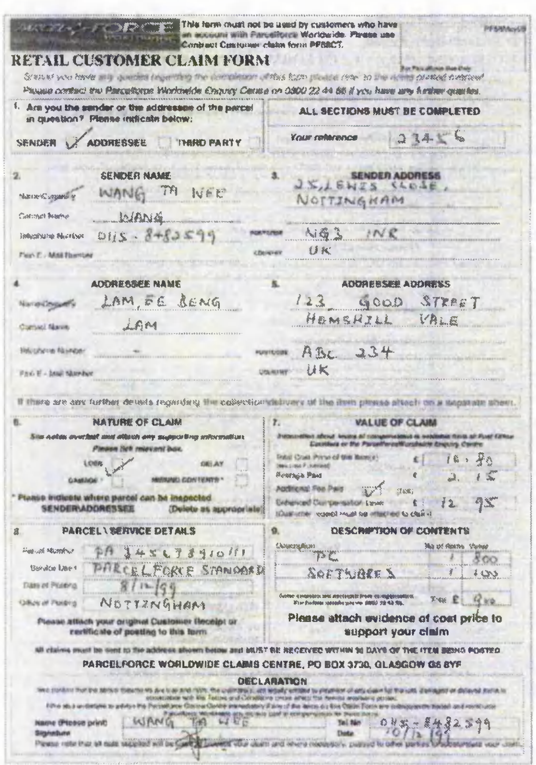
Description	No of Items	Value
PIH-LINED JACKET	1	100/=

Total £ 100/=

Please attach evidence of cost price to support your claim

All claims must be sent to the address shown below and MUST BE RECEIVED WITHIN 30 DAYS OF THE ITEM BEING POSTED.
PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION
Name (Please print): P. S. THENG
Tel No: (0115) 986 2176
Date: DECEMBER 3, 1999



This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF58CT.

RETAIL CUSTOMER CLAIM FORM

Should you have any queries regarding the completion of this form please refer to the notes printed overleaf.
Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ☑ ADDRESSEE ☐ THIRD PARTY ☐

ALL SECTIONS MUST BE COMPLETED
Your reference: 23456

2. SENDER NAME
Name/Company: WANG TA WEE
Contact Name: WANG
Telephone Number: 0115 - 8482599

3. SENDER ADDRESS
25, LEWIS CLOSE,
NOTTINGHAM
Postcode: NG3 1NR
Country: UK

4. ADDRESSEE NAME
Name/Company: LAM, EE BENG
Contact Name: LAM
Telephone Number: -

5. ADDRESSEE ADDRESS
123 GOOD STREET
HEMSHILL VALE
Postcode: ABC 234
Country: UK

If there are any further details regarding the collection/delivery of the item please attach on a separate sheet.

6. NATURE OF CLAIM
LOSS ☑ DELAY ☐ DAMAGE* ☐ MISSING CONTENTS* ☐
* Please indicate where parcel can be inspected SENDER/ADDRESSEE (Delete as appropriate)

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 16.80
Postage Paid: £ 2.15
Additional Fee Paid ☑
Enhanced Compensation Level: £ 12.95

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA 34567891011
Service Used: PARCELFORCE STANDARD
Date of Posting: 8/12/99
Office of Posting: NOTTINGHAM
Please attach your original Customer Receipt or certificate of posting to this form

9. DESCRIPTION OF CONTENTS

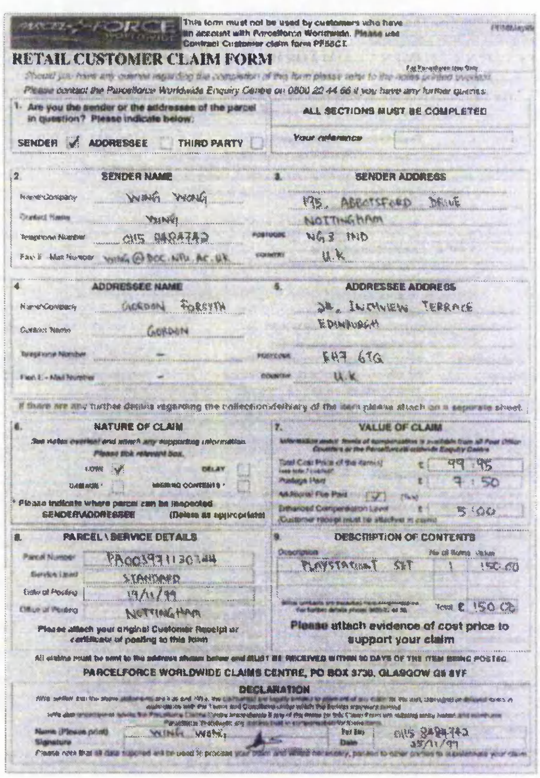
Description	No of Items	Value
PC	1	800
SOFTWARES	1	100

Total £ 900

Please attach evidence of cost price to support your claim

All claims must be sent to the address shown below and MUST BE RECEIVED WITHIN 30 DAYS OF THE ITEM BEING POSTED.
PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION
Name (Please print): WANG TA WEE
Tel No: 0115 - 8482599
Date: 10/12/99



This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF58CT.

RETAIL CUSTOMER CLAIM FORM

Should you have any queries regarding the completion of this form please refer to the notes printed overleaf.
Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ☑ ADDRESSEE ☐ THIRD PARTY ☐

ALL SECTIONS MUST BE COMPLETED
Your reference

2. SENDER NAME
Name/Company: WING WONG
Contact Name: WING
Telephone Number: 0115 8484742
Fax/E-Mail Number: wing@doc.ntu.ac.uk

3. SENDER ADDRESS
195, ABBOTSFORD DRIVE
NOTTINGHAM
Postcode: NG3 1ND
Country: U.K

4. ADDRESSEE NAME
Name/Company: GORDON FORSYTH
Contact Name: GORDON
Telephone Number: -
Fax/E-Mail Number: -

5. ADDRESSEE ADDRESS
28, INCHVIEW TERRACE
EDINBURGH
Postcode: EH7 6TG
Country: U.K

If there are any further details regarding the collection/delivery of the item please attach on a separate sheet.

6. NATURE OF CLAIM
LOSS ☑ DELAY ☐ DAMAGE* ☐ MISSING CONTENTS* ☐
* Please indicate where parcel can be inspected SENDER/ADDRESSEE (Delete as appropriate)

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 99.95
Postage Paid: £ 7.50
Additional Fee Paid ☑
Enhanced Compensation Level: £ 5.00

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA00397130744
Service Used: STANDARD
Date of Posting: 19/11/99
Office of Posting: NOTTINGHAM
Please attach your original Customer Receipt or certificate of posting to this form

9. DESCRIPTION OF CONTENTS

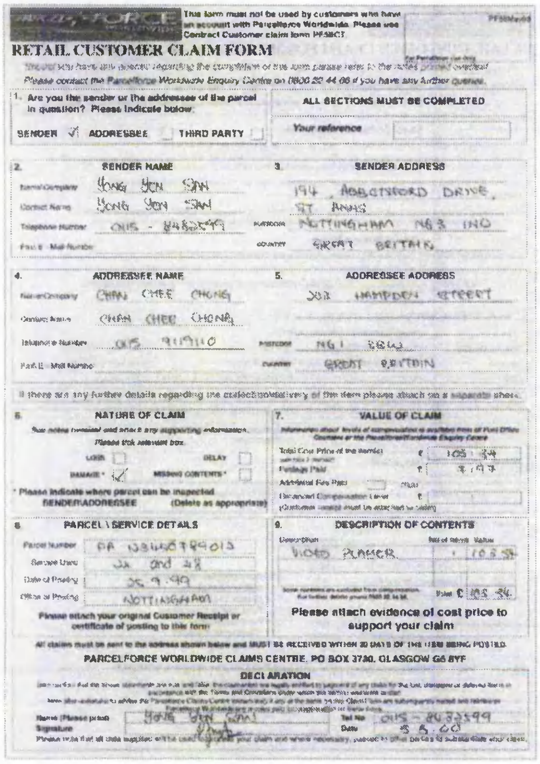
Description	No of Items	Value
PLAYSTATION SET	1	150.00

Total £ 150.00

Please attach evidence of cost price to support your claim

All claims must be sent to the address shown below and MUST BE RECEIVED WITHIN 30 DAYS OF THE ITEM BEING POSTED.
PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION
Name (Please print): WING WONG
Tel No: 0115 8484742
Date: 25/11/99


RETAIL CUSTOMER CLAIM FORM

This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF56CT.

Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ✓ ADDRESSEE THIRD PARTY

ALL SECTIONS MUST BE COMPLETED

Your reference

2. SENDER NAME
Name/Company: YONG YEN SAN
Contact Name: YONG YEN SAN
Telephone Number: 0115 - 8482599

3. SENDER ADDRESS
194, ABBOTSFORD DRIVE
ST ANNS
NOTTINGHAM NG3 1NQ
Country: GREAT BRITAIN

4. ADDRESSEE NAME
Name/Company: CHAN CHEE CHONG
Contact Name: CHAN CHEE CHONG
Telephone Number: 0115 9119110

5. ADDRESSEE ADDRESS
303 HAMPDEN STREET
NG1 3BU
Country: GREAT BRITAIN

6. NATURE OF CLAIM
LOSS / DELAY / DAMAGE* ✓ / MISSING CONTENTS*

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 105.84
Postage Paid: £ 7.97

8. PARCEL \ SERVICE DETAILS
Parcel Number: PA 03448789013
Service Used: 24 and 48
Date of Posting: 25.9.99
Office of Posting: NOTTINGHAM

Please attach your original Customer Receipt or certificate of posting to this form

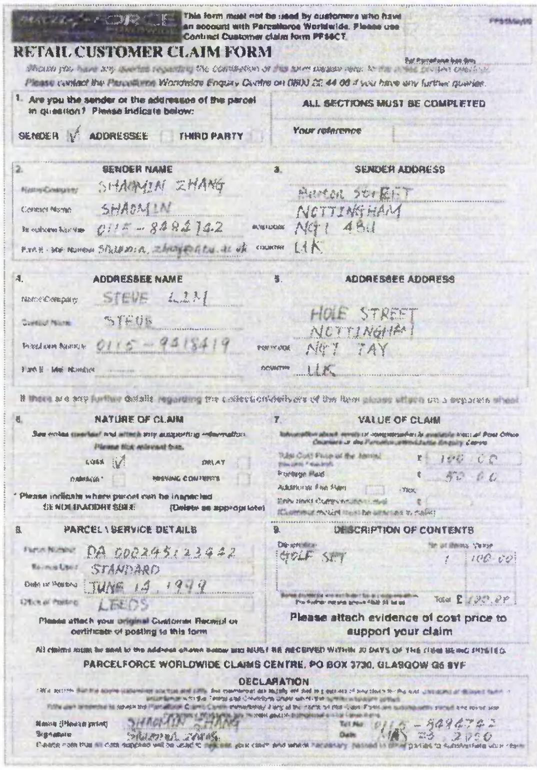
9. DESCRIPTION OF CONTENTS
VIDEO PLAYER 1 105.84
Total £ 105.84

Please attach evidence of cost price to support your claim

PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION

Name (Please print): YONG YEN SAN
Tel No: 0115 - 8482599
Date: 5.5.00


RETAIL CUSTOMER CLAIM FORM

This form must not be used by customers who have an account with Parcelforce Worldwide. Please use Contract Customer claim form PF56CT.

Please contact the Parcelforce Worldwide Enquiry Centre on 0800 22 44 66 if you have any further queries.

1. Are you the sender or the addressee of the parcel in question? Please indicate below:
SENDER ✓ ADDRESSEE THIRD PARTY

ALL SECTIONS MUST BE COMPLETED

Your reference

2. SENDER NAME
Name/Company: SHAOMIN ZHANG
Contact Name: SHAOMIN
Telephone Number: 0115 - 8482142
Fax/E-Mail Number: Shaomin.zhang@ntu.ac.uk

3. SENDER ADDRESS
Burton STREET
NOTTINGHAM
NG1 4BU
Country: UK

4. ADDRESSEE NAME
Name/Company: STEVE LIM
Contact Name: STEVE
Telephone Number: 0115 - 9518419

5. ADDRESSEE ADDRESS
HOLE STREET
NOTTINGHAM
NG1 7AY
Country: UK

6. NATURE OF CLAIM
LOSS ✓ / DELAY / DAMAGE* / MISSING CONTENTS*

7. VALUE OF CLAIM
Total Cost Price of the item(s): £ 100.00
Postage Paid: £ 50.00

8. PARCEL \ SERVICE DETAILS
Parcel Number: DA 002245123422
Service Used: STANDARD
Date of Posting: JUNE 14, 1999
Office of Posting: LEEDS

Please attach your original Customer Receipt or certificate of posting to this form

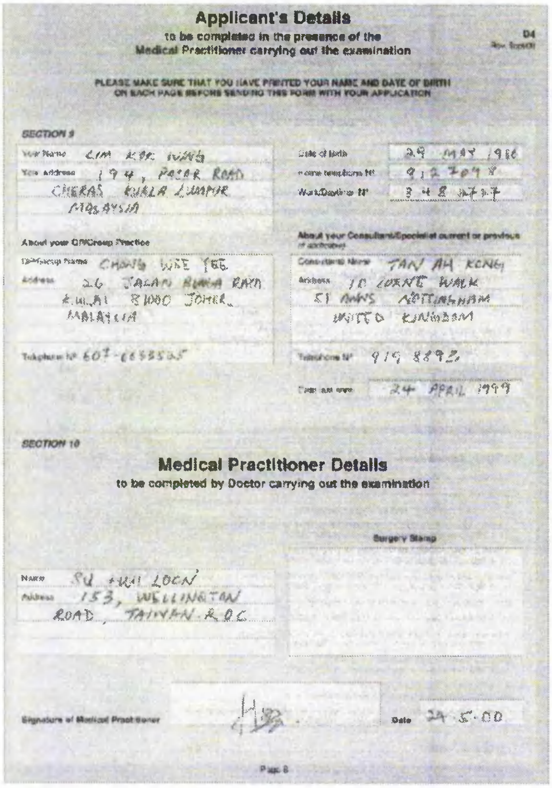
9. DESCRIPTION OF CONTENTS
GOLF SET 1 100.00
Total £ 100.00

Please attach evidence of cost price to support your claim

PARCELFORCE WORLDWIDE CLAIMS CENTRE, PO BOX 3730, GLASGOW G5 8YF

DECLARATION

Name (Please print): SHAOMIN ZHANG
Tel No: 0115 - 8484742
Date: MAY 23, 2000


Applicant's Details
to be completed in the presence of the Medical Practitioner carrying out the examination

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9
Your Name: LIM KOK WING
Your Address: 194, PASAR ROAD, CHERAS KUALA LUMPUR MALAYSIA
Date of Birth: 29 MAY 1966
Home telephone Nº: 912 7098
Work/Daytime Nº: 348 3727

About your GP/Group Practice
GP/Group Name: CHANG WEE YEE
Address: 26 JALAN RUMAH RAYA KULAI 81000 JOHOR MALAYSIA
Telephone Nº: 607-6633525

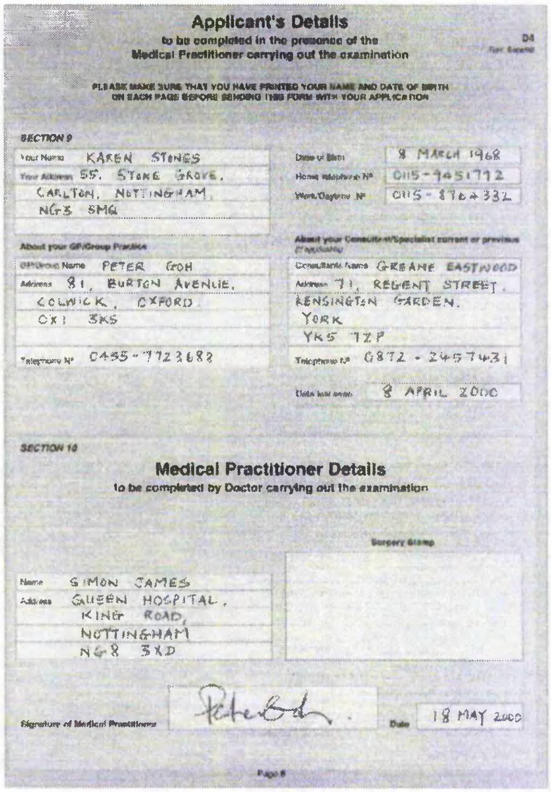
About your Consultant/Specialist current or previous
Consultant's Name: TAN AH KENG
Address: 10 LORNE WALK, ST ANNS NOTTINGHAM, UNITED KINGDOM
Telephone Nº: 919 8893
Date last seen: 24 APRIL 1999

SECTION 10
Medical Practitioner Details
to be completed by Doctor carrying out the examination

Surgery Stamp

Name: SU HUI LOON
Address: 153, WELLINGTON ROAD, TAIWAN R.O.C.

Signature of Medical Practitioner
Date: 24.5.00


Applicant's Details
to be completed in the presence of the Medical Practitioner carrying out the examination

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9
Your Name: KAREN STONES
Your Address: 55, STONE GROVE, CARLTON, NOTTINGHAM, NG3 5MQ
Date of Birth: 8 MARCH 1968
Home telephone Nº: 0115-9451792
Work/Daytime Nº: 0115-8764332

About your GP/Group Practice
GP/Group Name: PETER GOH
Address: 81, BURTON AVENUE, COLWICK, OXFORD, OX1 3KS
Telephone Nº: 0455-7723682

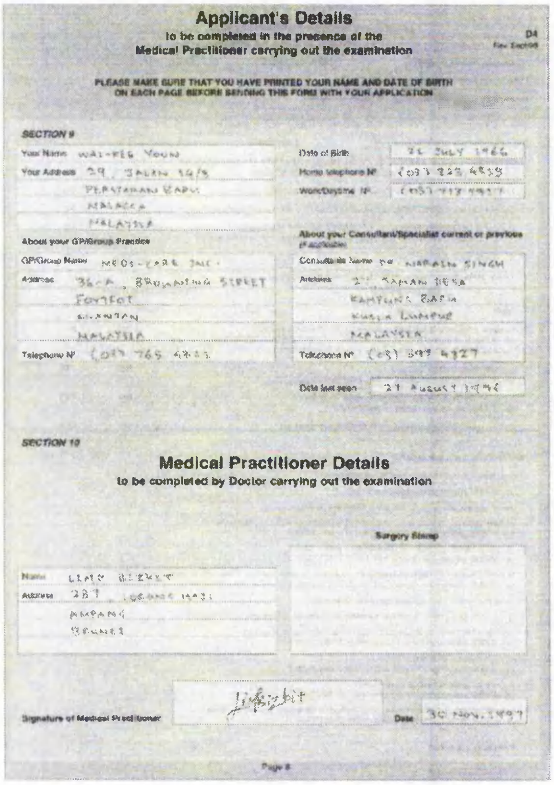
About your Consultant/Specialist current or previous
Consultant's Name: GREANE EASTWOOD
Address: 71, REGENT STREET, KENSINGTON GARDEN, YORK YK5 7ZP
Telephone Nº: 0872-2457431
Date last seen: 8 APRIL 2000

SECTION 10
Medical Practitioner Details
to be completed by Doctor carrying out the examination

Surgery Stamp

Name: SIMON JAMES
Address: QUEEN HOSPITAL, KING ROAD, NOTTINGHAM NG8 3XD

Signature of Medical Practitioner
Date: 18 MAY 2000



Applicant's Details
to be completed in the presence of the
Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name WAI-KEE YOUNG
Your Address 29, JALAN 14/5
PERSTARAN BARU
MALACCA
MALAYSIA

Date of Birth 21 JULY 1966
Home telephone Nº (03) 323 4559
Work/Daytime Nº (03) 717 9927

About your GP/Group Practice
GP/Group Name MEDI-CARE INC.
Address 36-A, BROWNING STREET
FORTFOT
KLANTAN
MALAYSIA
Telephone Nº (03) 765 4833

About your Consultant/Specialist current or previous (if applicable)
Consultants Name DR NIRAJ SINGH
Address 27 TAMAN DESA
KAMPUNG BARU
KUALA LUMPUR
MALAYSIA
Telephone Nº (03) 397 4927

Date last seen 21 AUGUST 1996

SECTION 10

Medical Practitioner Details
to be completed by Doctor carrying out the examination

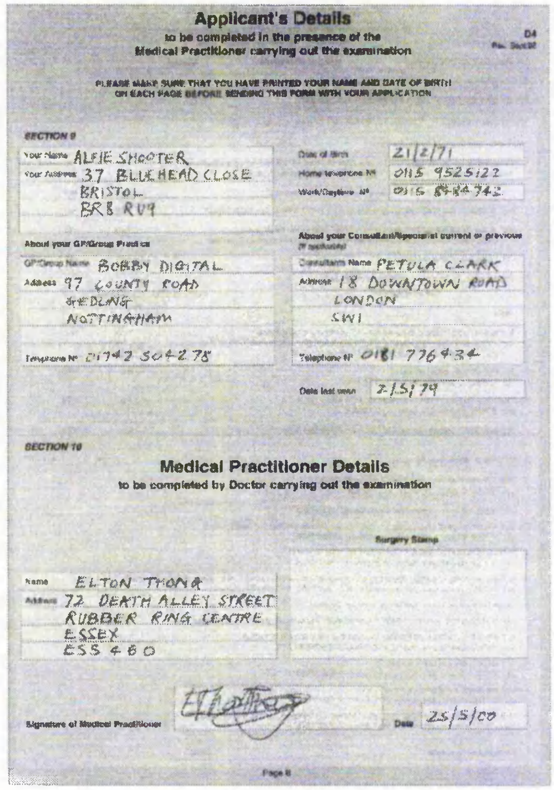
Surgery Stamp

Name LIMA BIRKET
Address 287 ... ROAD
KAMPANG
BRUNEI

Signature of Medical Practitioner

Date 30 NOV. 1997

Page 8



Applicant's Details
to be completed in the presence of the
Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name ALFIE SHOOTER
Your Address 37 BLUEHEAD CLOSE
BRISTOL
BR8 RU9

Date of Birth 21/2/71
Home telephone Nº 0115 9525122
Work/Daytime Nº 0115 8484742

About your GP/Group Practice
GP/Group Name BOBBY DIGITAL
Address 97 COUNTY ROAD
GEDLING
NOTTINGHAM
Telephone Nº 01742 504278

About your Consultant/Specialist current or previous (if applicable)
Consultants Name PETULA CLARK
Address 18 DOWNTOWN ROAD
LONDON
SW1
Telephone Nº 0181 776434

