- S. Kahan, T. Pavlidis & H. S. Baird.
  On the Recognition of Printed Characters of Any Font and Size.
  IEEE T-PAMI, Vol. 9, No.2, p. 274-288, March 1987.

- J. Mantas.
  An Overview of Character Recognition Methodologies.
  Pattern Recognition, Vol. 19, No 6, p. 425-430, 1986.

- S. Mori C.Y. Suen & K. Yamamoto.
  Historical Review of OCR research and Development.
  IEEE Proceedings, special issue on OCR, p. 1029-1057, July 1992.

- G. Nagy.
  At the Frontiers of OCR.
  IEEE Proceedings, special issue on OCR, p.1093-1100, July 1992.

- T. Pavlidis.
  Recognition of printed text under realistic conditions.
  Pattern Recognition Letters 14, p. 317-326, 1993.

- T. Pavlidis, J. Swartz & Y. P. Wang.
  Fundamentals of Bar Code Information Theory.
  IEEE Computer.

- R. Plamondon & G. Lorette.
  Automatic Signature Verification and Writer Identification - The State of the Art.
  Pattern Recognition, Vol. 22,No 2, p. 107-131, 1989.

- H. F. Schantz.
  The History of OCR.
  Recognition Technology Users Association, VT, 1982.

- J. Scurmann.
  Reading Machines.
  Proceedings IJCPR, Munich, p. 1031-1044, 1982.

- C.Y. Suen, M. Berthod & S. Mori.
  Automatic Recognition of Handprinted Characters - The State of the Art.
  IEEE Proceedings, Vol. 68, No. 4, p.469-487, April 1980.

- A.A. Verikas, M.I. Bachauskene, S.J. Vilunas & D.R. Skaigiris.
  Adaptive Character Recognition System.
  Pattern Recognition Letters 13, p. 207-212, 1992.

- T. Y. Young & K-S Fu.
  Handbook of Pattern Recognition and Image Processing.
  Academic Press, 1986.

Bibliography

- H.S. Baird & R. Fossey.
  A 100-Font Classifier.
  Proceedings ICDAR-91, Vol. 1, p. 332-340, 1991.

- M. Bokser.
  Omnidocument Technologies.
  IEEE Proceedings, special issue on OCR, p. 1066-1078, July 1992.

- R. Bradford & T. Nartker.
  Error Correlation in Contemporary OCR Systems.
  Proceedings ICDAR-91, Vol. 2, p. 516-524, 1991.

- J-P. Caillot.
  Review of OCR Techniques.
  NR-note, BILD/08/087.

- R. G. Casey & K. Y. Wong.
  Document-Analysis Systems and Techniques.
  Image Analysisi Applications, eds: R. Kasturi & M. Tivedi, p. 1-36.
  New York: Marcel Dekker, 1990.

- R. H. Davis & J. Lyall.
  Recognition of Handwritten Characters - a Review.
  Image and Vision Computing, Vol. 4, No. 4, p. 208-218, nov. 1986.

- S. Diehl & H. Eglowstein.
  Tame the Paper Tiger.
  Byte, p. 220-238, April 1991.

- G. Dimauro, S. Impedovo & G. Pirlo.
  From Character to Cursive Script Recognition: Future Trends in Scientific Research.
  Proceedinngs, IAPR'92, The Hague, Vol. 2, p. 516-519, 1992.

- R. C. Gonzalez & R. E. Woods.
  Digital Image Processing.
  Addison-Wesley, 1992.

- V. K. Govindan & A.P. Shivaprasad.
  Character Recognition - a Review.
  Pattern Recognition, Vol. 23, No &, P. 671-683, 1990.

- L. Haaland.
  Automatisk identifikasjon - den glemte muligheten.
  Teknisk Ukeblad, nr 39, 1992.

- S. Impedovo & L. Ottaviano & S. Occhinegro.
  Optical Character Recognition - A survey.
  Int. Journal of PRAI, Vol. 5, No 1& 2, p. 1-24, 1991.

Chapter 7

Summary

Character recognition techniques associate a symbolic identity with the image of character. Character recognition is commonly referred to as optical character recognition (OCR), as it deals with the recognition of optically processed characters. The modern version of OCR appeared in the middle of the 1940's with the development of the digital computers. OCR machines have been commercially available since the middle of the 1950's. Today OCR-systems are available both as hardware devices and software packages, and a few thousand systems are sold every week.