Date last seen 2/5/79

SECTION 10

Medical Practitioner Details
to be completed by Doctor carrying out the examination

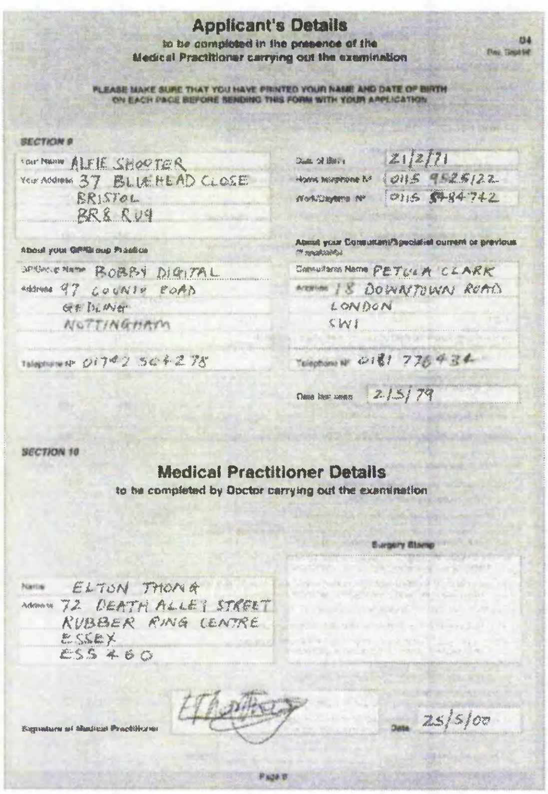
Surgery Stamp

Name ELTON THONG
Address 72 DEATH ALLEY STREET
RUBBER RING CENTRE
ESSEX
ESS 460

Signature of Medical Practitioner

Date 25/5/00

Page 8



Applicant's Details
to be completed in the presence of the
Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name ALFIE SHOOTER
Your Address 37 BLUEHEAD CLOSE
BRISTOL
BR8 RU9

Date of Birth 21/2/71
Home telephone Nº 0115 9525122
Work/Daytime Nº 0115 8484742

About your GP/Group Practice
GP/Group Name BOBBY DIGITAL
Address 97 COUNTY ROAD
GEDLING
NOTTINGHAM
Telephone Nº 01742 504278

About your Consultant/Specialist current or previous (if applicable)
Consultants Name PETULA CLARK
Address 18 DOWNTOWN ROAD
LONDON
SW1
Telephone Nº 0181 776434

Date last seen 2/5/79

SECTION 10

Medical Practitioner Details
to be completed by Doctor carrying out the examination

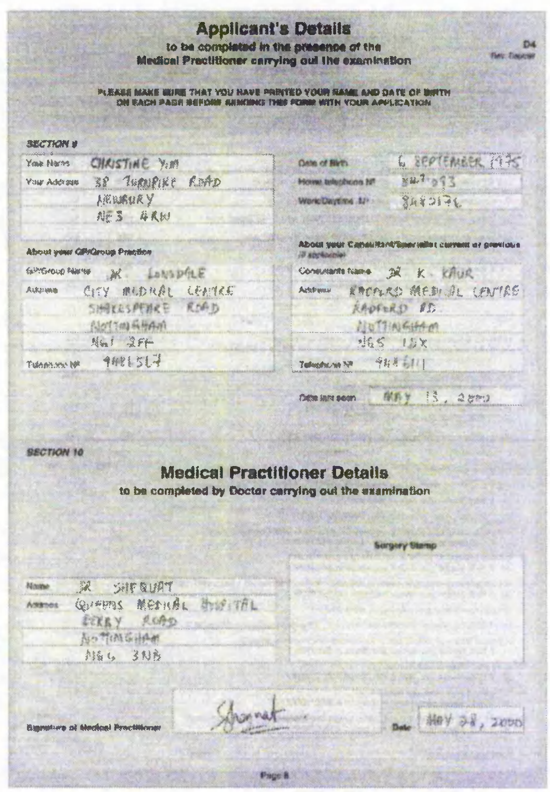
Surgery Stamp

Name ELTON THONG
Address 72 DEATH ALLEY STREET
RUBBER RING CENTRE
ESSEX
ESS 460

Signature of Medical Practitioner

Date 25/5/00

Page 8



Applicant's Details
to be completed in the presence of the
Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name CHRISTINE YIM
Your Address 38 TURNPIKE ROAD
NEWBURY
NE3 4RW

Date of Birth 6 SEPTEMBER 1975
Home telephone Nº 847093
Work/Daytime Nº 8442176

About your GP/Group Practice
GP/Group Name DR. LONSDALE
Address CITY MEDICAL CENTRE
SHAKESPEARE ROAD
NOTTINGHAM
NG1 2FF
Telephone Nº 9481517

About your Consultant/Specialist current or previous (if applicable)
Consultants Name DR K. KAUR
Address RADFORD MEDICAL CENTRE
RADFORD RD.
NOTTINGHAM
NG5 1DX
Telephone Nº 9448111

Date last seen MAY 13, 2000

SECTION 10

Medical Practitioner Details
to be completed by Doctor carrying out the examination

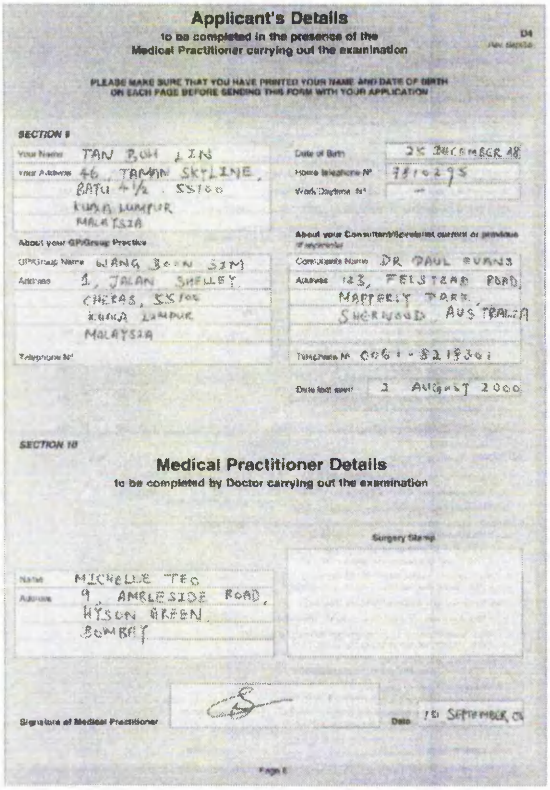
Surgery Stamp

Name DR SHERQUAT
Address QUEENS MEDICAL HOSPITAL
DERBY ROAD
NOTTINGHAM
NG6 3NB

Signature of Medical Practitioner

Date MAY 28, 2000

Page 8



Applicant's Details

to be completed in the presence of the Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name: TAN BOH LIN
Your Address: 46, TAMAN SKYLINE, BATU 4½, 55100 KUALA LUMPUR MALAYSIA

Date of Birth: 25 DECEMBER 48
Home telephone Nº: 7810295
Work/Daytime Nº: –

About your GP/Group Practice

GP/Group Name: WANG BOON SIM
Address: 1, JALAN SHELLEY CHERAS, 55100 KUALA LUMPUR MALAYSIA
Telephone Nº:

About your Consultant/Specialist current or previous (if applicable)

Consultants Name: DR PAUL EVANS
Address: 123, FELSTEAD ROAD, MAPPERLY PARK, SHERWOOD, AUSTRALIA
Telephone Nº: 0061-8219361

Date last seen: 2 AUGUST 2000

SECTION 10

Medical Practitioner Details

to be completed by Doctor carrying out the examination

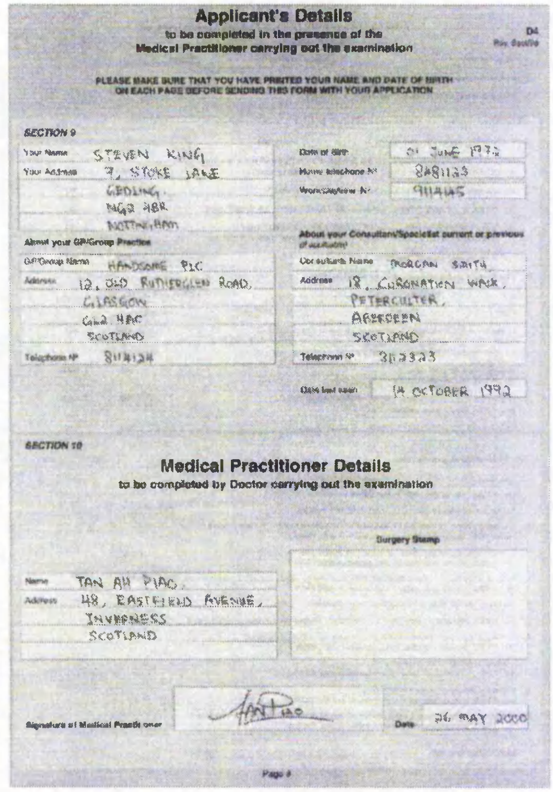
Surgery Stamp

Name: MICHELLE TEO
Address: 9, AMBLESIDE ROAD, HYSON GREEN, BOMBAY

Signature of Medical Practitioner

Date: 15 SEPTEMBER 00

Page 6



Applicant's Details

to be completed in the presence of the Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name: STEVEN KING
Your Address: 7, STOKE LANE GEDLING, NG2 4BR NOTTINGHAM

Date of Birth: 01 JUNE 1972
Home telephone Nº: 8481133
Work/Daytime Nº: 9114145

About your GP/Group Practice

GP/Group Name: HANDSOME PLC
Address: 12, OLD RUTHERGLEN ROAD, GLASGOW G42 4AC SCOTLAND
Telephone Nº: 8114134

About your Consultant/Specialist current or previous (if applicable)

Consultants Name: MORGAN SMITH
Address: 18, CORONATION WALK, PETERCULTER, ABERDEEN SCOTLAND
Telephone Nº: 3123323

Date last seen: 14 OCTOBER 1992

SECTION 10

Medical Practitioner Details

to be completed by Doctor carrying out the examination

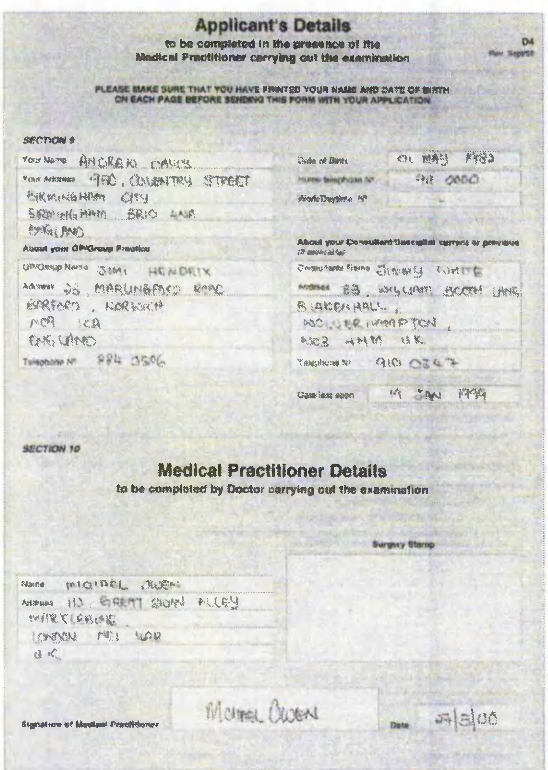
Surgery Stamp

Name: TAN AH PIAO
Address: 48, EASTFIELD AVENUE, INVERNESS SCOTLAND

Signature of Medical Practitioner: TAN PIAO

Date: 26 MAY 2000

Page 8



Applicant's Details

to be completed in the presence of the Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name: ANDREW DAVIS
Your Address: 780, COUNTRY STREET BIRMINGHAM CITY BIRMINGHAM B10 4NB ENGLAND

Date of Birth: 01 MAY 1982
Home telephone Nº: 911 0000
Work/Daytime Nº: –

About your GP/Group Practice

GP/Group Name: JIMI HENDRIX
Address: 33, MARLINGFORD ROAD EARLHAM, NORWICH NOR 1CA ENGLAND
Telephone Nº: 884 0506

About your Consultant/Specialist current or previous (if applicable)

Consultants Name: JIMMY WHITE
Address: 83, WILLIAM SCOTT LANE, BLAKENHALL, WOLVERHAMPTON, NO3 4NM U.K.
Telephone Nº: 910 0347

Date last seen: 19 JAN 1999

SECTION 10

Medical Practitioner Details

to be completed by Doctor carrying out the examination

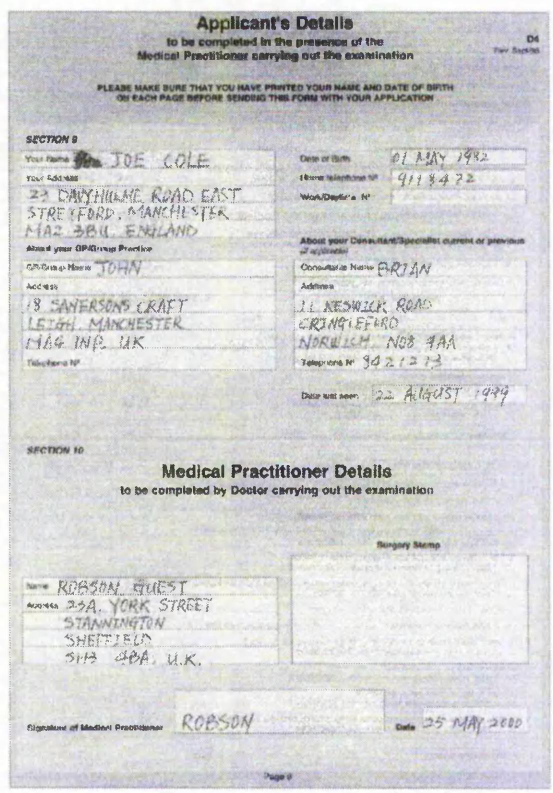
Surgery Stamp

Name: MICHAEL OWEN
Address: 113, GREAT SWAN ALLEY MARYLEBONE, LONDON NE1 4AB U.K.

Signature of Medical Practitioner: Michael Owen

Date: 27/5/00

Page 5



Applicant's Details

to be completed in the presence of the Medical Practitioner carrying out the examination

D4

PLEASE MAKE SURE THAT YOU HAVE PRINTED YOUR NAME AND DATE OF BIRTH ON EACH PAGE BEFORE SENDING THIS FORM WITH YOUR APPLICATION

SECTION 9

Your Name: JOE COLE
Your Address: 23 DAVYHULME ROAD EAST STRETFORD, MANCHESTER M32 3BU ENGLAND

Date of Birth: 01 MAY 1982
Home telephone Nº: 9113472
Work/Daytime Nº:

About your GP/Group Practice

GP/Group Name: JOHN
Address: 18 SAVERSONS CRAFT LEIGH, MANCHESTER M46 1NB UK
Telephone Nº:

About your Consultant/Specialist current or previous (if applicable)

Consultants Name: BRIAN
Address: 11 KESWICK ROAD CRINGLEFORD NORWICH NO8 4AA
Telephone Nº: 9421213

Date last seen: 22 AUGUST 1999

SECTION 10

Medical Practitioner Details

to be completed by Doctor carrying out the examination

Surgery Stamp

Name: ROBSON GUEST
Address: 23A, YORK STREET STANNINGTON SHEFFIELD S13 4BA, U.K.

Signature of Medical Practitioner: ROBSON

Date: 25 MAY 2000

Page 9

THIS TENANCY AGREEMENT

is made on the 31 day of MAY 2000

between

COMPANY NAME IMAGE SCAN HOLDINGS

ADDRESS 10 LORNE WALK STAINS NOTTINGHAM

POSTCODE NG3 4FX COUNTRY UK

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
CHONG WEE YEE	TAN JUI JIH
HOME ADDRESS	HOME ADDRESS
26 JALAN BUNGA RAYA	123 JALAN
KULAI JOHOR	ABU BAKAR
	KLUANG
POSTCODE 81000	POSTCODE 83812
TELEPHONE	TELEPHONE
6633304	7364528

The property let address 84 BURTON STREET

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month AUGUST 2000 to month JUNE 2001

Inclusive at a fixed rent of £ 120 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 17 day of 9 2000

between

COMPANY NAME SIEW JIUN LTD

ADDRESS 36, FREDERICK PLACE, BRIGHTON

POSTCODE BN1 4EA COUNTRY U.K

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
RUTH RENDELL	MARIO PUZO
HOME ADDRESS	HOME ADDRESS
16, WELLESBOURNE HOUSE,	18, GORWAY ROAD,
WALTON ROAD,	WALSALL,
WARWICK	WOLVERHAMPTON
POSTCODE CV35 9JB	POSTCODE WS1 3BD
TELEPHONE	TELEPHONE
01789 470007	01922 631200

The property let address 160, SOLLEY STREET, SHEFFIELD

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month OCTOBER 2000 to month JUNE 2001

Inclusive at a fixed rent of £ 100 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 23 day of 9 2000

between

COMPANY NAME AXIOMATIC

ADDRESS 23, LONG ROW AVE. NOTTINGHAM

POSTCODE NG1 4BG COUNTRY U.K.

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
KENNY LSEW	TERENCE TAN
HOME ADDRESS	HOME ADDRESS
63, JALAN KEPONG,	73 A, GLANMORGAN
AMPANG	STREET, SCOTLAND
SELANGOR	
POSTCODE 43659	POSTCODE S10 2FQ
TELEPHONE	TELEPHONE
(603) 618 6023	(0181) 375 7076

The property let address 14 ABBOT DR NOTTINGHAM

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month JUNE 2000 to month AUGUST 2001

Inclusive at a fixed rent of £ 750 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 12 day of MAY 2000

between

COMPANY NAME TNTU

ADDRESS 14 CITY AV LENTON ABBEY

POSTCODE NG9 2TL COUNTRY ENGLAND

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
NIMA	JAMSHID
HOME ADDRESS	HOME ADDRESS
UNIVERSITY ROAD	10 COLLEGE ROAD
BELFAST	BANGOR
NORTHERN IRELAND	WALES
POSTCODE BT7 1NN	POSTCODE LL57 2DG
TELEPHONE	TELEPHONE
028-90245133	01248-35451

The property let address 20 CHARLES AVE.

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month JUN 2000 to month FEB 2001

Inclusive at a fixed rent of £ 100 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 25 day of JULY 2000

between

COMPANY NAME NICEDAY LTD

ADDRESS 45 GEORGE HUDSON STREET

YORK

POSTCODE YO1 6ZG COUNTRY GB

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
ARNOLD HAMLET	BARRIE BETHALL
HOME ADDRESS	HOME ADDRESS
7 BOWLING HILL	21 CONWAY STREET
CHIPPING SODBURY	BIRKENHEAD
BRISTOL	MERSEYSIDE
POSTCODE BS37 6JX	POSTCODE L41 4FD
TELEPHONE	TELEPHONE
01454 863333	0151 748888

The property let address 42 TRUMPET STREET

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month APRIL 2000 to month JUNE 2001

Inclusive at a fixed rent of £ 100 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 3 day of MARCH 2000

between

COMPANY NAME STANNARD CORPORATION

ADDRESS 25 BRICKYARD LANE

EAST BRIDGFORD

POSTCODE NG13 7NF COUNTRY U.K.

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
PENNY CHIN	WEBIE SIN
HOME ADDRESS	HOME ADDRESS
148 PANDAN JAYA	6 SWALEDALE CL.
PUCHONG	MELBOURNE PARK
KUALA LUMPUR	NOTTINGHAM
POSTCODE 58200	POSTCODE NG8 2SX
TELEPHONE	TELEPHONE
7834150	9700946

The property let address 829 WESTERN BOULEVARD

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month JULY 2000 to month JULY 2001

Inclusive at a fixed rent of £ 250 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 23 day of JUNE 2000

between

COMPANY NAME TRAVAZL GROUP

ADDRESS 4, NOTTINGHAM ROAD,

LONG EATON, NOTTINGHAM

POSTCODE NG10 1MP COUNTRY UK.

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
TONY SHARPE	JON RIDING
HOME ADDRESS	HOME ADDRESS
26, WILTON TERRACE	46, TAMAN SKYLINE
CRESCENT, NEW BASFORD	BATU EMPAT,
LONDON, UK	JALAN KLANG LAMA,
POSTCODE GN4 3GT	POSTCODE 55100
TELEPHONE	TELEPHONE
022-9613130	03-7810295

The property let address 25, LEWIS CLOSE, ST ANNS

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month JULY 2000 to month JUNE 2001

Inclusive at a fixed rent of £ 30 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 28 day of MAY 2000

between

COMPANY NAME IT TECHNOLOGY

ADDRESS 1522, JALAN TIGA EMPAT

JINJANG NORTH, KUALA LUMPUR

POSTCODE 52000 COUNTRY MALAYSIA

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
HILTON KONG	MARHILL STEVEN
HOME ADDRESS	HOME ADDRESS
28, IVY CLOSE	1, CARLTON SQ.
GEDLING	ASHWOOD
NOTTINGHAM	OXFORD
POSTCODE NG4 1DA	POSTCODE OX2 1AA
TELEPHONE	TELEPHONE
0115-8481128	0219-3412186

The property let address 32, ABBOTSFORD DRIVE, ST ANNS

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month JUNE 2000 to month DECEMBER 2001

Inclusive at a fixed rent of £ 40 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 23 day of JUNE 2000

between

COMPANY NAME ALL ADVANTAGE

ADDRESS 520, BLOCK P/B

TAMAN CHERAS

POSTCODE 56100 COUNTRY SABAH

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
YONG YEN SAN	CHOW YUAN KAI
HOME ADDRESS	HOME ADDRESS
54, JALAN HANG	11-4A JALAN 3/12A
LORONG ZIFFA	TAMAN BARU
POSTCODE 76503	POSTCODE 20355
TELEPHONE	TELEPHONE
03-9119111	05-3787634

The property let address 194, ABBOTSFORD DRIVE

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month MAY 2000 to month JULY 2001

Inclusive at a fixed rent of £ 100.00 per person per week inclusive water rates

THIS TENANCY AGREEMENT

is made on the 15 day of JULY 2000

between

COMPANY NAME MICROSOFT

ADDRESS 2 HIGH STREET

CHELTENHAM

POSTCODE GL52 3AY COUNTRY U.K

(Hereinafter called the "Landlord") of the one part and

NAME	NAME
DAVID BECKHAM	EMMA WOODHOUSE
HOME ADDRESS	HOME ADDRESS
25 LONDON STREET	31 JEWRY STREET
BEDFORD	LONDON
POSTCODE MK42 9AP	POSTCODE EC3N 2EY
TELEPHONE	TELEPHONE
01234 363222	020 7320 1000

The property let address 2 RENFEW STREET, GLASGOW

Together with the right to use the furniture, articles and things about the said premises.

The tenant will hold the property for the period

From month MARCH 2000 to month AUGUST 2001

Inclusive at a fixed rent of £ 50 per person per week inclusive water rates

APPENDIX D – 15 Filled Forms (Single writer)

National Health Service

Form FP95 (Rev. April 1997)

APPLICATION FOR CERTIFICATE OF PREPAYMENT OF PRESCRIPTION CHARGES

Do not write anything in this space

Certificate No. C40165670S

VALID from 04.04.2000

to (inclusive) 04.10.2002

SURNAME Mr. ~~Mrs~~ Miss Ms CHUANG

FIRST NAME WUKUNG

OTHER NAMES NEIL

DATE OF BIRTH

Day	Month	Year
30	11	1972

Number of last certificate of prepayment of prescription charges (if any)

No.

National Health Service Number (as shown on medical card)

ADDRESS FLAT 5, 20 MILNER
SHERWOOD, NOTTINGHAM

PERMANENT ADDRESS (if different from above)
107, SHU-LIN ST., TAIWAN

PREVIOUS ADDRESS
(if changed since last medical card issued)
If recently arrived from abroad state date
If recently discharged from H.M. Forces state date

Your National Health Service general practitioner (if none, write 'NONE') DOCTOR'S NAME AND ADDRESS

To the Health Authority
I enclose *Postal Order/*Cheque for £ 50 (insert amount)[1] (made payable to "**HM Paymaster General**" and crossed "**Payee Only**")
in prepayment of prescription charges for *FOUR MONTHS *TWELVE MONTHS (*delete as required)
starting from (date) 02.04.2001
I have read and understood the notes overleaf, and I am fully aware of the circumstances and time limits in which a refund may be made.

Signed Wukung Chung Date 02.04.2001

[1] Note: You can find out the amount of the charge from leaflet HC12 available from your doctor, main post office or Social Security office.

DEPARTMENT OF COMPUTING
REQUISITION FORM

The Nottingham Trent University

Date 28.02.2001
Order Number ES6600113
Requested By WING WONG

SUPPLIERS DETAILS

CONTACT NAME SHAOMIN TAN
COMPANY NAME TRENT COMPUTER
ADDRESS 10 FEUERIL STREET
GEDING
NOTTINGHAM NG4 1BU

TEL 0115 - 8484966
FAX 0115 - 9491001

FOR WHO/WHAT/WHERE/WHY

NASSER SHERKAT / MANCHESTER

WHO PAYS DAPID BROWN

THE NOTTINGHAM TRENT UNIVERSITY
REQUISITION FOR CROSS CHARGED SERVICES

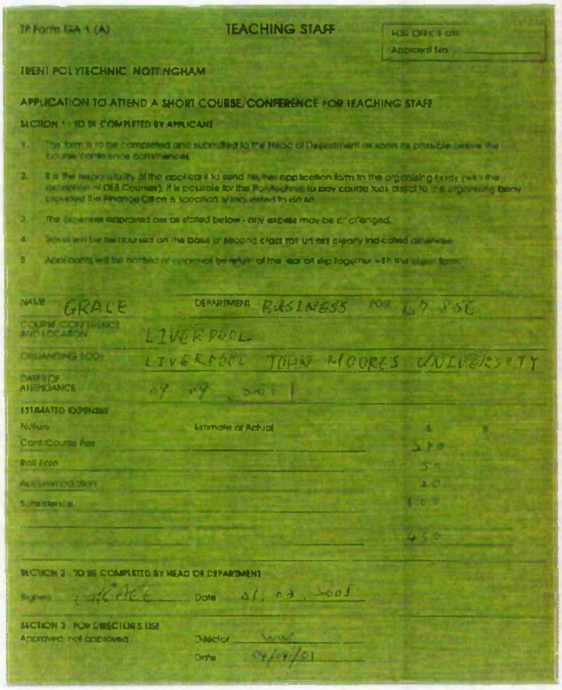
DEPT. PROVIDING SERVICE COMPUTING NAME NEIL
DETAILS UP GRADE SYSTEM
DATE 20.03.2001
DEPT. REQUESTING SERVICE MECHANICAL NAME ROBIN

FINANCE CODES FOR CROSS CHARGING

CHARGE TO (EXPENDITURE)	£	p	CREDIT TO (INCOME)	£	p
	86	50			
	70	00			

SIGNED (AUTHORISED BUDGET HOLDER) Wukung Chung

DISTRIBUTION
REQUESTING DEPT: RETAIN YELLOW, SEND BLUE/WHITE TO PROVIDING DEPT.
PROVIDING DEPT: RETAIN BLUE, SEND WHITE TO MANAGEMENT ACCOUNTS



TP Form GA 1 (A)

TEACHING STAFF

FOR OFFICE USE
Approval No.

TRENT POLYTECHNIC NOTTINGHAM

APPLICATION TO ATTEND A SHORT COURSE/CONFERENCE FOR TEACHING STAFF

SECTION 1 - TO BE COMPLETED BY APPLICANT

1. This form is to be completed and submitted to the Head of Department as soon as possible before the course/conference commences.
2. It is the responsibility of the applicant to send his/her application form to the organising body (with the exception of [illegible] Courses). It is possible for the Polytechnic to pay course fees direct to the organising body provided the Finance Office is [illegible] requested to do so.
3. The expenses approved are as stated below - only excess may be [illegible] charged.
4. Travel will be reimbursed on the basis of second class rail unless clearly indicated otherwise.
5. Applicants will be notified of approval by return of the [illegible] slip together with the claim form.

NAME GRACE DEPARTMENT BUSINESS [illegible]

COURSE/CONFERENCE AND LOCATION LIVERPOOL

ORGANISING BODY LIVERPOOL JOHN MOORES UNIVERSITY

DATES OF ATTENDANCE 09 09 2001

ESTIMATED EXPENSES

Nature	Estimate or Actual	£	p
Conf/Course Fee		270	
Rail Fare		50	
Accommodation		20	
Subsistence		100	
		450	

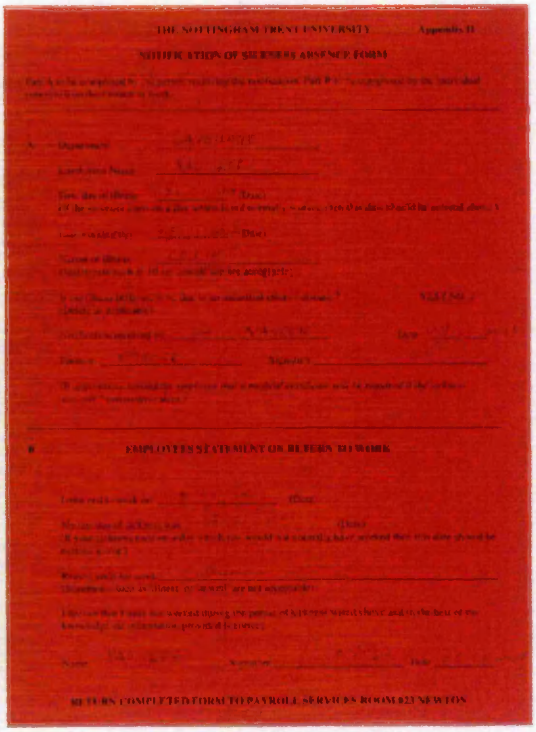
SECTION 2 - TO BE COMPLETED BY HEAD OF DEPARTMENT

Signed GRACE Date 01. 03. 2001

SECTION 3 - FOR DIRECTOR'S USE

Approved, not approved Director [illegible]

Date 04/09/01

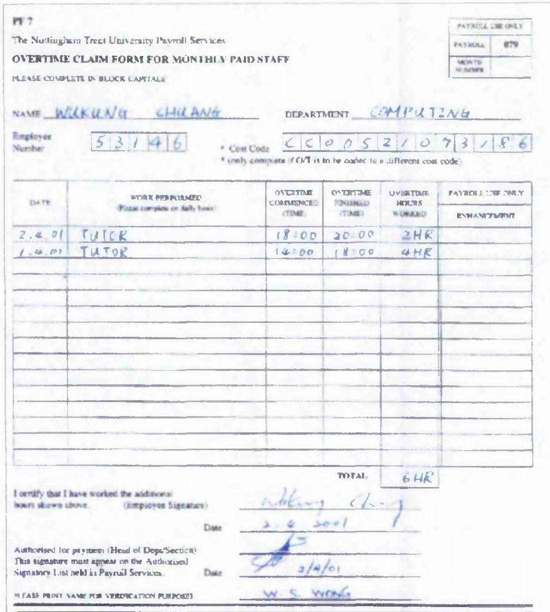


THE NOTTINGHAM TRENT UNIVERSITY Appendix II

NOTIFICATION OF SICKNESS ABSENCE FORM

B EMPLOYEES STATEMENT OF RETURN TO WORK

RETURN COMPLETED FORM TO PAYROLL SERVICES ROOM 023 NEWTON



PF 7

The Nottingham Trent University Payroll Services

OVERTIME CLAIM FORM FOR MONTHLY PAID STAFF

PLEASE COMPLETE IN BLOCK CAPITALS

PAYROLL USE ONLY	
PAYROLL	079
MONTH NUMBER	

NAME WUKUNG CHUANG DEPARTMENT COMPUTING

Employee Number 5 3 1 4 6 * Cost Code C C 0 0 5 2 / 0 7 3 / 8 6

* (only complete if O/T is to be coded to a different cost code)

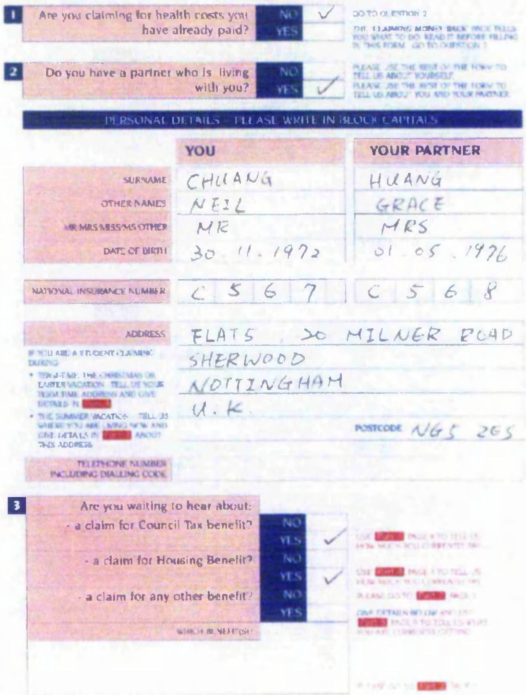
DATE	WORK PERFORMED (Please complete on daily basis)	OVERTIME COMMENCED (TIME)	OVERTIME FINISHED (TIME)	OVERTIME HOURS WORKED	PAYROLL USE ONLY ENHANCEMENT
2.4.01	TUTOR	18:00	20:00	2HR	
1.4.01	TUTOR	14:00	18:00	4HR	
			TOTAL	6HR	

I certify that I have worked the additional hours shown above. (Employee Signature)

Date 2. 4 2001

Authorised for payment (Head of Dept/Section)
This signature must appear on the Authorised Signatory List held in Payroll Services. Date 2/4/01

PLEASE PRINT NAME FOR VERIFICATION PURPOSES W. S. WONG



1 Are you claiming for health costs you have already paid? NO ✓ YES

GO TO QUESTION 2

THE CLAIMING MONEY BACK PAGE TELLS YOU WHAT TO DO. READ IT BEFORE FILLING IN THIS FORM. GO TO QUESTION 2

2 Do you have a partner who is living with you? NO YES ✓

PLEASE USE THE REST OF THE FORM TO TELL US ABOUT YOURSELF

PLEASE USE THE REST OF THE FORM TO TELL US ABOUT YOU AND YOUR PARTNER

PERSONAL DETAILS - PLEASE WRITE IN BLOCK CAPITALS

	YOU	YOUR PARTNER
SURNAME	CHUANG	HUANG
OTHER NAMES	NEIL	GRACE
MR/MRS/MISS/MS/OTHER	MR	MRS
DATE OF BIRTH	30. 11. 1972	01. 05. 1976
NATIONAL INSURANCE NUMBER	C 5 6 7	C 5 6 8

ADDRESS FLAT 5, 20 MILNER ROAD
SHERWOOD
NOTTINGHAM
U.K.

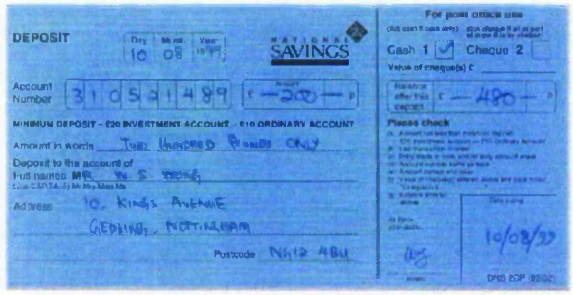
POSTCODE NG5 2GS

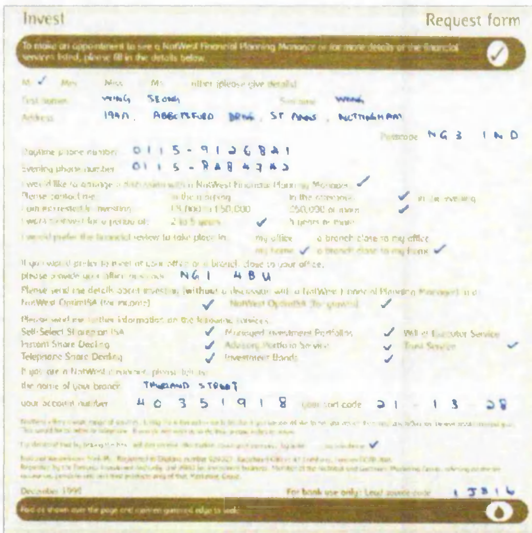
TELEPHONE NUMBER INCLUDING DIALLING CODE

3 Are you waiting to hear about:
- a claim for Council Tax benefit? NO YES ✓
- a claim for Housing Benefit? NO YES ✓
- a claim for any other benefit? NO YES

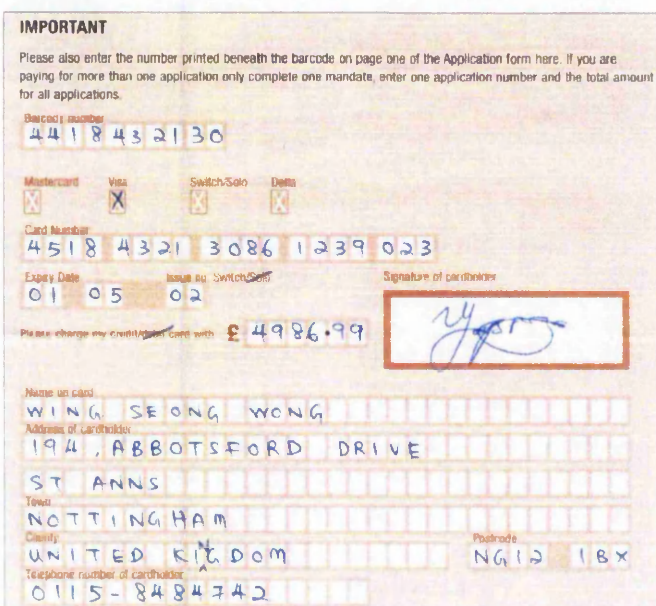
WHICH BENEFIT(S)?


Despatch Certificate (PPU 25)
PARCEL FORCE WORLDWIDE
Standard
Number of parcels received 1
Enhanced Compensation (Mark 'X' if required)
Bar-code
Name and Address (as written on parcel)
Postcode
0208472591
W. S. WONG, 194, ABBOTSFORD DRIVE ST ANNS NOTTINGHAM
NG3 1ND
11/1/00
PLEASE KEEP THIS CERTIFICATE SAFE. IT MUST BE PRODUCED IN THE EVENT OF A CLAIM
11/1/00


DEPOSIT
10 08 1999
NATIONAL SAVINGS
Account Number 3 1 0 5 2 1 4 8 9
£ — 200 — p
MINIMUM DEPOSIT - £20 INVESTMENT ACCOUNT - £10 ORDINARY ACCOUNT
Amount in words TWO HUNDRED POUNDS ONLY
Deposit to the account of Full names MR. W S WONG
Address 10, KINGS AVENUE GEDLING, NOTTINGHAM
Postcode NG12 4BU
For post office use
Cash 1 ✓ Cheque 2
£ — 480 —
Please check
10/08/99


Invest
Request form
WING SEONG
WONG
194A, ABBOTSFORD DRIVE, ST ANNS, NOTTINGHAM
NG3 1ND
0115-9126841
0115-8484742
NG1 4BU
THURLAND STREET
40351918
21-13-28


Apply for your Tesco credit card today
TESCO Personal Finance
MR WING S
WONG
HSBC
403519
01 03 31
02 03
194 ABBOTSFORD DRIVE
ST ANNS
4019184321308712 33
Employment details
THE NOTTINGHAM TRENT UNI
BURTON STREET
0115 8484742
25500
Card Protection
Additional Cardholder
MRS YAP SIEW
SIEW SIAN YAP
11/01/01


IMPORTANT
Please also enter the number printed beneath the barcode on page one of the Application form here. If you are paying for more than one application only complete one mandate, enter one application number and the total amount for all applications.
Barcode number
441843 2130
Mastercard
Visa X
Switch/Solo
Delta
Card Number
4518 4321 3086 1239 023
Expiry Date 01 05
Issue no. Switch/Solo 02
Signature of cardholder
Please charge my credit/debit card with £ 4986.99
Name on card
WING SEONG WONG
Address of cardholder
194, ABBOTSFORD DRIVE
ST ANNS
Town
NOTTINGHAM
County
UNITED KINGDOM
Postcode
NG12 1BX
Telephone number of cardholder
0115-8484742

02 Read Note 02

Date of travel

13 05 2001

03 Read Note 03

Details of applicant (intended passport holder)

WONG

WING SEONG

WS

194, ABBOTSFORD DRIVE

ST ANNS

NOTTINGHAM

UNITED KINGDOM

NG14 14X

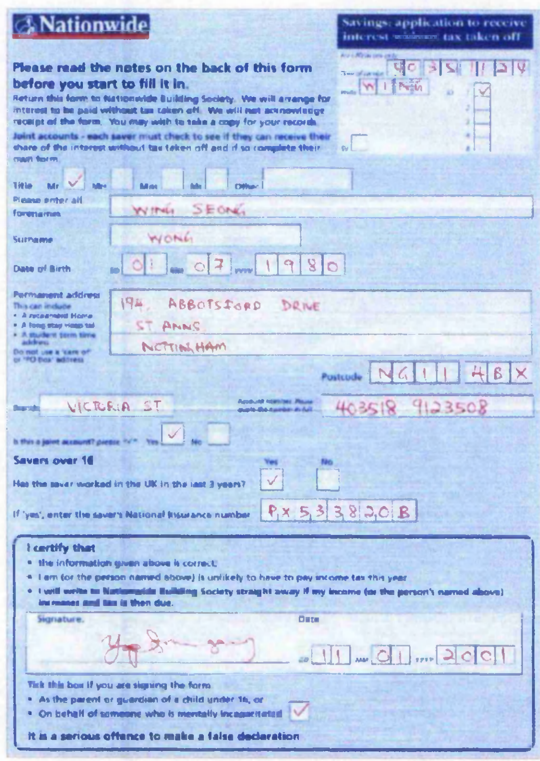
01 07 1972

KUALA LUMPUR

MALAYSIA

0115-848 4742

0115-8484743



Nationwide

Savings: application to receive interest without tax taken off

Please read the notes on the back of this form before you start to fill it in.

Return this form to Nationwide Building Society. We will arrange for interest to be paid without tax taken off. We will not acknowledge receipt of the form. You may wish to take a copy for your records.

Joint accounts - each saver must check to see if they can receive their share of the interest without tax taken off and if so complete their own form.

Title Mr ✓ Mrs Miss Ms Other

Please enter all forenames: WING SEONG

Surname: WONG

Date of Birth: 01 07 1980

Permanent address: 194, ABBOTSFORD DRIVE, ST ANNS, NOTTINGHAM

Postcode: NG11 4BX

Branch: VICTORIA ST

Account number: 403518 9123508

Is this a joint account? Yes ✓ No

Savers over 16

Has the saver worked in the UK in the last 3 years? Yes ✓ No

If 'yes', enter the saver's National Insurance number: PX533820B

I certify that

- the information given above is correct;
- I am (or the person named above) is unlikely to have to pay income tax this year
- I will write to Nationwide Building Society straight away if my income (or the person's named above) increases and tax is then due.

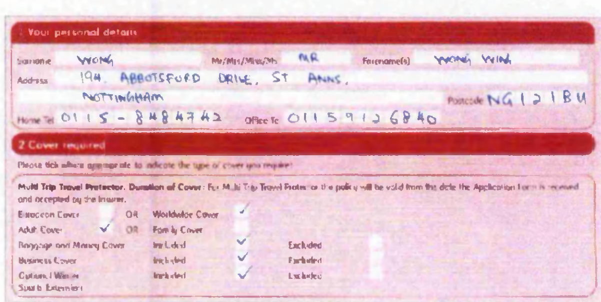
Signature

Date 11 01 2001

Tick this box if you are signing the form

- As the parent or guardian of a child under 16, or
- On behalf of someone who is mentally incapacitated ✓

It is a serious offence to make a false declaration



Your personal details

Surname: WONG Mr/Mrs/Miss/Ms: MR Forename(s): WONG WING

Address: 194, ABBOTSFORD DRIVE, ST ANNS, NOTTINGHAM

Postcode: NG12 1BU

Home Tel: 0115-8484742 Office Tel: 01159126840

2 Cover required

Please tick where appropriate to indicate the type of cover you require

Multi Trip Travel Protector. Duration of Cover: For Multi Trip Travel Protector the policy will be valid from the date the Application Form is received and accepted by the Insurer.