In a typical OCR systems input characters are digitized by an optical scanner. Each character is then located and segmented, and the resulting character image is fed intoa preprocessor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. After classification the identified characters are grouped to reconstruct the original symbol strings, and context may then be applied to detect and correct errors.

Optical character recognition has many different practical applications. The main areas where OCR has been of importance, are text entry (office automation), data entry (banking environment) and process automation (mail sorting).

The present state of the art in OCR has moved from primitive schemes for limited character sets, to the application of more sophisticated techniques for omnifont and handprint recognition. The main problems in OCR usually lie in the segmentation of degraded symbols which are joined or fragmented. Generally, the accuracy of an OCR system is directly dependent upon the quality of the input document. Three figures are used in ratings of OCR systems; correct classification rate, rejection rate and error rate. The performance should be rated from the systems error rate, as these errors go by undetected by the system and must be manually located for correction.

In spite of the great number of algorithms that have been developed for character recognition, the problem is not yet solved satisfactory, especially not in the cases when there are no strict limitations on the handwriting or quality of print. Up to now, no recognition algorithm may compete with man in quality. However, as the OCR machine is able to read much faster, it is still attractive.

In the future the area of recognition of constrained print is expected to decrease. Emphasis will then be on the recognition of unconstrained writing, like omnifont and handwriting. This is a challenge which requires improved recognition techniques. The potential for OCR algorithms seems to lie in the combination of different methods and the use of techniques that are able to utilize context to a much larger extent than current methodologies.

may be exchanged electronically or printed in a more computer readable form, for instance barcodes.

The applications for future OCR-systems lie in the recognition of documents where control over the production process is impossible. This may be material where the recipient is cut off from an electronic version and has no control of the production process or older material which at production time could not be generated electronically. This means that future OCR-systems intended for reading printed text must be omnifont.

Another important area for OCR is the recognition of manually produced documents. Within postal applications for instance, OCR must focus on reading of addresses on mail produced by people without access to computer technology. Already, it is not unusual for companies etc., with access to computer technology to mark mail with barcodes. The relative importance of handwritten text recognition is therefore expected to increase.

Chapter 6

The Future of OCR

Through the years, the methods of character recognition has improved from quite primitive schemes, suitable only for reading stylized printed numerals, to more complex and sophisticated techniques for the recognition of a great variety of typeset fonts and also handprinted characters. Below the future of OCR when it comes to both research and areas of applications, is briefly discussed.

6.1   Future improvements

New methods for character recognition are still expected to appear, as the computer technology develops and decreasing computational restrictions open up for new approaches. There might for instance be a potential in performing character recognition directly on grey level images. However, the greatest potential seems to lie within the exploitation of existing methods, by mixing methodologies and making more use of context.

Integration of segmentation and contextual analysis can improve recognition of joined and split characters. Also, higher level contextual analysis which look at the semantics of entire sentences may be useful. Generally there is a potential in using context to a larger extent than what is done today. In addition, combinations of multiple independent feature sets and classifiers, where the weakness of one method is compensated by the strength of another, may improve the recognition of individual characters.

The frontiers of research within character recognition have now moved towards the recognition of cursive script, that is handwritten connected or calligraphic characters. Promising techniques within this area, deal with the recognition of entire words instead of individual characters.

6.2   Future needs

Today optical character recognition is most successful for constrained material, that is documents produced under some control. However, in the future it seems that the need for constrained OCR will be decreasing. The reason for this is that control of the production process usually means that the document is produced from material already stored on a computer. Hence, if a computer readable version is already available, this means that data

There is usually a tradeoff between the different recognition rates. A low error rate may lead to a higher rejection rate and a lower recognition rate. Because of the time required to detect and correct OCR errors, the error rate is the most important when evaluating whether an OCR system is cost-effective or not. The rejection rate is less critical. An example from barcode reading may illustrate this. Here a rejection while reading a barcoded price tag will only lead to rescanning of the code or manual entry, while a misdecoded pricetag might result in the customer being charged for the wrong amount. In the barcode industry the error rates are therefore as low as one in a million labels, while a rejection rate of one in a hundred is acceptable.

In view of this, it is apparent that it is not sufficient to look solely on the recognition rates of a system. A correct recognition rate of 99%, might imply an error rate of 1%. In the case of text recognition on a printed page, which on average contains about 2000 characters, an error rate of 1% means 20 undetected errors per page. In postal applications for mail sorting, where an address contains about 50 characters, an error rate of 1% implies an error on every other piece of mail.

Feature extraction.