European Cover OR Worldwide Cover ✓

Adult Cover ✓ OR Family Cover

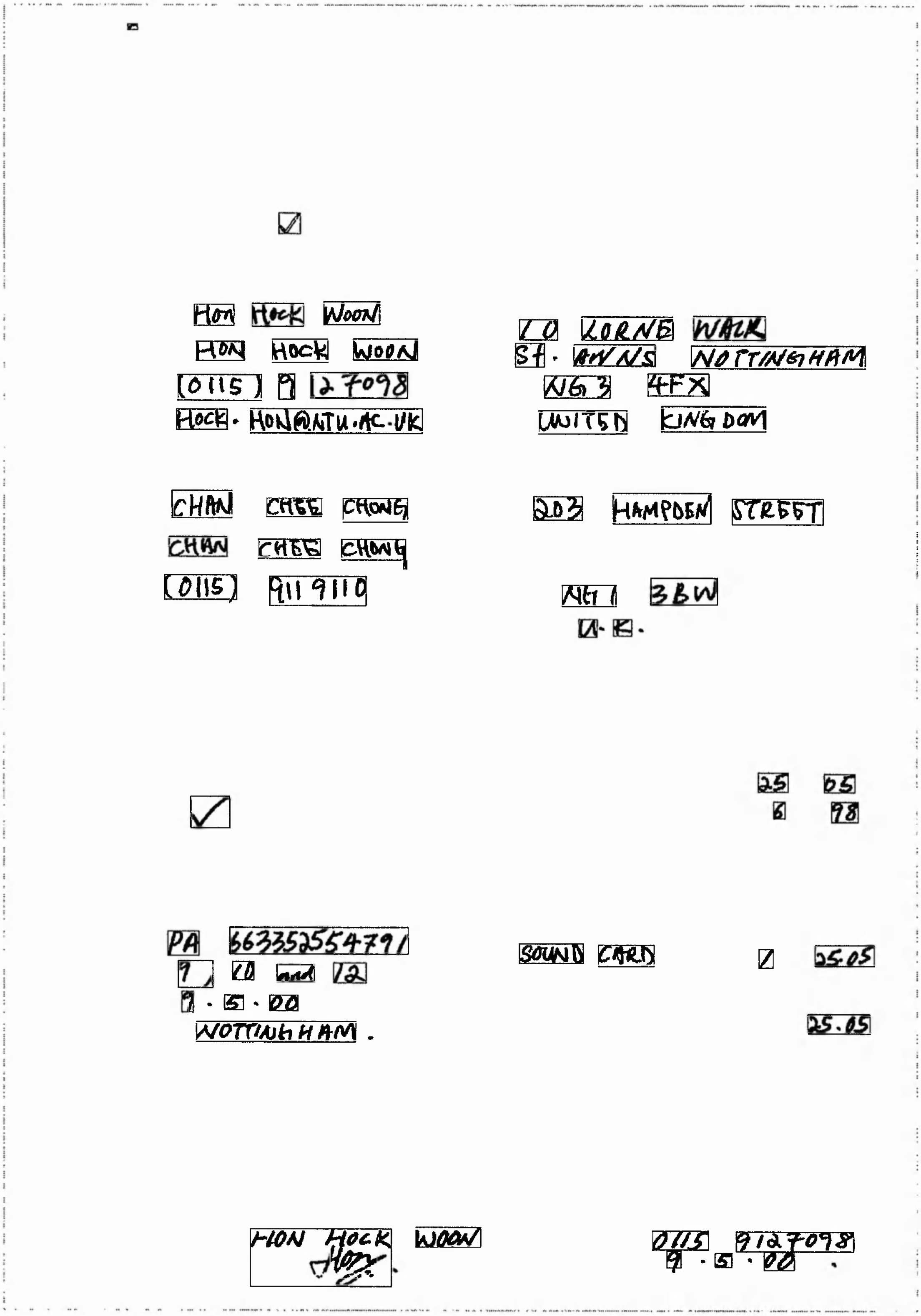
Baggage and Money Cover Included ✓ Excluded

Business Cover Included ✓ Excluded

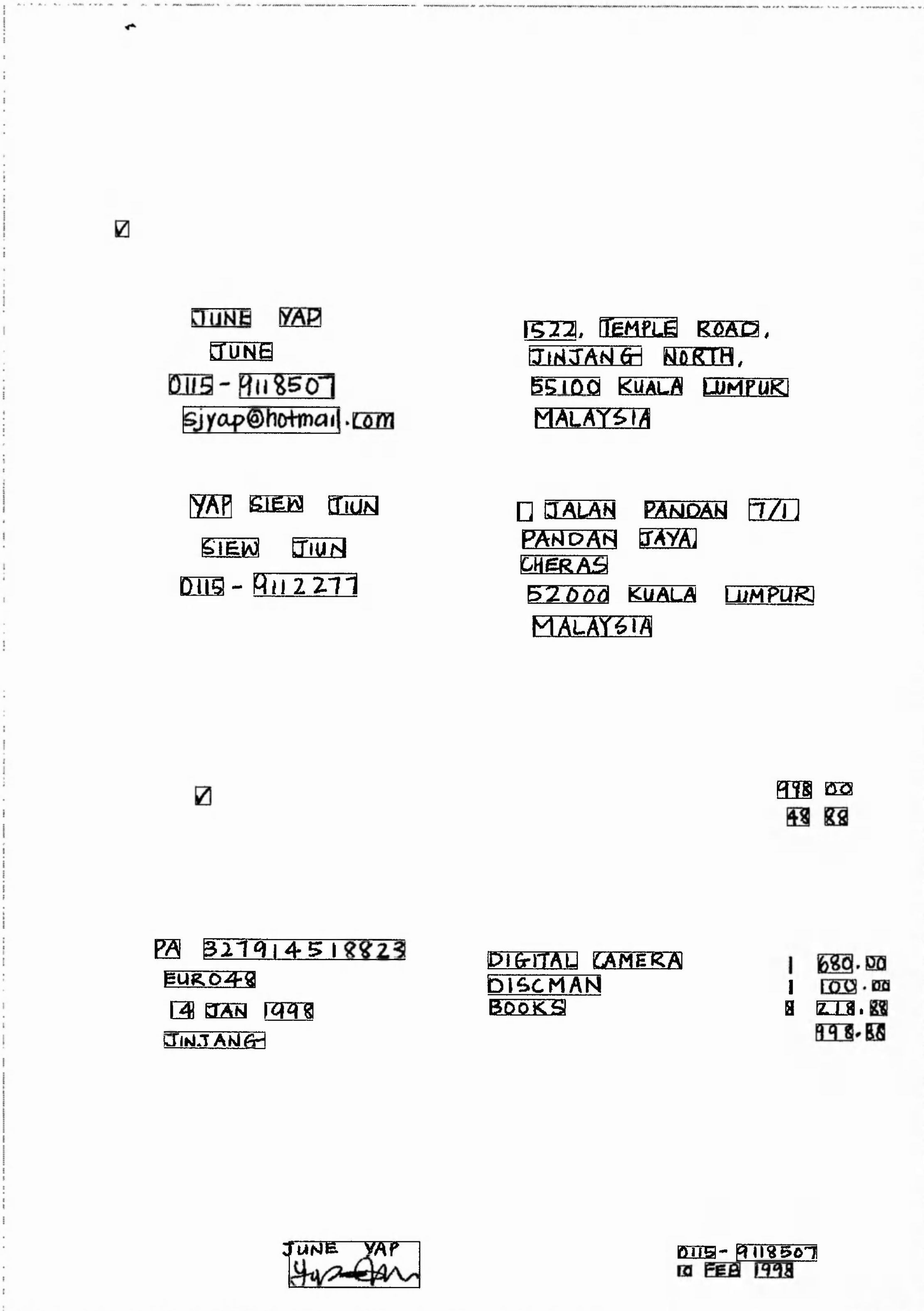
Optional Winter Sports Extension Included ✓ Excluded