Even if a character is printed, scanned and segmented correctly, it may be incorrectly classified. This may happen if the character shapes are similar and the selected features are not sufficiently efficient in separating the different classes, or if the features are difficult to extract and has been computed incorrectly.

Classification.

Incorrect classification may also be due to poor design of the classifier. This may happen if the classifier has not been trained on a sufficient number of test samples representing all the possible forms of each character.

Grouping.

Finally, errors may be introduced by the postprocessing, when the isolated symbols are associated to reconstruct the original words as characters may be incorrectly grouped. These problems may occur if the text is skewed, in some cases of proportional spacing and for symbols having subscripts or superscripts.

As OCR devices employ a wide range of approaches to character recognition, all systems are not equally affected by the above types of complexities. The different systems have their particular strengths and weaknesses. In general, however, the problems of correct segmentation of isolated characters are the ones most difficult to overcome, and recognition of joined and split characters are usually the weakest link of an OCR-system.

5.4   OCR performance evaluation

No standardized test sets exist for character recognition, and as the performance of an OCR system is highly dependent on the quality of the input, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often given, and usually presented as the percentage of characters correctly classified. However, this does not say anything about the errors committed. Therefore in evaluation of OCR system, three different performance rates should be investigated:

- *Recognition rate.*
  The proportion of correctly classified characters.
- *Rejection rate.*
  The proportion of characters which the system were unable to recognize. Rejected characters can be flagged by the OCR-system, and are therefore easily retraceable for manual correction.
- *Error rate.*
  The proportion of characters erroneously classified. Misclassified characters go by undetected by the system, and manual inspection of the recognized text is necessary to detect and correct these errors.

Script.

All the methods for character recognition described in this document treat the problem of recognition of isolated characters. However, to humans it might be of more interest if it were possible to recognize entire words consisting of cursively joined characters. Script recognition deals with this problem of recognizing unconstrained handwritten characters which may be connected or cursive.

In signature verification and identification the objective is to establish the identity of the writer, irrespective of the handwritten contents. Identification establishes the identity of the writer by comparing specific attributes of the pattern describing the signature, with those of a list of writers stored in a reference database. When performing signature verification the claimed identity of the writer is known, and the signature pattern is matched against the signature stored in the database for this person. Some systems of this kind are starting to appear.

A more difficult problem is script recognition where the contents of the handwriting must be recognized. This is one of the really challenging areas of optical character recognition. The variations in shape of handwritten characters are infinite and depend on the writing habit, style, education, mood, social environment and other conditions of the writer. Even the best trained optical readers, the humans, make about 4% errors when reading in the lack of context. Recognition of characters written without any constraint is still quite remote. For the time being, recognition of handwritten script seems to belong only to on-line products where writing tablets are used to extract real-time information and features to aid recognition.

5.3 Typical errors in OCR

The accuracy of OCR systems is, in practice, directly dependent upon the quality of the input documents. The main difficulties encountered in different documents may be classified as follows:

- *Variations in shape*, due to serifs and style variations.
- *Deformations*, caused by broken characters, smudged characters and speckle.
- *Variations in spacing*, due to subscripts, superscripts, skew and variable spacing.
- *Mixture of text and graphics.*

These imperfections may affect and cause problems in different parts of the recognition process of an OCR-system, resulting in rejections or misclassifications.

Segmentation.

The majority of errors in OCR-systems are often due to problems in the scanning process and the following segmentation, resulting in joined or broken characters. Errors in the segmentation process may also result in confusion between text and graphics or between text and noise.

Constrained handwriting.

Recognition of constrained handwriting deals with the problem of unconnected normal handwritten characters. Optical readers with such capabilities are not yet very common, but do exist. However, these systems require well-written characters, and most of them can only recognize digits unless certain standards for the hand-printed characters are followed (see figure 10). The characters should be printed as large as possible to retain good resolution, and entered in specified boxes. The writer is also instructed to keep to certain models provided, avoiding gaps and extra loops. Commercially the term ICR (Intelligent Character Recognition) is often used for systems able to recognize handprinted characters.
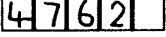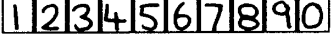
Figure 10 :   Instructions for OCR handwriting

5.2   OCR capabilities

The sophistication of the OCR system depends on the type and number of fonts recognized. Below a classification, by the order of difficulty, based on the OCR systems' capability to recognize different character sets, is presented.

Fixed font.