APPENDIX E - Extracted Output

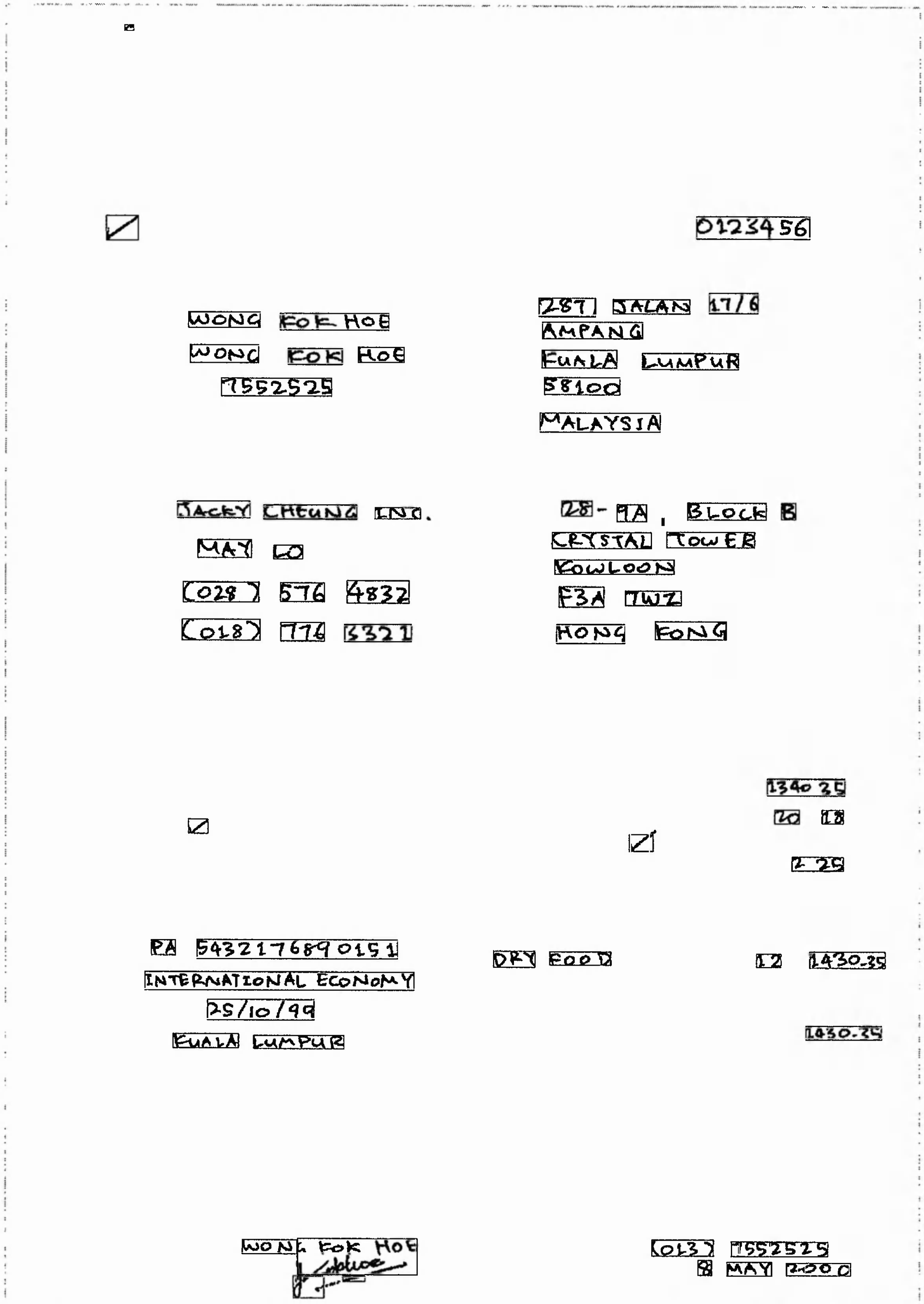
Form01 (writer 1) extracted words



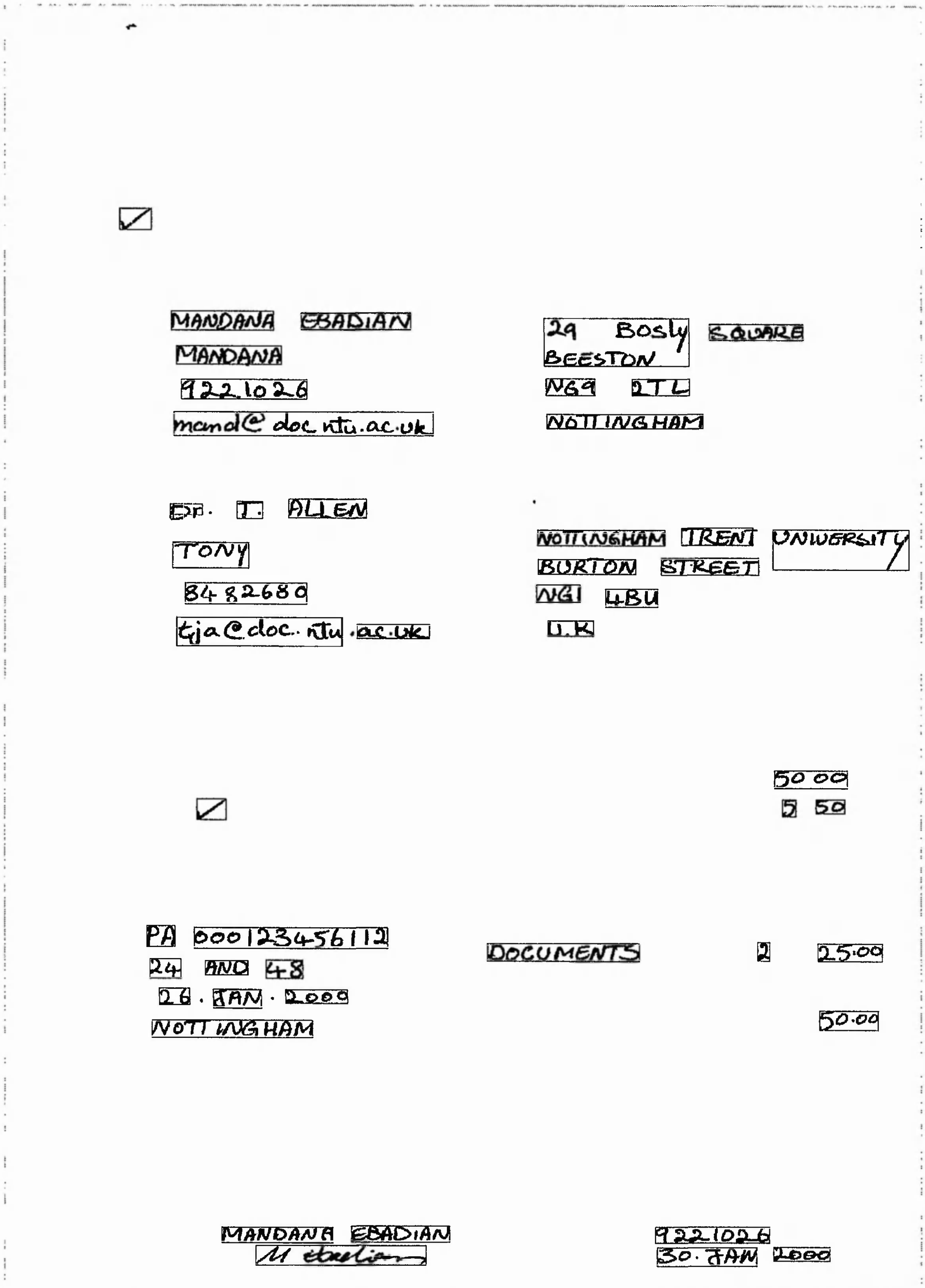
Form01 (writer 2) extracted words



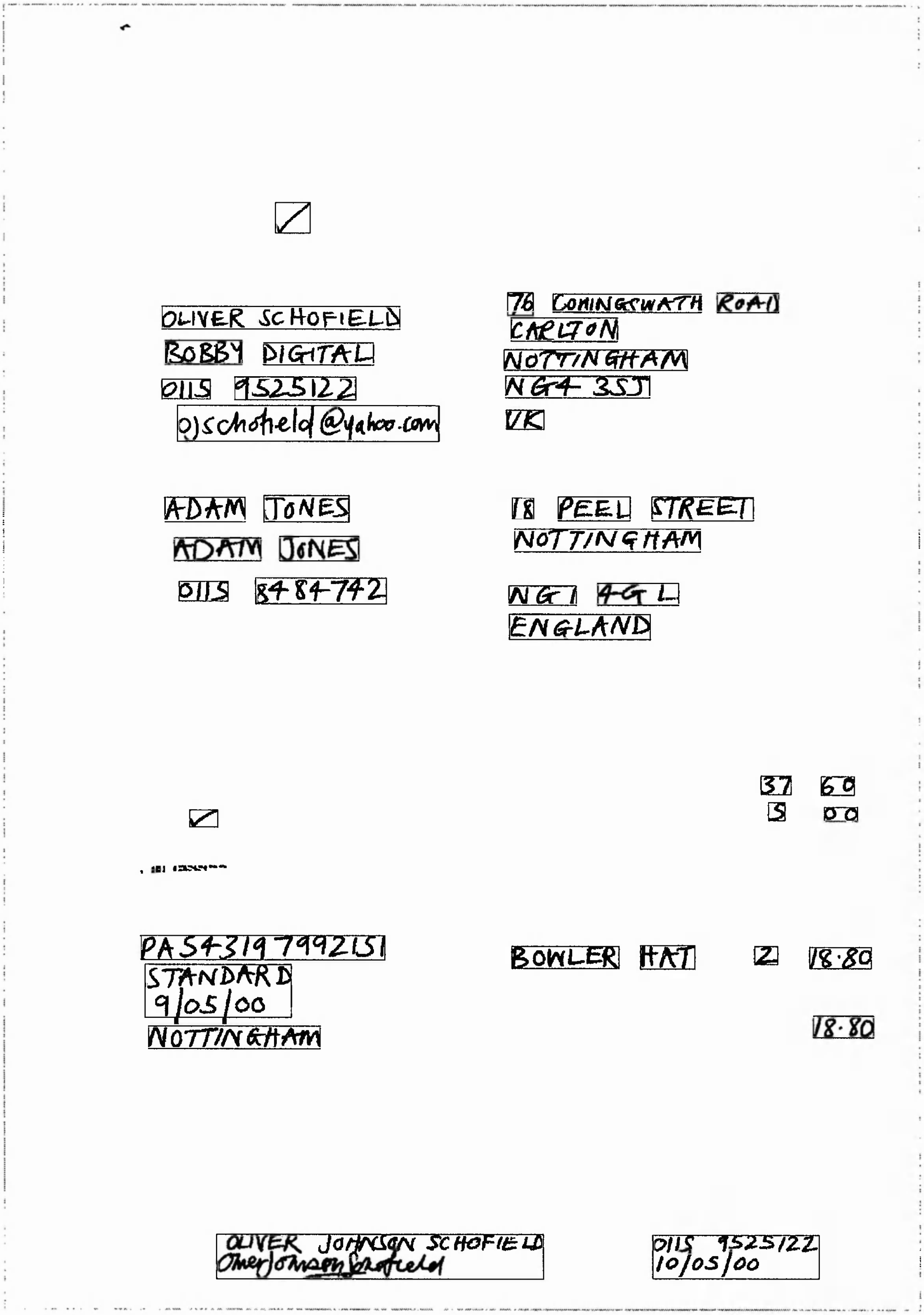
Form01 (writer 3) extracted words



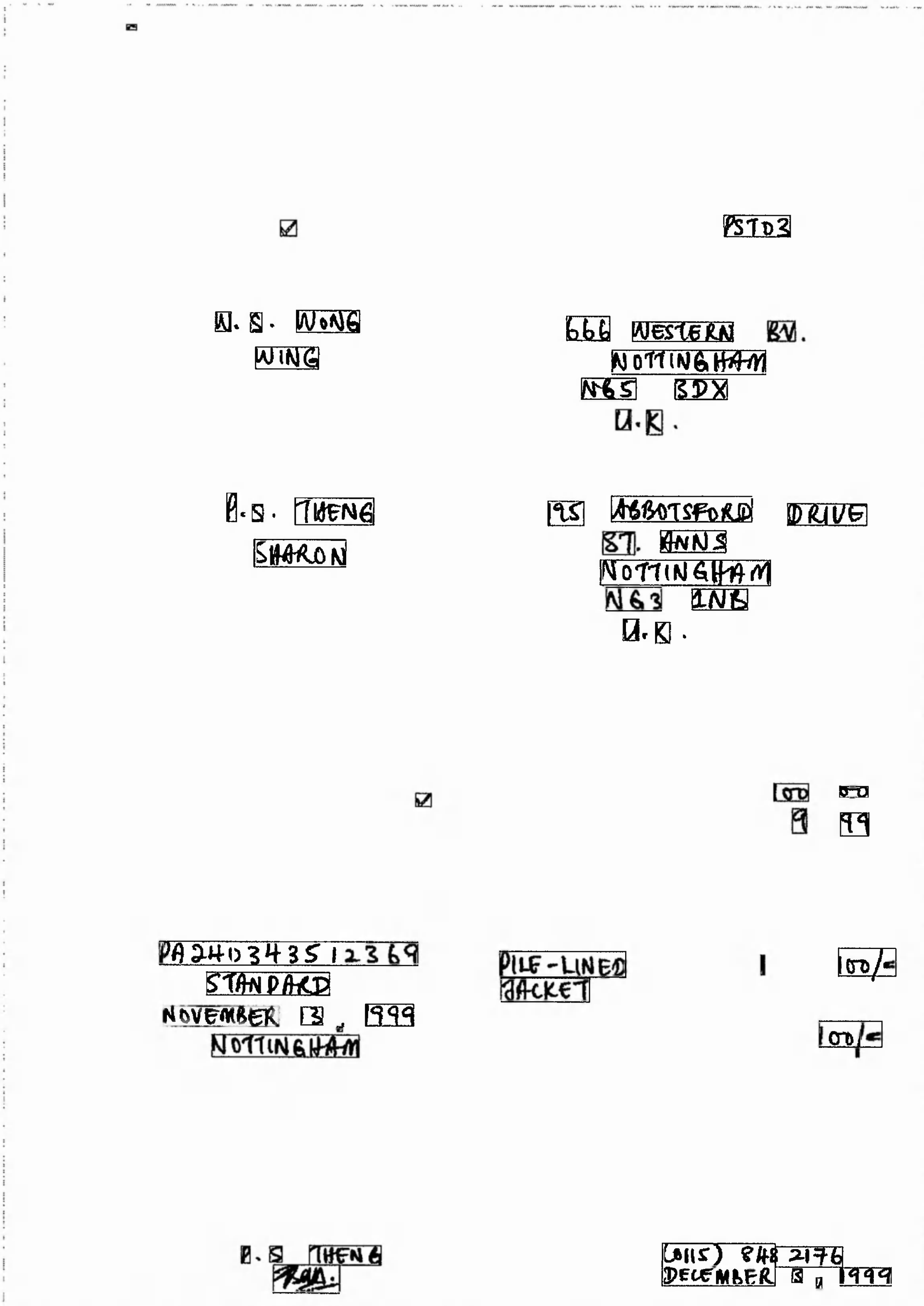
Form01 (writer 4) extracted words



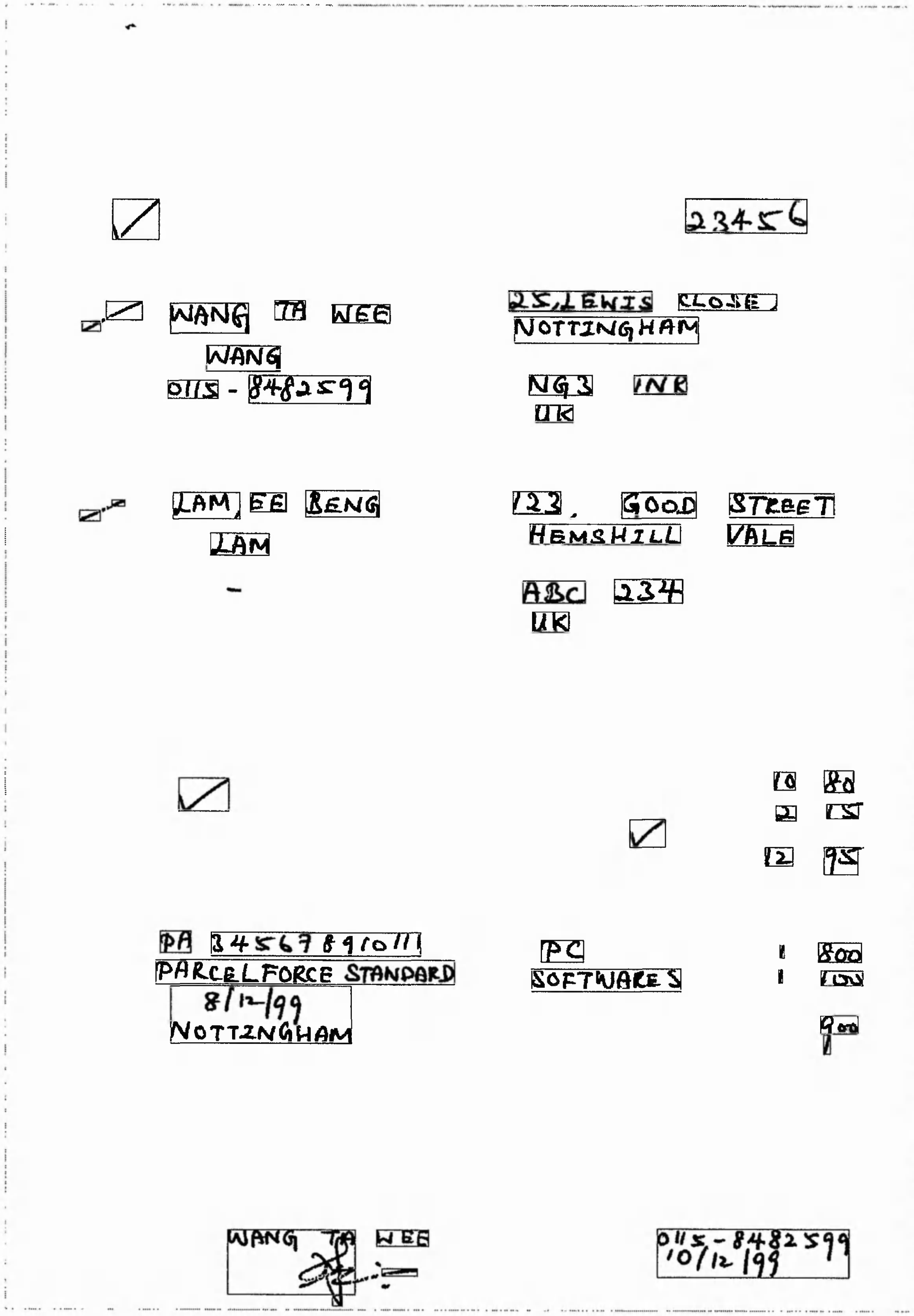
Form01 (writer 5) extracted words



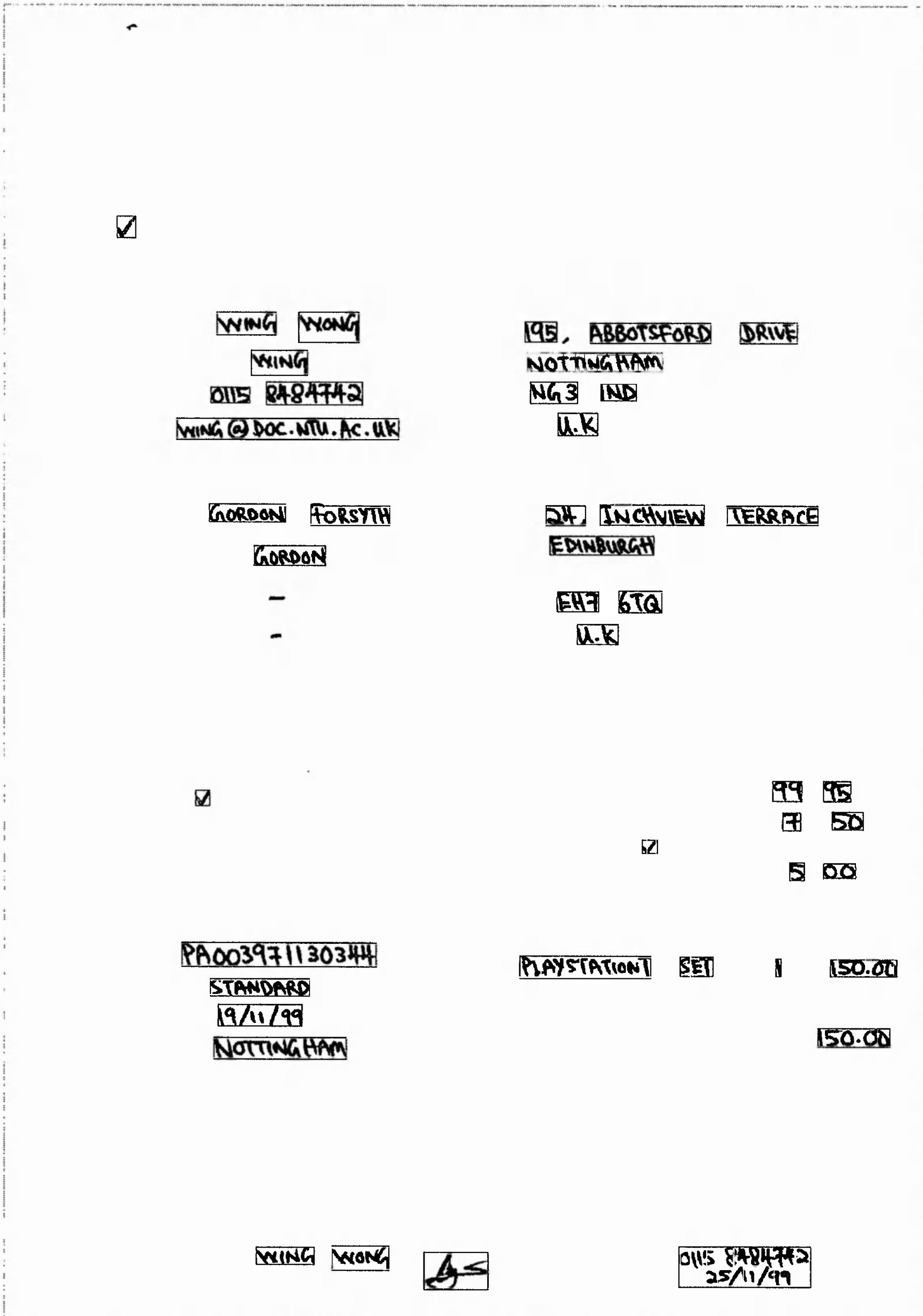
Form01 (writer 6) extracted words



Form01 (writer 7) extracted words



Form01 (writer 8) extracted words



WING WONG
WING
0115 8484742
WING@DOC.NTU.AC.UK

195, ABBOTSFORD DRIVE
NOTTINGHAM
NG3 1ND
U.K

GORDON FORSYTH
GORDON
-
-

24, INCHVIEW TERRACE
EDINBURGH
EH7 6TQ
U.K

99 95
7 50
5 00

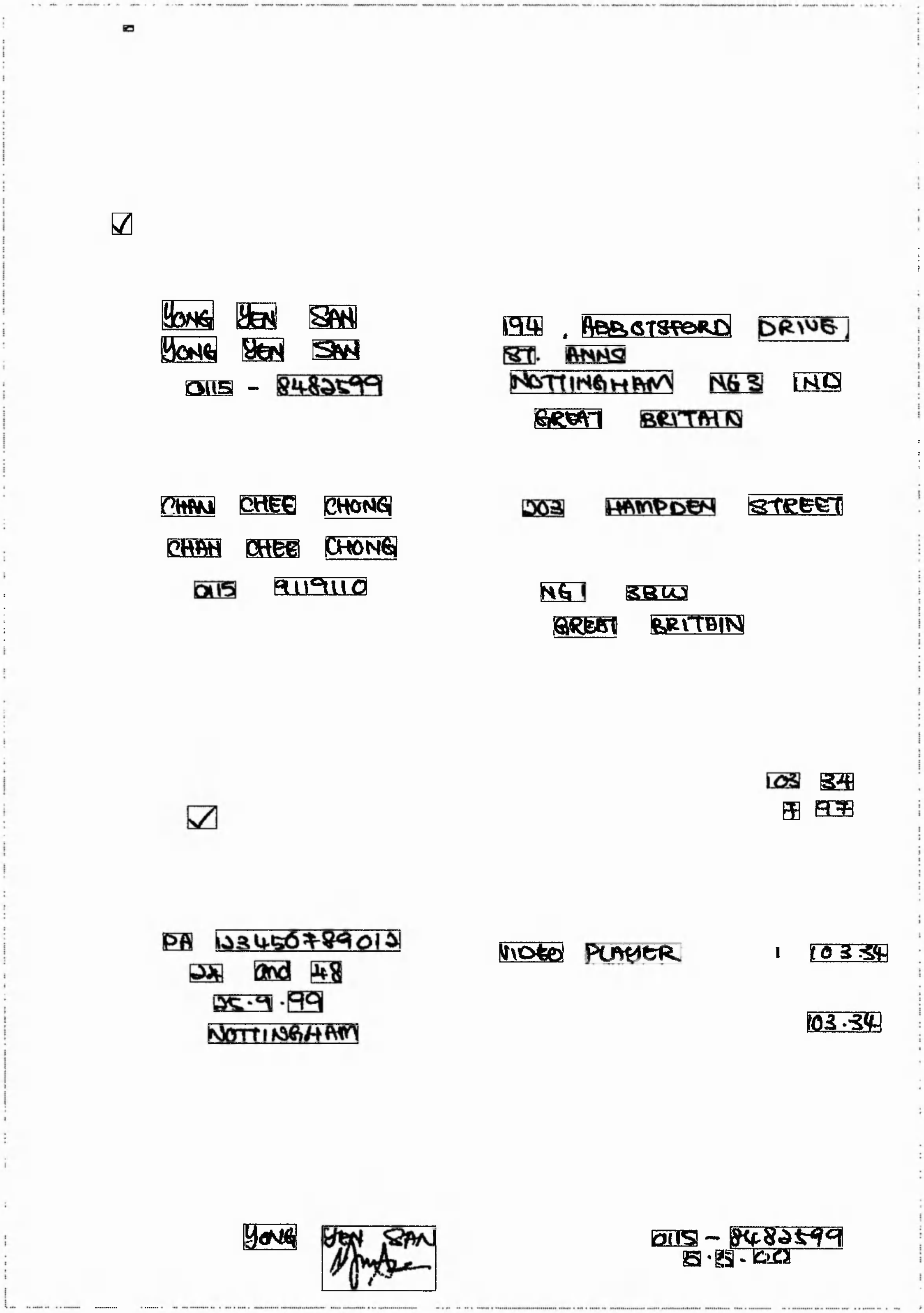
PA003971130344
STANDARD
19/11/99
NOTTINGHAM

PLAYSTATION SET 1 150.00
150.00

WING WONG

0115 8484742
25/11/99

Form01 (writer 9) extracted words



Form01 (writer 10) extracted words

☑

SHAOMIN ZHANG
SHAOMIN
0115 - 8484742
shaomin.zhang@ntu.ac.uk

Burton StrEET
NOTTINGHAM
NG1 4BU
UK

STEVE LIM
STEVE
0115 - 9418419

HOLE STREET
NOTTINGHAM
NG7 7AY
UK

☑

100 00
50 00

PA 000245123442
STANDARD
JUNE 14, 1999
LEEDS

GOLF SET

1 100.00

100.00

SHAOMIN ZHANG
Shaomin zhang

0115-8484742
MAY 23, 2000

Form02 (writer 1) extracted words

LIM KOK WING
194 , PASAR ROAD
CHERAS KUALA LUMPUR
MALAYSIA

29 MAY 1966
912709
48 2727

CHONG WEE YEE
26 JALAN BUNGA RAYA
KULAI 81000 JOHOR
MALAYSIA

TAN AH KONG
10 LORNE WALK
ST. ANN'S NOTTINGHAM
UNITED KINGDOM

607-6633525

919 8893

24 APRIL 1999

SU HUI LOON
153, WELLINGTON
ROAD, TAIWAN. R.O.C

24.5.00

Form02 (writer 2) extracted words

KAREN STONES
55 STOKE GROVE
CARLTON, NOTTINGHAM
NG3 5MQ

8 MARCH 1968
0115-9451792
0115-8764332

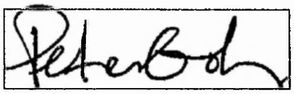
PETER GOH
81 BURTON AVENUE
COLWICK, OXFORD.
OX1 3KS

0455-7723688

GREAME EASTWOOD
71, REGENT STREET
KENSINGTON GARDEN
YORK
YK5 1ZP
0872 - 2457431

8 APRIL 2000

SIMON JAMES
QUEEN HOSPITAL.
KING ROAD
NOTTINGHAM
NG8 3XD

 .

18 MAY 2000

Form02 (writer 3) extracted words

WAI KEE YOON

29, JALAN 14/8

PERSIARAN BARU

MALACCA

MALAYSIA

21 JULY 1966

(03) 323 4318

(03) 718 7917

MEDI-CARE INC.

36-A, BROWNING STREET

FOXTROT

KUANTAN

MALAYSIA

(03) 769 4321

DR. NARAIN SINGH

27, TAMAN DESA

KAMPUNG BARU

KUALA LUMPUR

MALAYSIA

(03) 397 4827

21 AUGUST 1996

LIMP BIZKIT

287, LORONG HAJI

AMPANG

BRUNEI

LimpBizkit

30 NOV. 1997

Form02 (writer 4) extracted words

NIMA DEHMSHKI

15 PEVERILE RD

BEESTON NOTTINGHAM

NG7 2TL

5 MAY 1992

0115 - 9262225

0115 - 9826209

IRIN TERAM

80 CENTRAL AV.

BEESTON

NG7 2TL

0115 - 9250865

CLAIR CLARK

28 CHARLES AV.

CHILWELL

NOTTINGHAM

NG9 2SD

0804 - 2625981

8 MAY 1999

NASSER SHERKAT

9 ABBEY MEDICAL CENTRE

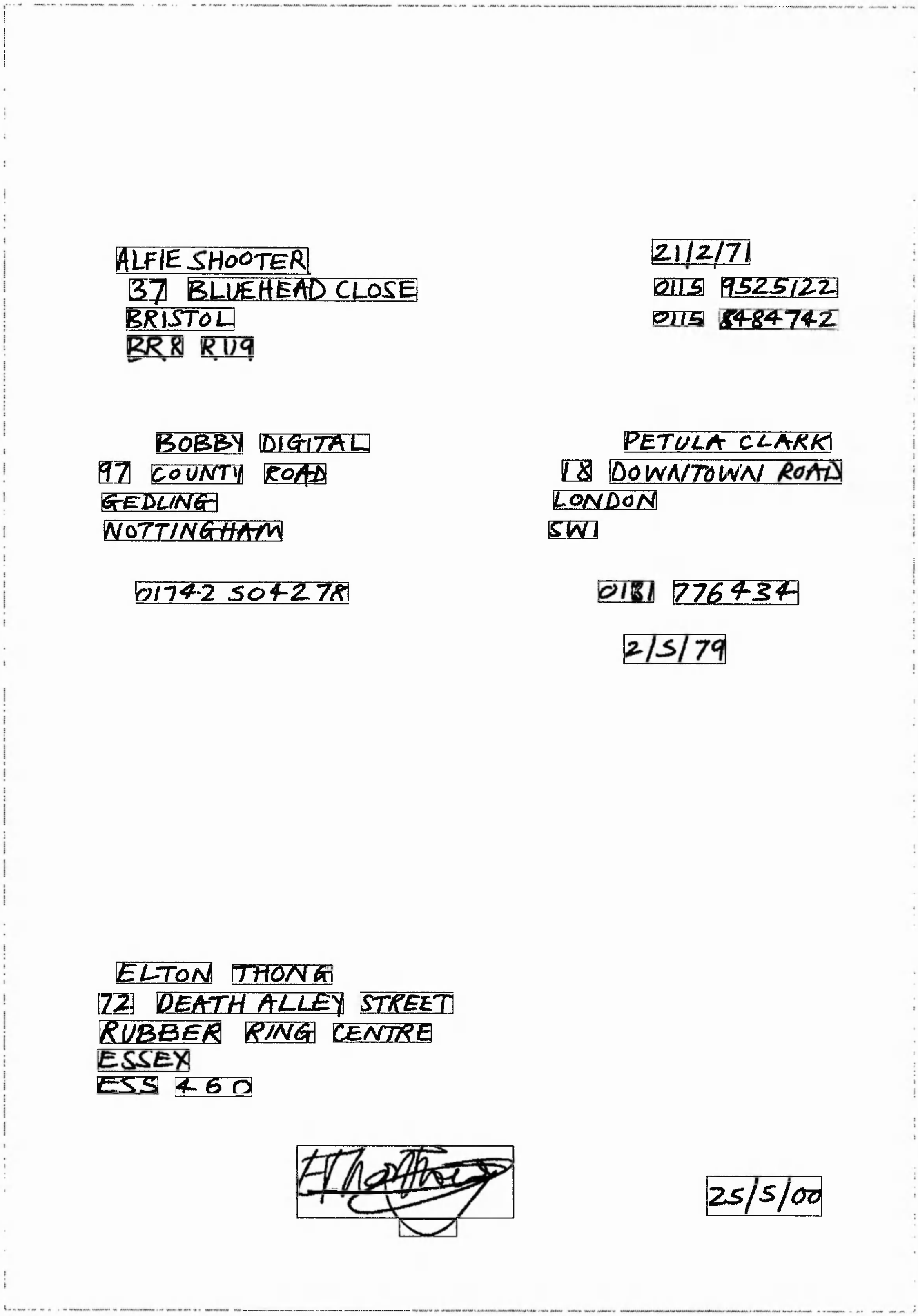
CENTRAL AV.

BEESTON

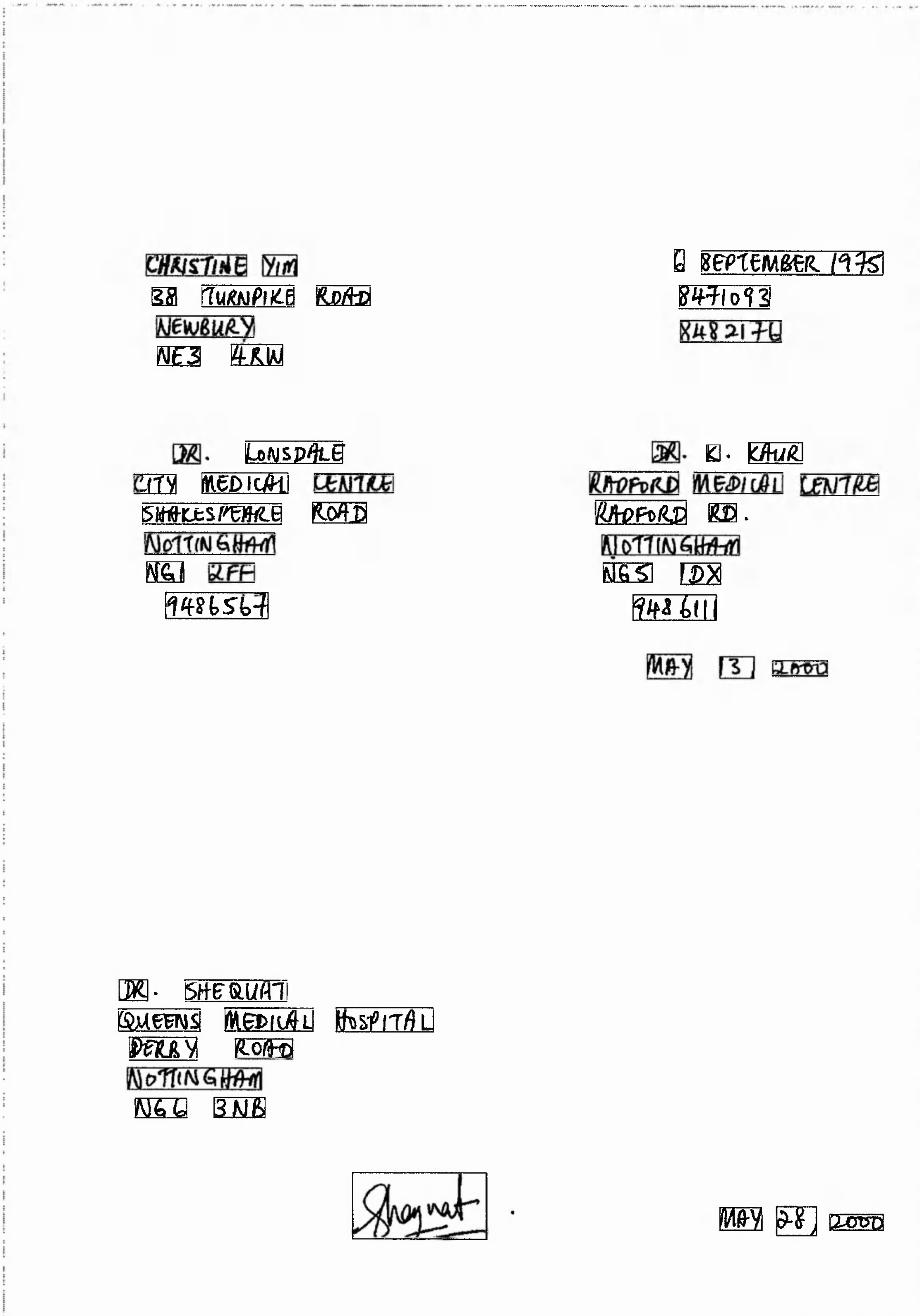
NG7 2TS

18 May 2000

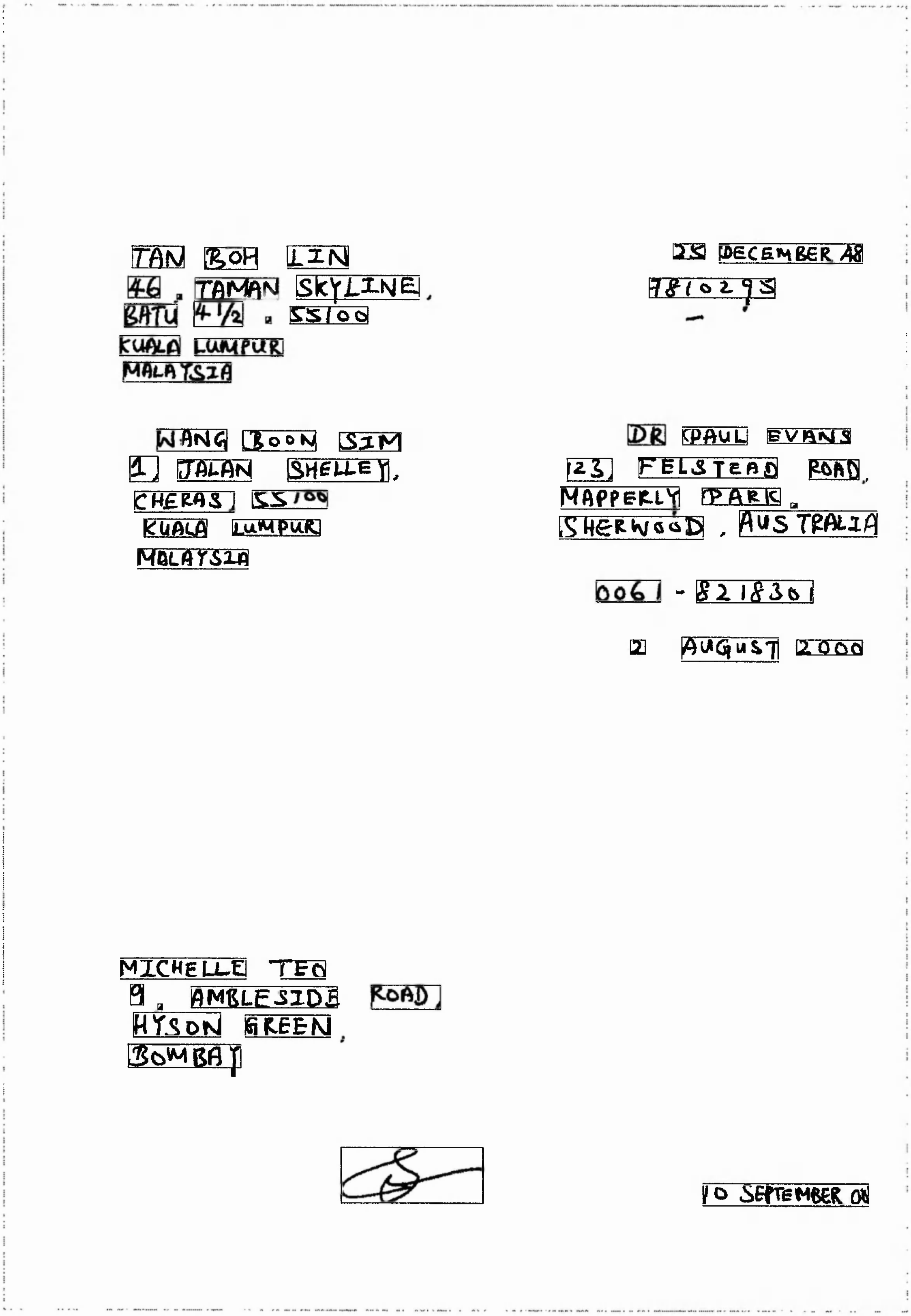
Form02 (writer 5) extracted words



Form02 (writer 6) extracted words



Form02 (writer 7) extracted words


TAN BOH LIN
46, TAMAN SKYLINE,
BATU 4 1/2, 55100
KUALA LUMPUR
MALAYSIA

25 DECEMBER 48
781029S

WANG BOON SIM
1, JALAN SHELLEY,
CHERAS, 55100
KUALA LUMPUR
MALAYSIA

DR PAUL EVANS
123, FELSTEAD ROAD,
MAPPERLY PARK,
SHERWOOD, AUSTRALIA

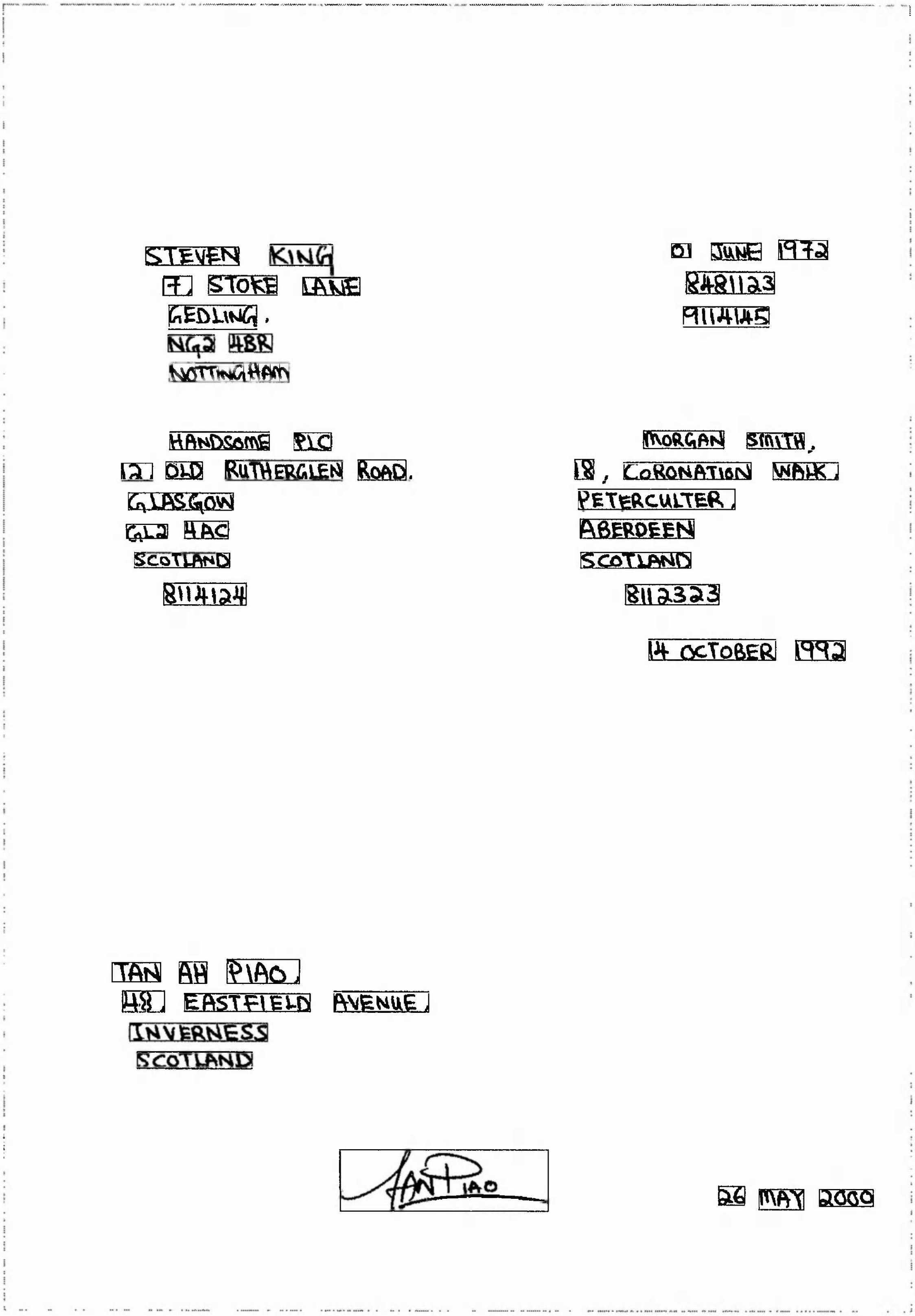
0061 - 8218301

2 AUGUST 2000

MICHELLE TEO
9, AMBLESIDE ROAD,
HYSON GREEN,
BOMBAY

10 SEPTEMBER 00

Form02 (writer 8) extracted words



STEVEN KING
7 STOKE LANE
GEDLING,
NG2 4BR
NOTTINGHAM

01 JUNE 1972
8481123
9114145

HANDSOME PLC
12 OLD RUTHERGLEN ROAD.
GLASGOW
GL2 4AC
SCOTLAND
8114124

MORGAN SMITH,
18, CORONATION WALK
PETERCULTER
ABERDEEN
SCOTLAND
8112323

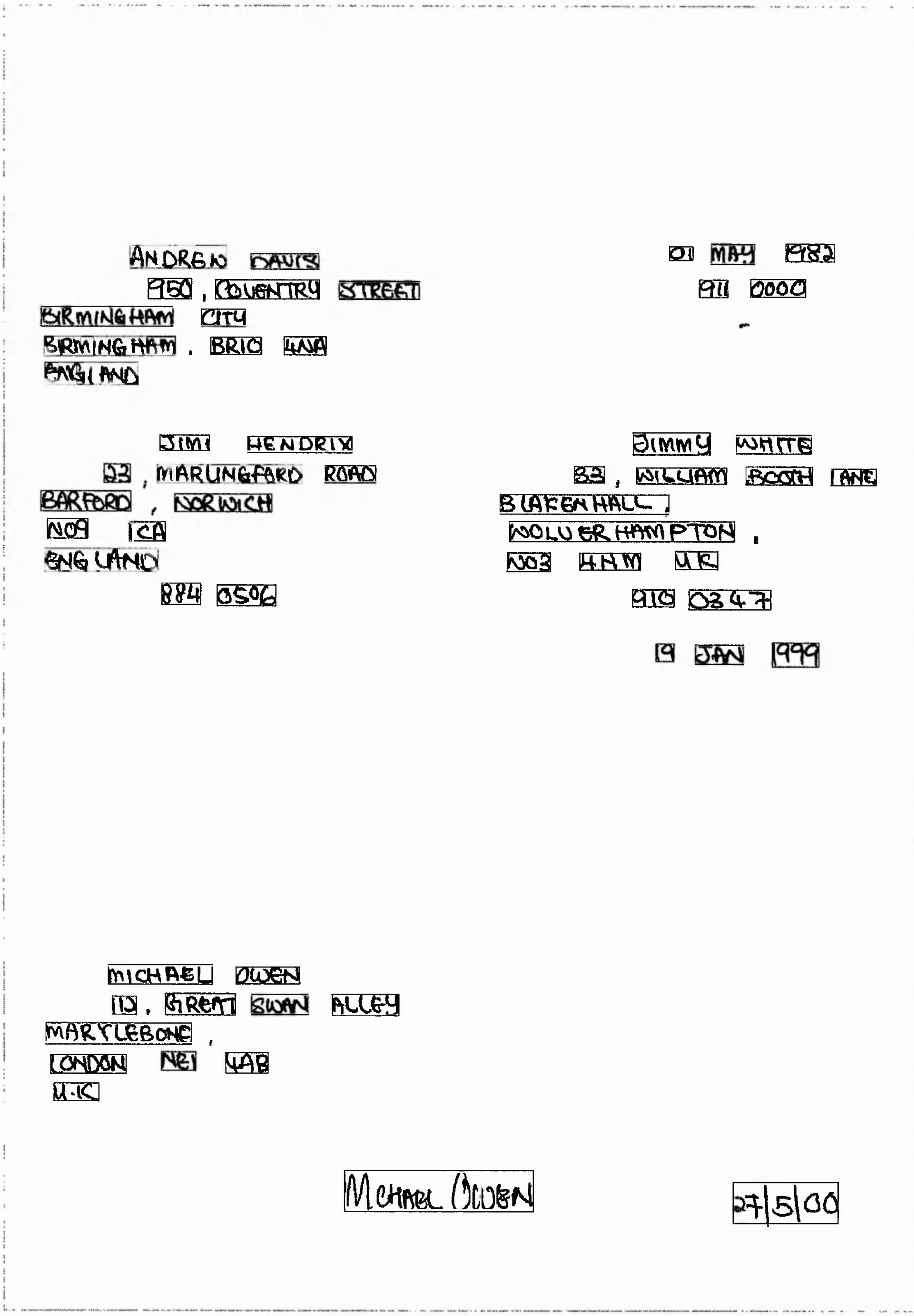
14 OCTOBER 1992

TAN AH PIAO
48 EASTFIELD AVENUE
INVERNESS
SCOTLAND

TAN PIAO

26 MAY 2000

Form02 (writer 9) extracted words



Form02 (writer 10) extracted words

[redacted] JOE COLE

01 MAY 1982
9118472

23 DAVYHULME ROAD EAST
STRETFORD, MANCHESTER
MA2 3BU, ENGLAND

JOHN

BRIAN

18 SANERSONS CRAFT
LEIGH, MANCHESTER
MA4 INP, UK

KESWICK ROAD
CRINGLEFORD
NORWICH NO8 4AA
842 12 13

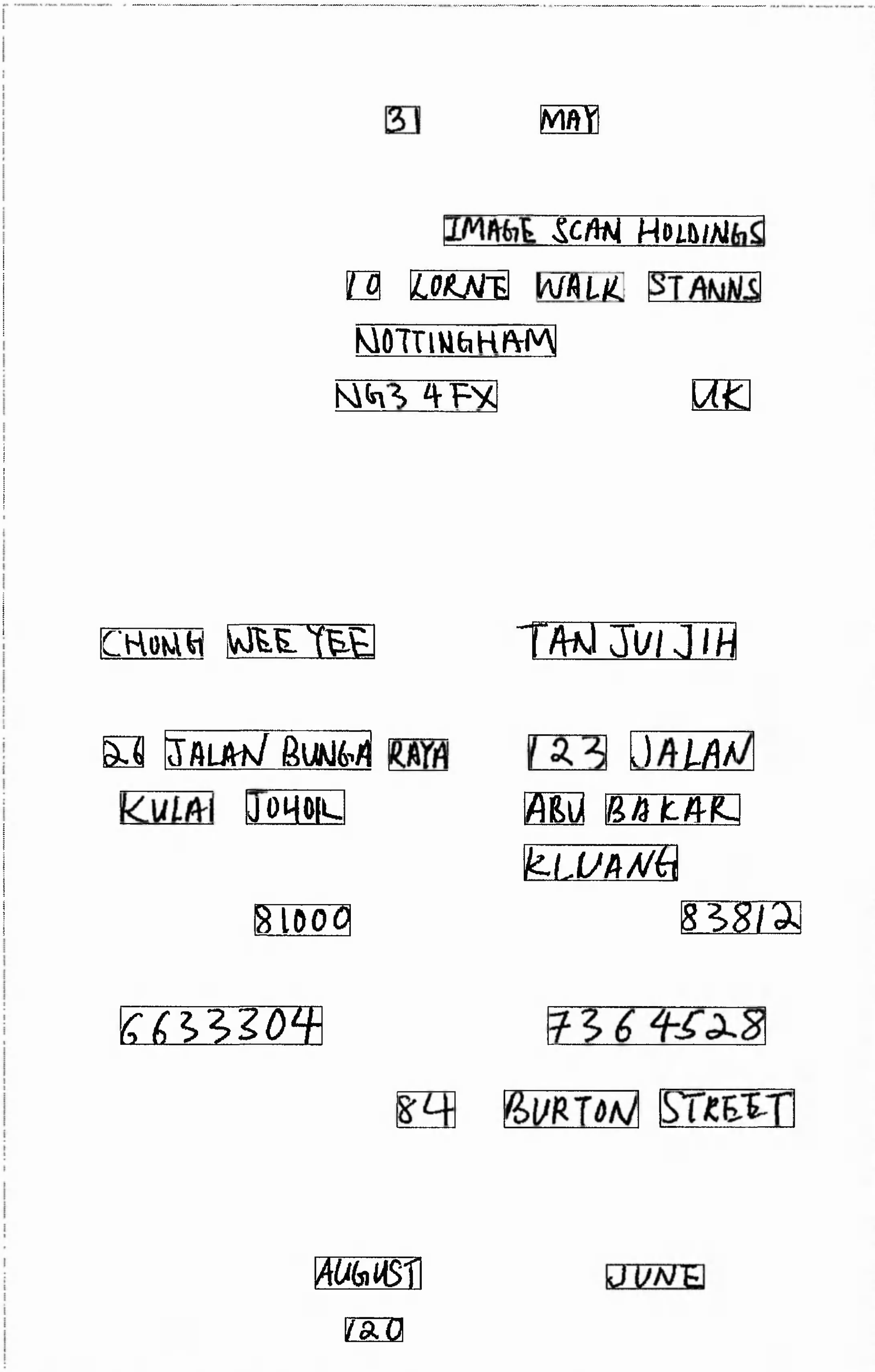
22 AUGUST 1999

ROBSON GUEST
23A, YORK STREET
STANNINGTON
SHEFFIELD
SH3 4BA U.K.

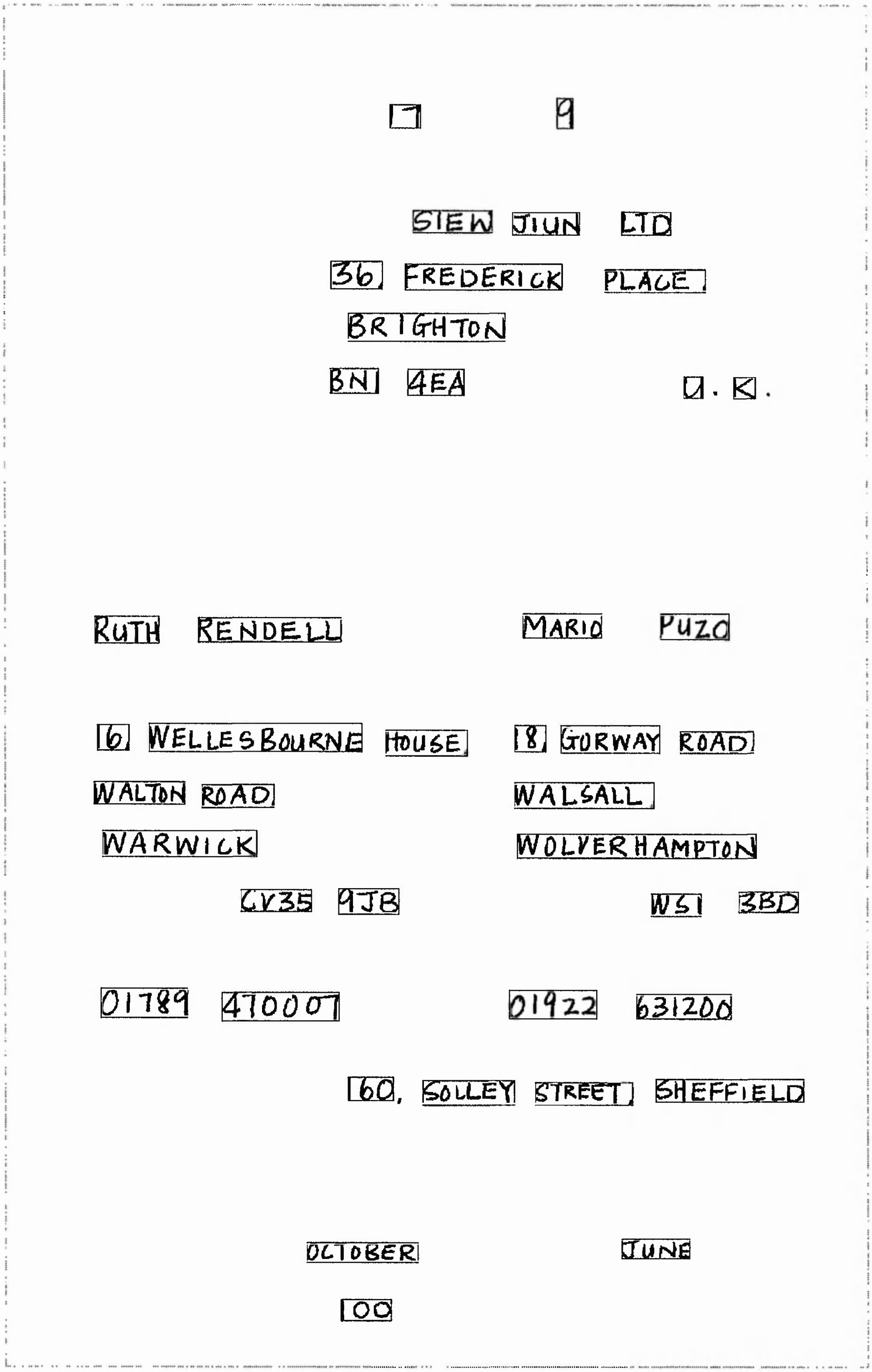
ROBSON

25 MAY 2000

Form03 (writer 1) extracted words



Form03 (writer 2) extracted words



Form03 (writer 3) extracted words

23 9

AXIOMATIC

23, LONG ROW AVE.,

NOTTINGHAM

NG1 4BW U.K.