OCR machines of this category deals with the recognition of one specific typewritten font. Such fonts are OCR-A, OCR-B, Pica, Elite, etc. These fonts are characterized by fixed spacing between each character. The OCR-A and OCR-B are the American and European standard fonts specially designed for optical character recognition, where each character has a unique shape to avoid ambiguity with other characters similar in shape. Using these character sets, it is quite common for commercial OCR machines to achieve a recognition rate as high as 99.99% with a high reading speed.

The systems of the first OCR generation were fixed font machines, and the methods applied were usually based on template matching and correlation.

Multifont.

Multifont OCR machines recognize more than one font, as opposed to a fixed font system, which could only recognize symbols of one specific font. However, the fonts recognized by these machines are usually of the same type as those recognized by a fixed font system.

These machines appeared after the fixed-font machines. They were able to read up to about ten fonts. The limit in the number of fonts were due to the pattern recognition algorithm, template matching, which required that a library of bit map images of each character from each font was stored. The accuracy is quite good, even on degraded images, as long as the fonts in the library are selected with care.

Omnifont.

An omnifont OCR machine can recognize most nonstylized fonts without having to maintain huge databases of specific font information. Usually omnifont-technology is characterized by the use of feature extraction. The database of an omnifont system will contain a description of each symbol class instead of the symbols themselves. This gives flexibility in automatic recognition of a variety of fonts.

Although omnifont is the common term for these OCR systems, this should not be understood literally as the system being able to recognize all existing fonts. No OCR machine performs equally well, or even usably well, on all the fonts used by modern typesetters. A lot of current OCR-systems claim to be omnifont.

Chapter 5

Status of OCR

A wide variety of OCR systems are currently commercially available. In this chapter we take a look at the capabilities of OCR systems and the main problems encountered. We also discuss the problem of evaluating the performance of an OCR system.

5.1   OCR systems

OCR systems may be subdivided into two classes. The first class includes the special purpose machines dedicated to specific recognition problems. The second class covers the systems that are based on a PC and a low-cost scanner.

5.1.1   Dedicated hardware systems

The first recognition machines were all hardwired devices. Because this hardware was expensive, throughput rates had to be high to justify the cost, and parallelism was exploited. Today such systems are used in specific applications were speed is of high importance, for instance within the areas of mail-sorting and check-reading. The cost of these machines are still high, up to a million dollars, and they may recognize a wide range of fonts.

5.1.2   Software based PC versions

Advancements in the computer technology has made it possible to fully implement the recognition part of OCR in software packages which work on personal computers. Present PC systems are comparable to the large scaled computers of the early days, and as little additional equipment is required, the cost of such systems are low. However, there are some limitations in such OCR software, especially when it comes to speed and the kinds of character sets read.

Hand held scanners for reading do also exist. These are usually limited to the reading of numbers and just a few additional letters or symbols of fixed fonts. They often read a line at a time and transmits it to application programs.

Three commercial software products are dominant within the area of recognition of European languages. These are systems produced by Caera Corporation, Kurzweil and Calera Corporation, with prices in the range of $500 - $1000. The speed of these systems is about 40 characters per second

Automatic cartography.

Character recognition from maps presents special problems within character recognition. The symbols are intermixed with graphics, the text may be printed at different angles and the characters may be of several fonts or even handwritten.

Form readers.

Such systems are able to read specially designed forms. In such forms all the information irrelevant to the reading machine is printed in a colour "invisible" to the scanning device. Fields and boxes indicating where to enter the text is printed in this invisible colour. Characters should be entered in printed or hand written upper case letters or numerals in the specified boxes. Instructions are often printed on the form as how to write each character or numeral. The processing speed is dependent on the amount of data on each form, but may be about a few hundred forms per minute. Recognition rates are seldom given for such systems.

Signature verification and identification.

This is an application specially useful for the banking environment. Such a system establishes the identity of the writer without attempting to read the handwriting. The signature is simply considered as a pattern which is matched with signatures stored in a reference database.

4.3 Process automation.

Within this area of application the main concern is not to read what is printed, but rather to control some particular process. This is actually the technology of automatic address reading for mail sorting. Hence, the goal is to direct each letter into the appropriate bin regardless of whether each character was correctly recognized or not. The general approach is to read all the information available and use the postcode as a redundancy check.

The acceptance rate of these systems is obviously very dependent on the properties of the mail. This rate therefore varies with the percentage of handwritten mail. Although, the reject rate for mail sorting may be large, the missort rate is usually close to zero. The sorting speed is typically about 30.000 letters per hour.