KENNY LIEW

63 JALAN KEPONG

AMPANG

SELANGOR

43634

(603) 618 6023

TERENCE TAN

73-A, GLANMORGAN

STREET, SCOTLAND

S10 2FQ

(0181) 375 7076

14 ABBOT DR., NOTTINGHAM

JUNE AUGUST

750

Form03 (writer 4) extracted words

12 MAY

TNTU

14 CITY AV.

LENTON ABBEY

NG9 2TL ENGLAND

NIMA JAMSHID

UNIVERSITY ROAD 10 COLLEGE ROAD

BELFAST BANGOR

NORTHERN IRELAND WALES

BT7 1NN LL57 2DG

028 - 90245133 01248 351151

20 CHARLES AVE.

JUN FEB

100

Form03 (writer 5) extracted words

25 JULY

NICEDAY LTD
45 GEORGE HUDSON STREET
YORK
YO1 6ZG GB

ARNOLD HAMLET BARRIE BETHALL

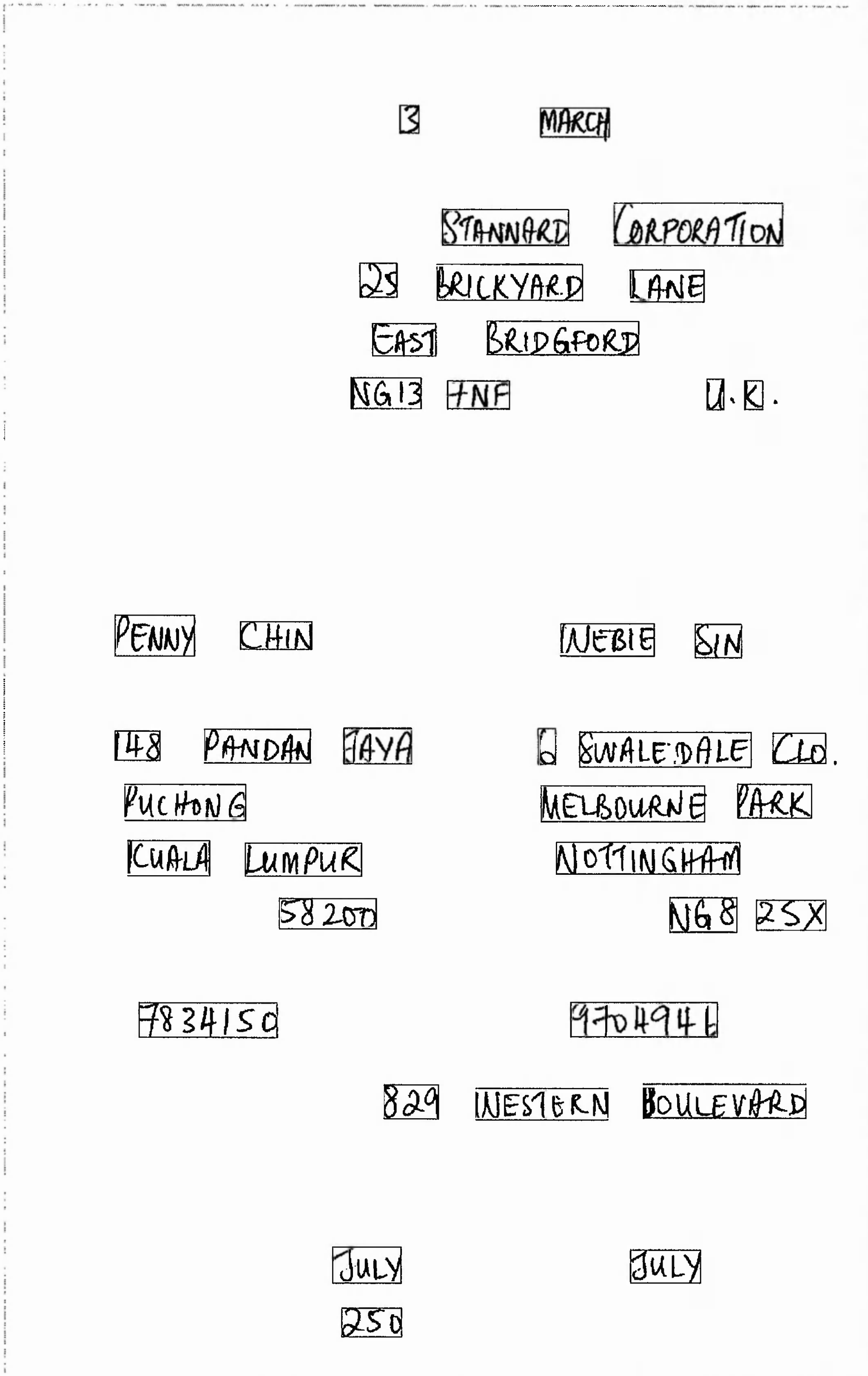
7 BOWLING HILL 21 CONWAY STREET
CHIPPING SODBURY BIRKENHEAD
BRISTOL MERSEYSIDE
BS37 6JX L41 4FD

01454 863333 0151 798888

42 TRUMPET STREET

APRIL JUNE
100

Form03 (writer 6) extracted words



Form03 (writer 7) extracted words

23 JUNE

TRAVAIL GROUP
4, NOTTINGHAM ROAD,
LONG EATON, NOTTINGHAM
NG10 1HP UK.

TONY SHARPE
26, WILTON TERRACE
CRESCENT, NEW BASFORD
LONDON, UK.
GN4 3GT

JON RIDING
46, TAMAN SKYLINE
BATU EMPAT,
JALAN KLANG LAMA,
55100

022 -9613130

03 -7810295.

25, LEWIS CLOSE, STANNS

JULY

JUNE

30

Form03 (writer 8) extracted words

28 MAY

IT TECHNOLOGY

522 JALAN TIGA EMPAT

JINJANG NORTH KUALA LUMPUR

52000 MALAYSIA

HILTON KONG MARHILL STEVEN

28 IVY CLOSE, 1, CARLTON SQ.

GEDLING ASHWOOD

NOTTINGHAM OXFORD

NG4 1DA OX2 1AA

0115 - 8481128 0717 3412186

32 ABBOTSFORD DRIVE, ST ANNS.

JUNE DECEMBER

40

Form03 (writer 9) extracted words

03 JUNE

ALL ADVANTAGE

330, BLOCK P/B

TAMAN CHERAS

56100 SABAH

YONG YEN SAN CHOW YUAN KAI

54, JALAN HANG 14-4A, JALAN 2 12A

LORONG ZIFFA TAMAN BARU

76503 20355

03-9119111 05-3787634

194, ABBOTSFORD DRIVE

MAY JULY

100.00

Form03 (writer 10) extracted words

15 JULY

MICROSOFT

2 HIGH STREET

CHELTENHAM

GL52 3AY U.K

DAVID BECKHAM EMMA WOODHOUSE

25 LONDON STREET 31 JEWRY STREET

BEDFORD LONDON

MK42 9AP EC3N 2EY

01234 363222 020 7320 1000

2 RENFEW STREET GLASGOW

MARCH AUGUST

50

APPENDIX F – Cue Word Linked Results

Name
CHONG
CHEE
CHAN
CHONG
CHEE
CHAN
WOON
Woon
Hock
Hock
HON
Hon
Address
STREET
HAMPDEN
203
WALK
NOTTINGHAM
LORNE
AVNS
0
/
St
Postcode
3BW
/
NG
4FX
3
NG
Country
K
U
KINGDOM
UNITED
Telephone
9127098
0115
9110
911
(0115)
127098
9
)
(0115
Month
Pound/£
05
25
98
6
05
25
Fax/Email
UK
AC
HON@NTU
Hock
Signature
Date
00
5
9
00
5
9
Value
2505
/
Description
CARD
SOUND
Parcel Number
66335255479/
PA
Office of Posting
NOTTINGHAM
Service Used
12
and
10
/
9
Bad Word
WOON
HON Hock
Not Tagged
✓
✓

Name
JIUN
SIEW
JIUN
SIEW
YAP
JUNE
YAP
JUNE
Address
JAYA,
CHERAS
PANDAN
7/1
PANDAN
JALAN
1,
NORTH
JINJANG
ROAD
TEMPLE
1522
Postcode
LUMPUR
KUALA
52000
LUMPUR
KUALA
55100
Country
MALAYSIA
MALAYSIA
Telephone
9118507
0115
9112277
0115
9118507
0115
Month
Pound/£
88
998
88
48
00
998
Fax/Email
com
sjyap@hotmail
Signature
Date
1998
FEB
10
1998
JAN
14
Value
88
218
8
00
100
1
00
680
1
Description
BOOKS
DISCMAN
CAMERA
DIGITAL
Parcel Number
18823
3279145
PA
Office of Posting
JINJANG
Service Used
EUR048
Bad Word
JUNE YAP
Not Tagged
✓
✓

Name
WONI
LO
MAY
INC
CHEUNG
JACKY
HOE
FOK
WONG
FOK HOE
WONG
Address
KOWLOON
TOWER
CRYSTAL
B
BLOCK
,
9A
28
LUMPUR
KUALA
17/6
JALAN
AMPANG
287,
Postcode
TWZ
F3A
58100
Country
FONG
HONG
MALAYSIA
Telephone
7552525
(013)
4832
576
)
(028
7552525
Month
Pound/£
1430.35
25
2
✓
18
70
1340 35
Fax/Email
3321
776
(018)
Signature
Date
2000
MAY
8
25/10/99
Value
1430.35
12
Description
FOOD
DRY
Parcel Number
543217689 0151
PA
Office of Posting
LUMPUR
KUALA
Service Used
ECONOMY
INTERNATIONAL
Bad Word
7
6 FOK HOE
1
Not Tagged
✓
0123456
✓

Name
EBADIAN
MANDANA
TONY
ALLEN
T.
Dr
EBADIAN
MANDANA
MANDANA
Address
UNIVERSITY
TRENT
STREET
BURTON
NOTTINGHAM
SQUARE
Bosly
29
BEESTON
Postcode
4BU
NG1
2TL
NG9
Country
K
U
NOTTINGHAM
Telephone
9221026
82680
7
81
9221026
Month
Pound/£
50.00
50
5
00
50
Fax/Email
ac.uk
ntu
tja@doc
ac.uk
mand@doc.ntu
Signature
M
Date
2000
JAN
30
2000
JAN
26
Value
25.00
2
Description
DOCUMENTS
Parcel Number
12
0001234561
PA
Office of Posting
NOTTINGHAM
Service Used
48
AND
24
Bad Word
Not Tagged
✓
✓

Name
JONES
ADAM
JONES
ADAM
DIGITAL
SCHOFIELD
BOBBY
OLIVER
Address
NOTTINGHAM
STREET
PEEL
18
ROAD
COMINGSWATH
CARLTON
76
NOTTINGHAM
Postcode
L
4G
1
NG
3SJ
NG4
Country
ENGLAND
UK
Telephone
8484742
0115
9525122
0115
Month
Pound/£
80
18
00
5
60
37
Fax/Email
ojschofield@yahoo.com
Signature
Date
Value
80
18
2
Description
HAT
BOWLER
Parcel Number
PA543197992151
Office of Posting
NOTTINGHAM
Service Used
Bad Word
0115 9525122
10/05/00
HOFIELD
OLIVER JOHNSON SC
OliverJohnsonSchofield
STANDARD
9/05/00
Not Tagged
✓
✓

Name
S
P
SHARON
THENG
S
P
WING
WONG
S
W
Address
ANNS
ST
NOTTINGHAM
DRIVE
ABBOTSFORD
195
NOTTINGHAM
BV
WESTERN
666
Postcode
1NB
NG3
3DX
NG5
Country
K
U
K
U
Telephone
Month
Pound/£
100/=
99
9
00
100
Fax/Email
Signature
Date
1999
1
3
1999
13
NOVEMBER
Value
1
✓
Description
100/=
LINED
JACKET
PILE
Parcel Number
69
123
PA2403435
Office of Posting
NOTTINGHAM
Service Used
STANDARD
Bad Word
2176
(0115) 848
DECEMBER
THENG
Not Tagged
PSTD3
✓

Name
WEE
LAM
BENG
EE
LAM,
WANG
CLOSE
NOTTINGHAM
25, LEWIS
WEE
TA
WANG
Address
VALE
HEMSHILL
STREET
GOOD
123
Postcode
234
ABC
1NR
NG3
Country
UK
UK
Telephone
8482599
0115
Month
Pound/£
1
900
15
80
95
2
10
12
✓
Fax/Email
Signature
Date
Value
100
1
1
Description
800
PC
SOFTWARES
Parcel Number
8910111
34567
PA
Office of Posting
Service Used
STANDARD
PARCELFORCE
Bad Word
8482599
199
0115-
10/12
WANG TA
8/12/99
NOTTINGHAM
Not Tagged
✓
23456
✓

Name
WING
GORDON
FORSYTH
GORDON
WING
WONG
WING
Address
EDINBURGH
TERRACE
INCHVIEW
24
NOTTINGHAM
DRIVE
ABBOTSFORD
195
Postcode
6TQ
EH7
1ND
NG3
Country
U.K
U.K
Telephone
011!
8484742
0115
Month
Pound/£
150.00
00
5
✓
50
7
95
99
Fax/Email
UK
AC
NTU
WING@DOC
Signature
Date
19/11/99
Value
150.00
1
Description
SET
PLAYSTATION
Parcel Number
PA003971130344
Office of Posting
NOTTINGHAM
Service Used
STANDARD
Bad Word
5 8484742
25/11/99
WONG
Not Tagged
✓
✓

Name
YONG
CHONG
CHEE
CHAN
CHONG
CHEE
CHAN
SAN
SAN
YEN
YEN
YONG
YONG
Address
STREET
HAMPDEN
203
DRIVE,
ABBOTSFORD
ANNS
ST
194
Postcode
3BW
1
NG
1ND
NG 3
NOTTINGHAM
Country
BRITAIN
GREAT
BRITAIN
GREAT
Telephone
8482599
0115
9119110
0115
8482599
0115
Month
Pound/£
34
103
97
7
34
103
Fax/Email
Signature
Date
00
5
5
99
9
25
Value
103.34
1
Description
PLAYER
VIDEO
Parcel Number
1234567890012
PA
Office of Posting
NOTTINGHAM
Service Used
48
and
24
Bad Word
YEN SAN
Not Tagged
✓
✓

Name
STEVE
LIM
STEVE
SHAOMIN
ZHANG
SHAOMIN
Address
StREET
Burton
Postcode
HOLE STREET
NOTTINGHAM
NG7 7AY
/
NOTTINGHAM
NG1 4BU
Country
UK
Telephone
9418419
15
01
8484742
0115
Month
Pound/£
00
100
0
0
50
00
100
Fax/Email
ac.uk
zhang@ntu
shaomin
Signature
Date
1999
14
JUNE
Value
00
100
/
Description
SET
GOLF
Parcel Number
000245123442
PA
Office of Posting
LEEDS
Service Used
STANDARD
Bad Word
-8484742
23, 2000
0115
MAY
SHAOMIN ZHANG
Shaomin zhang
NG1
UK
Not Tagged
✓
✓

Name
LOON
HUI
SU
KONG
AH
TAN
YEE
WEE
CHONG
WING
KOK
LIM
Address
OC
R
TAIWAN
ROAD
WELLINGTON
153
KINGDOM
UNITED
MALAYSIA
NOTTINGHAM
AMNS
51
JOHOR
81000
KULAI
WALK
LORNE
10
RAYA
BUNGA
JALAN
26
MALAYSIA
LUMPUR
KUALA
CHERAS
ROAD
PASAR
4
9
1
Postcode
Country
Telephone
8893
919
6633525
607
2727
8
4
8
8
709
2
1
9
Month
Pound/£
Fax/Email
Signature
Date
00
5
24
1999
APRIL
24
1966
MAY
29
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
JAMES
SIMON
EASTWOOD
GREAME
GOH
PETER
STONES
KAREN
Address
3XD
NG8
NOTTINGHAM
ROAD,
KING
,
HOSPITAL
QUEEN
P
7Z
YK5
YORK
3KS
1
OX
GARDEN,
KENSINGTON
OXFORD
,
COLWICK
STREET
REGENT
1
7
AVENUE,
BURTON
,
1
8
5MQ
NG3
NOTTINGHAM
CARLTON
GROVE,
STOKE
55,
Postcode
Country
Telephone
2457431
0872
7723688
0455
8764332
0115
9451792
0115
Month
Pound/£
Fax/Email
Signature
Date
2000
MAY
8
1
2000
APRIL
8
1968
MARCH
8
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
BIZKIT
LIMP
SINGH
NARAIN
DR
INC
CARE
MEDI
YOON
KEE
WAI
Address
BRUNEI
AMPANG
HAJI
LORONG
287
MALAYSIA
MALAYSIA
LUMPUR
KUALA
KUANTAN
BARU
KAMPUNG
FOXTROT
DESA
TAMAN
27
STREET
BROWNING
A
36
MALAYSIA
MALACCA
BARU
PERSIARAN
14/8
JALAN
29
Postcode
Country
Telephone
4827
397
(03)
4321
765
(03)
7917
718
(03)
4818
323
(03)
Month
Pound/£
Fax/Email
Signature
Jimbizkit
Date
1997
NOV
30
1996
AUGUST
21
1966
JULY
21
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
SHERKAT
NASSER
CLARK
CLAIR
JERAM
IRIN
DEHMSHKI
NIMA
Address
5
27
NG7
BEESTON
AV
CENTRAL
CENTRE
MEDICAL
ABBEY
9
25D
NG9
NOTTINGHAM
2TL
NG7
CHILWELL
BEESTON
AV
CHARLES
28
AV
CENTRAL
80
2TL
NG7
NOTTINGHAM
,
BEESTON
RD
PEVERILE
15
Postcode
Country
Telephone
2625981
0804
865
250
9
0115
9826209
0115
9262225
0115
Month
Pound/£
Fax/Email
Signature
Date
2000
May
18
1999
MAY
8
1992
MAY
5
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
THONG
ELTON
CLARK
PETULA
DIGITAL
BOBBY
ALFIE SHOOTER
Address
0
6
4
ESS
ESSEX
CENTRE
RING
RUBBER
STREET
ALLEY
DEATH
72
SW1
NOTTINGHAM
LONDON
GEDLING
ROAD
DOWNTOWN
18
ROAD
COUNTY
97
RU9
RR8
BRISTOL
CLOSE
BLUEHEAD
37
Postcode
Country
Telephone
776434
0181
504278
01742
8484742
0115
9525122
0115
Month
Pound/£
Fax/Email
Signature
Date
25/5/00
2/5/79
21/2/71
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
SHEQUAT
DR
KAUR
K
DR
LONSDALE
DR
YIM
CHRISTINE
Address
3NB
NG6
NOTTINGHAM
ROAD
DERBY
HOSPITAL
MEDICAL
QUEENS
1DX
NG5
2FF
NG1
NOTTINGHAM
NOTTINGHAM
RD
RADFORD
ROAD
SHAKESPEARE
CENTRE
MEDICAL
RADFORD
CENTRE
MEDICAL
CITY
4RW
NE3
NEWBURY
ROAD
TURNPIKE
38
Postcode
Country
Telephone
948 6111
9486567
8482176
8471093
Month
Pound/£
Fax/Email
Signature
Date
2000
28,
MAY
2000
,
13
MAY
1975
SEPTEMBER
6
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
FO
,
MICHELLE
EVANS
PAUL
DR
SIM
BOON
WANG
LIN
BOH
TAN
Address
,
BOMBAY
GREEN
HYSON
ROAD,
AMBLESIDE
9
MALAYSIA
AUSTRALIA
PARK
SHERWOOD
MAPPERLY
55100
LUMPUR
KUALA
CHERAS,
ROAD
FELSTEAD
123,
SHELLEY
JALAN
1,
MALAYSIA
LUMPUR
KUALA
55100
,
4 1/2
BATU
SKYLINE
TAMAN
46
Postcode
Country
Telephone
,
1830
82
,
006
95
2
0
781
Month
Pound/£
Fax/Email
Signature
Date
SEPTEMBER 08
10
2000
AUGUST
2
DECEMBER 48
25
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name
PIAO
AH
TAN
SMITH
MORGAN
PLC
HANDSOME
,
KING
STEVEN
Address
SCOTLAND
INVERNESS
AVENUE
EASTFIELD
48
SCOTLAND
SCOTLAND
ABERDEEN
4AC
G12
PETERCULTER
GLASGOW
WALK
CORONATION
18
ROAD
RUTHERGLEN
OLD
12
NOTTINGHAM
4BR
NG2
GEDLING
LANE
STOKE
7,
Postcode
Country
Telephone
8112323
8114124
9114145
8481123
Month
Pound/£
Fax/Email
Signature
Date
2000
MAY
26
1992
OCTOBER
14
1972
JUNE
01
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged

Name

OWEN
MICHAEL
WHITE
JIMMY
HENDRIX
JIMI
ENGLAND
4NA
BR10
BIRMINGHAM
CITY
BIRMINGHAM
DAVIS
ANDREW

Address

K
U
4AB
NB1
LONDON
MARYLEBONE
ALLEY
SWAN
GREAT
10
UK
4HM
WO3
ENGLAND
'
WOLVERHAMPTON
1CA
NO9
BLAKENHALL
NORWICH
BARFORD
LANE
BOOTH
WILLIAM
83
ROAD
MARLINGFORD
23
STREET
COVENTRY
950

Postcode

Country

Telephone

0347
910
0506
884
0000
911

Month

Pound/£

Fax/Email

Signature

Michael Owen

Date

27/5/00
1999
JAN
19
1982
MAY
01

Value

Description

Parcel Number

Office of Posting

Service Used

Bad Word

Not Tagged

Name

GUEST
ROBSON
BRIAN
JOHN
ENGLAND
3BU
MA2
MANCHESTER
STRETFORD
1
EAS
ROAD
DAVYHULME
23
COLE
[redacted] JOE

Address

K
U
4BA
5H3
STANNINGTON
SHEFFIELD
STREET
YORK
23A.
'
4AA
NO8
NORWICH
UK
INP
MA4
CRINGLEFORD
MANCHESTER
LEIGH
ROAD
KESWICK
1
1
CRAFT
SANERSONS
18

Postcode

Country

Telephone

13
12
842
72
84
911

Month

Pound/£

Fax/Email

Signature

ROBSON

Date

'
MAY 2000
25
'
1999
AUGUST
22
1982
MAY
01

Value

Description

Parcel Number

Office of Posting

Service Used

Bad Word

Not Tagged

Name
IH
J
JUI
AN
I
WEE YEE
CHONG
HOLDINGS
SCAN
IMAGE
Address
STREET
BURTON
84
VANG
KI
KAR
BA
ABU
JOHOR
KULAI
JALAN
23
I
RAYA
BUNGA
JALAN
26
NOTTINGHAM
ST ANNS
WALK
LORNE
0
I
Postcode
3812
8
1000
8
4FX
NG3
Country
UK
Telephone
4528
6
3
7
33304
66
Month
JUNE
AUGUST
Pound/£
120
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
MAY
31

Name
MARIO
RENDELL
RUTH
LTD
JIUN
SIEW
Address
SHEFFIELD
STREET
SOLLEY
160
WOLVERHAMPTON
WARWICK
WALSALL
ROAD
WALTON
ROAD
GORWAY
18,
HOUSE
WELLESBOURNE
16,
IGHTON
BR
PLACE
FREDERICK
36
Postcode
3BD
WS1
9JB
CV35
4EA
BN1
Country
K
U
Telephone
631200
01922
470007
01789
Month
JUNE
OCTOBER
Pound/£
100
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
PUZO
9
17

Name
TAN
TERENCE
LIEW
KENNY
AXIOMATIC
Address
NOTTINGHAM
DR
ABBOT
14
SELANGOR
SCOTLAND
STREET
AMPANG
GLANMORGAN
L
A
73
KEPONG,
JALAN
,
63
NOTTINGHAM
AVE
ROW
LONG
23
Postcode
2FQ
S10
43634
4BG
NG1
Country
K
U
Telephone
7076
375
(0181)
6023
618
(603)
Month
AUGUST
JUNE
Pound/£
750
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
9
23

Name
JAMSHID
NIMA
TNTU
Address
AVE
CHARLES
20
WALES
IRELAND
NORTHERN
BANGOR
BELFAST
ROAD
COLLEGE
10
ROAD
UNIVERSITY
ABBEY
LENTON
AV
CITY
14
Postcode
2DG
LL57
INN
7
BT
2TL
NG9
Country
ENGLAND
Telephone
1
35115
01248
90245133
028
Month
FEB
JUN
Pound/£
100
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
MAY
12

Name
BETHALL
BARRIE
HAMLET
ARNOLD
LTD
NICEDAY
Address
STREET
TRUMPET
42
MERSEYSIDE
BRISTOL
BIRKENHEAD
SODBURY
CHIPPING
STREET
CONWAY
21
HILL
BOWLING
7
YORK
HUDSON STREET
GEORGE
45
Postcode
4FD
L41
6JX
BS37
6ZG
YO1
Country
GB
Telephone
798888
0151
01454 863333
Month
JUNE
APRIL
Pound/£
100
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
JULY
25

Name
SIN
WEBIE
CHIN
PENNY
CORPORATION
STANNARD
Address
BOULEVARD
WESTERN
829
NOTTINGHAM
LUMPUR
KUALA
PARK
MELBOURNE
PUCHONG
CLO
SWALEDALE
6
JAYA
PANDAN
148
BRIDGFORD
EAST
LANE
BRICKYARD
25
Postcode
7834150
2SX
NG8
58200
7NF
NG13
Country
K
U
Telephone
9704946
Month
JULY
JULY
Pound/£
250
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
MARCH
3

Name
RIDING
JON
SHARPE
TONY
GROUP
TRAVAIL
Address
ANNS
'
S
,
CLOSE
LEWIS
25,
LAMA,
KLANG
JALAN
EMPAT,
BATU
BASFORD
NEW
'
CRESCEN
SKYLINE
TAMAN
46,
TERRACE
WILTON
26,
NOTTINGHAM
EATON
LONG
ROAD,
NOTTINGHAM
,
4
Postcode
55100
3GT
GN4
1HP
NG10
Country
UK
Telephone
7810295
03
9613130
022
Month
'
JUNE
JULY
Pound/£
30
Fax/Email
Signature
Date
Value
Description
Parcel Number
Office of Posting
Service Used
Bad Word
Not Tagged
'
UK
LONDON,
,
JUNE
23

Name

STEVEN
MARHILL
KONG
HILTON
TECHNOLOGY
IT

Address

ANNS
ST
DRIVE,
ABBOTSFORD
32
OXFORD
NOTTINGHAM
,
ASHWOOD
GEDLING
SQ
CARLTON
~~RD~~
1,
CLOSE
IVY
28
LUMPUR
KUALA
NORTH
,
JINJANG
EMPAT
TIGA
JALAN
1522

Postcode

1AA
OX2
1DA
NG4
52000

Country

MALAYSIA

Telephone

3412186
0717
8481128
0115

Month

DECEMBER
JUNE

Pound/£

40

Fax/Email

Signature

Date

Value

Description

Parcel Number

Office of Posting

Service Used

Bad Word

Not Tagged

,
MAY
28

Name

KAI
YUAN
CHOW
SAN
YEN
YONG
ADVANTAGE
ALL

Address

DRIVE
ABBOTSFORD
,
194
BARU
TAMAN
ZIFFA
LORONG
2/12A
JALAN
4A
14
HANG
JALAN
54,
CHERAS
TAMAN
P/B
BLOCK
320

Postcode

20355
76503
56100

Country

SABAH

Telephone

87634
37
05
9119111
03

Month

JULY
MAY

Pound/£

00
100

Fax/Email

Signature

Date

Value

Description

Parcel Number

Office of Posting

Service Used

Bad Word

Not Tagged

JUNE
23

Name

WOODHOUSE
EMMA
BECKHAM
DAVID
MICROSOFT

Address

GLASGOW
RENFEW STREET,
2
LONDON
BEDFORD
STREET
JEWRY
1
3
STREET
LONDON
25
ENHAM
CHELT
STREET
HIGH
2

Postcode

,
2EY
EC3N
9AP
MK42
3AY
GL52

Country

K
U

Telephone

1000
7320
020
363222
01234

Month

AUGUST
MARCH

Pound/£

50

Fax/Email

Signature

Date

Value

Description

Parcel Number

Office of Posting

Service Used

Bad Word

Not Tagged

JULY
5
1

APPENDIX G – List of Lexicons used in Experiment IV (section 4.7)

SERVICE USED (5 words)

ECONOMY, EURO, INTERNATIONAL, PARCELFORCE, STANDARD

MONTH (20 words)

JANUARY, JAN, FEBRUARY, FEB, MARCH, APRIL, MAY, JUNE, JUN, JULY, AUGUST, AUG, SEPTEMBER, SEPT, OCTOBER, OCT, NOVEMBER, NOV, DECEMBER, DEC

ITEM DESCRIPTION (21 words)

BOOKS, BOWLER, CAMERA, CARD, DIGITAL, DISCMAN, DOCUMENTS, DRY, FOOD, GOLF, HAT, JACKET, LINED, PC, PILE, PLAYER, PLAYSTATION, SET, SOFTWARES, SOUND, VIDEO

NAME (167 words)

ADAM, ADVANTAGE, AH, ALFIE, ALL, ALLEN, ANDREW, ARNOLD, AXIOMETIC, BARRIE, BECKHAM, BENG, BETHALL, BIZKIT, BOBBY, BOH, BOON, BRIAN, CARE, CHAN, CHEE, CHEUNG, CHIN, CHONG, CHOW, CHRISTINE, CLAIR, CLARK, COLE, CORPORATION, DAVID, DAVIS, DEHMSHKI, DIGITAL, DR, EASTWOOD, EBADIAN, EE, ELTON, EMMA,

EVANS, FORSYTH, GOH, GORDON, GREAME, GROUP, GUEST, HAMLET, HANDSOME, HENDRIX, HILTON, HOCK, HOE, HOLDINGS, HON, HUI, IMAGE, INC, IRIN, IT, JACKY, JAMES, JAMSHID, JERAM, JIH, JIM, JIMMY, JIUN, JOE, JOHN, JOHNSON, JON, JONES, JUI, JUNE, K, KAI, KAREN, KAUR, KENNY, KING, KOK, KONG, L, LAM, LIEW, LIM, LIMP, LIN, LO, LONGSDALE, LOON, LTD, MANDANA, MARHILL, MARIO, MAY, MEDI, MICHAEL, MICHELLE, MICROSOFT, MORGAN, NARAIN, NASSER, NICEDAY, NIMA, OLIVER, OWEN, PAUL, PENNY, PETER, PETULA, PIAO, PLC, PUZO, RENDELL, RIDING, ROBSON, RUTH, SAN, SCAN, SCHOFIELD, SHAOMIN, SHARON, SHARPE, SHEQUAT, SHERKAT, SHOOTER, SIEW, SIM, SIMON, SIN, SINGH, SMITH, STANNARD, STEVE, STEVEN, STONES, SU, TA, TAN, TECHNOLOGY, TEO, TERENCE, THENG, THONG, TNTU, TONY, TONY, TRAVAIL, WAI, WANG, WEBIE, WEE, WHITE, WING, WONG, WOODHOUSE, WOON, YAP, YEE, YEN, YIM, YONG, YOON, YUAN, ZHANG

TOWN (163 words)

ABERDEEN, ALCESTER, ALEXANDRIA, ASHFORD, BAKAR, BAKEWELL, BALDOCK, BARFORD, BARNES, BASFORD, BATH, BATHGATE, BAYTON, BEDFORD, BEESTON, BELFAST, BIRKENHEAD, BIRMINGHAM, BLACKPOOL, BLAKENHALL, BOLLINGTON, BOSTON, BOWLING, BRICKYARD, BRIDGFORD, BRIGHTON, BRISTOL, BRIXTON, BROMLEY, BUCKINGHAM, CALLINGTON, CAMBERWELL, CAMBRIDGE, CAMDEN, CANTERBURY, CARDIFF, CHATHAM, CHELSEA, CHELTENHAM, CHESTER, CHILWELL, CHIPPING, CORBY, COVENTRY, CRESCENT, CRINGLEFORD, CROYDON, DARTFORD, DAVYHULME, DERBY, DORCHESTER, DUBLIN, DUNDEE, EALING, EAST, EASTBOURNE, EATON, EDINBURGH, ENFIELD, ESSEX, EXETER, FOLKESTONE, FOXTROT, FREDERICK, GEORGE, GLANMORGAN, GLASGOW, GORWAY, GRANTHAM, GREEN, GREENWICH, HALIFAX, HAMPSTEAD, HARLOW, HASTINGS, HATFIELD, HERTFORD, HONITON, HUDSON, HULL, HYTHE, INVERNESS, IPSWICH, JERSEY, JEWRY, JINJANG, KINGSWEAR, KLANG, KLUANG, KUALA, LANCASTER, LEEDS,

LEICESTER, LEIGH, LINCOLN, LIVERPOOL, LONDON, LOUGHBOROUGH, LUMPUR, LUTON, MAIDENHEAD, MAIDSTONE, MANCHESTER, MANSFIELD, MARLOW, MARYLEBONE, MELBOURNE, MERSEYSIDE, MIDDLESBROUGH, NEW, NEWARK, NEWBURY, NEWCASTLE, NEWHAM, NEWPORT, NEWTOWN, NORTH, NORTHAMPTON, NORTHERN, NORWICH, NOTTINGHAM, OXFORD, PENARTH, PERTH, PETERBOROUGH, PETERCULTER, PLYMOUTH, PORLOCK, PORTSMOUTH, READING, RENFEW, ROYSTON, RYE, SALISBURY, SANDHURST, SANDWICH, SEAFORD, SELANGOR, SHEFFIELD, SLOUGH, SODBURY, SOLLEY, SOUTHAMPTON, STANNINGTON, STRETFORD, SUDBURY, SWALEDALE, SWAN, SWANSEA, SWINDON, VAUXHALL, WALSALL, WALTON, WARWICK, WATERFORD, WELLESBOURNE, WELLINGTON, WESTERN, WILTON, WIMBLEDON, WINDSOR, WOLVERHAMPTON, YORK

COUNTRY (230 words)

AFGHANISTAN, AFRICA, AFRICAN, ALBANIA, ALGERIA, AND, ANDORRA, ANGOLA, ANTIGUA, ARAB, ARABIA, ARGENTINA, ARMENIA, AUSTRALIA, AUSTRIA, AZERBAIJAN, BAHAMAS, BAHRAIN, BANGLADESH, BARBADOS, BARBUDA, BELGIUM, BELIZE, BELORUSSIA, BENIN, BHUTAN, BOLIVIA, BOSNIA, BOTSWANA, BRAZIL, BRITAIN, BRUNEI, BULGARIA, BURKINA, BURMA, BURUNDI, CAMBODIA, CAMEROON, CANADA, CAPE, CENTRAL, CHAD, CHILE, CHINA, CITY, COAST, COLOMBIA, COMOROS, CONGO, COSTA, CROATIA, CUBA, CYPRUS, CZECH, DENMARK, DJIBOUTI, DOMINICA, DOMINICAN, ECUADOR, EGYPT, EL, EMIRATES, ENGLAND, EQUATORIAL, ERITREA, ESTONIA, ETHIOPIA, FEDERATION, FIJI, FINLAND, FRANCE, GABON, GAMBIA, GB, GEORGIA, GERMANY, GHANA, GREAT, GREECE, GRENADA, GRENADINES, GUATEMALA, GUINEA, GUINEA-BISSAU, GUYANA, HAITI, HERZEGOVINA, HONDURAS, HONG, HUNGARY, ICELAND, INDIA, INDONESIA, IRAN, IRAQ, IRELAND, ISLANDS, ISRAEL, ITALY, IVORY, JAMAICA, JAPAN, JORDAN, KAZAKHSTAN, KENYA, KINGDOM, KIRIBATI, KITTS, KONG, KOREA, KUWAIT, KYRGYZSTAN,

LANKA, LAOS, LATVIA, LEBANON, LEONE, LESOTHO, LIBERIA, LIBYA, LIECHTENSTEIN, LITHUANIA, LUCIA, LUXEMBOURG, MACEDONIA, MADAGASCAR, MALAWI, MALAYSIA, MALDIVES, MALI, MALTA, MARINO, MARSHALL, MAURITANIA, MAURITIUS, MEXICO, MICRONESIA, MOLDOVIA, MONACO, MONGOLIA, MOROCCO, MOZAMBIQUE, MSIA, MYANMAR, NAMIBIA, NAURU, NEPAL, NETHERLANDS, NEVIS, NEW, NICARAGUA, NIGER, NIGERIA, NORTH, NORWAY, OMAN, PAKISTAN, PALAU, PANAMA, PAPUA, PARAGUAY, PERU, PHILIPPINES, POLAND, PORTUGAL, PRINCIPE, QATAR, REPUBLIC, RICA, ROC, ROMANIA, RUSSIAN, RWANDA, SABAH, SAINT, SALVADOR, SAMOA, SAN, SAO, SAUDI, SENEGAL, SEYCHELLES, SIERRA, SINGAPORE, SLOVAKIA, SLOVENIA, SOLOMON, SOMALIA, SOUTH, SPAIN, SRI, STATES, SUDAN, SURINAME, SWAZILAND, SWEDEN, SWITZERLAND, SYRIA, TAIWAN, TAJIKISTAN, TANZANIA, THAILAND, THE, TOBAGO, TOGO, TOME, TONGA, TRINIDAD, TUNISIA, TURKEY, TURKMENISTAN, TUVALU, UGANDA, UK, UKRAINE, UNITED, URUGUAY, UZBEKISTAN, VANUATU, VATICAN, VENEZUELA, VERDE, VIETNAM, VINCENT, YEMEN, YUGOSLAVIA, ZAIRE, ZAMBIA, ZEALAND, ZIMBABWE

ADDRESS (524 words)

ABBEY, ABBOT, ABBOTSFORD, ABC, ABERDEEN, ABU, AFGHANISTAN, AFRICA, AFRICAN, ALBANIA, ALCESTER, ALEXANDRIA, ALGERIA, ALLEY, AMBLESIDE, AMPANG, AND, ANDORRA, ANGOLA, ANNS, ANTIGUA, ARAB, ARABIA, ARGENTINA, ARMENIA, ASHFORD, ASHWOOD, AUSTRALIA, AUSTRIA, AV, AVE, AVENUE, AZERBAIJAN, BAHAMAS, BAHRAIN, BAKAR, BAKEWELL, BALDOCK, BANGLADESH, BANGOR, BARBADOS, BARBUDA, BARFORD, BARNES, BARU, BASFORD, BATH, BATHGATE, BATU, BAYTON, BEDFORD, BEESTON, BELFAST, BELGIUM, BELIZE, BELORUSSIA, BENIN, BHUTAN, BIRKENHEAD, BIRMINGHAM, BLACKPOOL, BLAKENHALL, BLOCK, BLUEHEAD, BOLIVIA, BOLLINGTON, BOMBAY, BOOTH, BOSLY, BOSNIA, BOSTON, BOTSWANA, BOULEVARD, BOWLING, BRAZIL, BRICKYARD, BRIDGFORD, BRIGHTON, BRISTOL, BRITAIN, BRIXTON, BROMLEY,

BROWNING, BRUNEI, BUCKINGHAM, BULGARIA, BUNGA, BURKINA, BURMA, BURTON, BURUNDI, BV, CALLINGTON, CAMBERWELL, CAMBODIA, CAMBRIDGE, CAMDEN, CAMEROON, CANADA, CANTERBURY, CAPE, CARDIFF, CARLTON, CENTRAL, CENTRE, CHAD, CHARLES, CHATHAM, CHELSEA, CHELTENHAM, CHERAS, CHESTER, CHILE , CHILWELL, CHINA, CHIPPING, CITY, CLO, CLOSE, COAST, COLLEGE, COLOMBIA, COLWICK, COMOROS, CONGO, CONINGSWATH, CONWAY, CORBY, CORONATION, COSTA, COUNTY, COVENTRY, CRAFT, CRESCENT, CRINGLEFORD, CROATIA, CROYDON, CRYSTAL, CUBA, CYPRUS, CZECH, DARTFORD, DAVYHULME, DEATH, DENMARK, DERBY, DESA, DJIBOUTI, DOMINICA, DOMINICAN, DORCHESTER, DOWNTOWN, DR, DRIVE, DUBLIN, DUNDEE, EALING, EAST, EASTBOURNE, EASTFIELD, EATON, ECUADOR, EDINBURGH, EGYPT , EL, EMIRATES, EMPAT, ENFIELD, ENGLAND, EQUATORIAL, ERITREA, ESSEX, ESTONIA, ETHIOPIA, EXETER, FEDERATION, FELSTEAD, FIJI, FINLAND, FOLKESTONE, FOXTROT, FRANCE, FREDERICK, GABON, GAMBIA, GARDEN, GB, GEDLING, GEORGE, GEORGIA, GERMANY, GHANA, GLANMORGAN, GLASGOW, GOOD, GORWAY, GRANTHAM, GREAT, GREECE, GREEN, GREENWICH, GRENADA, GRENADINES, GROVE, GUATEMALA, GUINEA, BISSAU, GUYANA, HAITI, HAJI, HALIFAX, HAMPDEN, HAMPSTEAD, HANG, HARLOW, HASTINGS, HATFIELD, HEMSHILL, HERTFORD, HERZEGOVINA, HIGH, HILL, HOLE, HONDURAS, HONG, HONITON, HOSPITAL, HOUSE, HUDSON, HULL, HUNGARY, HYSON, HYTHE, ICELAND, INCHVIEW, INDIA, INDONESIA, INVERNESS, IPSWICH, IRAN, IRAQ, IRELAND, ISLANDS, ISRAEL, ITALY, IVORY, IVY, JALAN, JAMAICA, JAPAN, JAYA, JERSEY, JEWRY, JINJANG, JOHOR, JORDAN, KAMPUNG, KAZAKHSTAN, KENSINGTON, KENYA, KEPONG, KESWICK, KING, KINGDOM, KINGSWEAR, KIRIBATI, KITTS, KLANG, KLUANG, KONG, KOREA, KOWLOON, KUALA, KUANTAN, KULAI, KUWAIT, KYRGYZSTAN, LAMA, LANCASTER, LANE, LANKA, LAOS, LATVIA, LEBANON, LEEDS, LEICESTER, LEIGH, LENTON, LEONE, LESOTHO, LEWIS, LIBERIA, LIBYA, LIECHTENSTEIN, LINCOLN, LITHUANIA, LIVERPOOL, LONDON, LONG, LORNE, LORONG,

LOUGHBOROUGH, LUCIA, LUMPUR, LUTON, LUXEMBOURG, MACEDONIA, MADAGASCAR, MAIDENHEAD, MAIDSTONE, MALACCA, MALAWI, MALAYSIA, MALDIVES, MALI, MALTA, MANCHESTER, MANSFIELD, MAPPERLY, MARINO, MARLOW, MARSHALL, MARUNGFORD, MARYLEBONE, MAURITANIA, MAURITIUS, MEDICAL, MELBOURNE, MERSEYSIDE, MEXICO, MICRONESIA, MIDDLESBROUGH, MOLDOVIA, MONACO, MONGOLIA, MOROCCO, MOZAMBIQUE, MYANMAR, NAMIBIA, NAURU, NEPAL, NETHERLANDS, NEVIS, NEW, NEWARK, NEWBURY, NEWCASTLE, NEWHAM, NEWPORT, NEWTOWN, NICARAGUA, NIGER, NIGERIA, NORTH, NORTHAMPTON, NORTHERN, NORWAY, NORWICH, NOTTINGHAM, OLD, OMAN, OXFORD, PAKISTAN, PALAU, PANAMA, PANDAN, PAPUA, PARAGUAY, PARK, PASAR, PEEL, PENARTH, PERSIARAN, PERTH, PERU, PETERBOROUGH, PETERCULTER, PEVERILE, PHILIPPINES, PLACE, PLYMOUTH, POLAND, PORLOCK, PORTSMOUTH, PORTUGAL, PRINCIPE, PUCHONG, QATAR, QUEEN, QUEENS, RADFORD, RAYA, RD, READING, REGENT, RENFEW, REPUBLIC, RICA, RING, ROAD, ROC, ROMANIA, ROW, ROYSTON, RUBBER, RUSSIAN, RUTHERGLEN, RWANDA, RYE, SABAH, SAINT, SALISBURY, SALVADOR, SAMOA, SAN, SANDHURST, SANDWICH, SANERSONS, SAO, SAUDI, SCOTLAND, SEAFORD, SELANGOR, SENEGAL, SEYCHELLES, SHAKESPEARE, SHEFFIELD, SHELLEY, SHERWOOD, SIERRA, SINGAPORE, SKYLINE, SLOUGH, SLOVAKIA, SLOVENIA, SODBURY, SOLLEY, SOLOMON, SOMALIA, SOUTH, SOUTHAMPTON, SPAIN, SQ, SQUARE, SRI, ST, STANNINGTON, STATES, STOKE, STREET, STRETFORD, SUDAN, SUDBURY, SURINAME, SWALEDALE, SWAN, SWANSEA, SWAZILAND, SWEDEN, SWINDON, SWITZERLAND, SYRIA, TAIWAN, TAJIKISTAN, TAMAN, TANZANIA, TEMPLE, TERRACE, THAILAND, THE, TIGA, TOBAGO, TOGO, TOME, TONGA, TOWER, TRENT, TRINIDAD, TRUMPET, TUNISIA, TURKEY, TURKMENISTAN, TURNPIKE, TUVALU, UGANDA, UK, UKRAINE, UNITED, UNIVERSITY, URUGUAY, UZBEKISTAN, VALE, VANUATU, VATICAN, VAUXHALL, VENEZUELA, VERDE, VIETNAM, VINCENT, WALES, WALK, WALSALL, WALSALL, WALTON, WARWICK, WATERFORD, WELLESBOURNE,

WELLINGTON, WESTERN, WILLIAM, WILTON, WIMBLEDON, WINDSOR, WOLVERHAMPTON, YEMEN, YORK, YUGOSLAVIA, ZAIRE, ZAMBIA, ZEALAND, ZIFFA, ZIMBABWE