4.4 Other applications.

The above areas are the ones in which OCR has been most successful and most widely used. However, many other areas of applications exist, and some of these are mentioned below.

Aid for blind.
In the early days, before the digital computers and the need for input of large amounts of data emerged, this was the imagined area of application for reading machines. Combined with a speech synthesis system such a reader would enable the blind to understand printed documents. However, a problem has been the high costs of reading machines, but this may be an increasing area as the costs of microelectronics fall.

Automatic number-plate readers.
A few systems for automatic reading of number plates of cars exist. As opposed to other applications of OCR, the input image is not a natural bilevel image, and must be captured by a very fast camera. This creates special problems and difficulties although the character set is limited and the syntax restricted.

Chapter 4

Applications of OCR

The last years have seen a widespread appearance of commercial optical character recognition products meeting the requirements of different users. In this chapter we treat some of the different areas of application for OCR. Three main application areas are commonly distinguished; data entry, text entry and process automation.

4.1   Data entry.

This area covers technologies for entering large amounts of restricted data. Initially such document reading machines were used for banking applications. The systems are characterized by reading only an extremely limited set of printed characters, usually numerals and a few special symbols. They are designed to read data like account numbers, customers identification, article numbers, amounts of money etc. The paper formats are constrained with a limited number of fixed lines to read per document.

Because of these restrictions, readers of this kind may have a very high throughput of up to 150.000 documents per hour. Single character error and reject rates are 0.0001% and 0.01% respectively. Also, due to the limited character set, these readers are usually remarkably tolerant to bad printing quality. These systems are specially designed for their applications and prices are therefore high.

4.2   Text entry.

The second branch of reading machines is that of page readers for text entry, mainly used in office automation. Here the restrictions on paper format and character set are exchanged for constraints concerning font and printing quality. The reading machines are used to enter large amounts of text, often in a word processing environment. These page readers are in strong competition with direct key-input and electronic exchange of data. This area of application is therefore of diminishing importance.

As the character set read by these machines is rather large, the performance is extremely dependent on the quality of the printing. However, under controlled conditions the single character error and reject rates are about 0.01% and 0.1% respectively. The reading speed is typically in the order of a few hundred characters per second

the word, by saying for instance that after a period there should usually be a capital letter. Also, for different languages the probabilities of two or more characters appearing together in a sequence can be computed and may be utilized to detect errors. For instance, in the English language the probability of a "k" appearing after an "h" in a word is zero, and if such a combination is detected an error is assumed.

Another approach is the use of dictionaries, which has proven to be the most efficient method for error detection and correction. Given a word, in which an error may be present, the word is looked up in the dictionary. If the word is not in the dictionary, an error has been detected, and may be corrected by changing the word into the most similar word. Probabilities obtained from the classification, may help to identify the character which has been erroneously classified. If the word is present in the dictionary, this does unfortunately not prove that no error occurred. An error may have transformed the word from one legal word to another, and such errors are undetectable by this procedure. The disadvantage of the dictionary methods is that the searches and comparisons implied are time-consuming.

3.1.5.2   Structural Methods.

Within the area of structural recognition, syntactic methods are among the most prevalent approaches. Other techniques exist, but they are less general and will not be treated here.

Syntactic methods.

Measures of similarity based on relationships between structural components may be formulated by using grammatical concepts. The idea is that each class has its own grammar defining the composition of the character.A grammar may be represented as strings or trees, and the structural components extracted from an unknown character is matched against the grammars of each class. Suppose that we have two different character classes which can be generated by the two grammars G_1 and G_2, respectively. Given an unknown character, we say that it is more similar to the first class if it may be generated by the grammar G_1, but not by G_2.

3.1.6   Post processing.

Grouping.

The result of plain symbol recognition on a document, is a set of individual symbols. However, these symbols in themselves do usually not contain enough information. Instead we would like to associate the individual symbols that belong to the same string with each other, making up words and numbers. The process of performing this association of symbols into strings, is commonly referred to as grouping. The grouping of the symbols into strings is based on the symbols' location in the document. Symbols that are found to be sufficiently close are grouped together.

For fonts with fixed pitch the process of grouping is fairly easy as the position of each character is known. For typeset characters the distance between characters are variable. However, the distance between words are usually significantly larger than the distance between characters, and grouping is therefore still possible. The real problems occur for handwritten characters or when the text is skewed.

Error-detection and correction.

Up until the grouping each character has been treated separately, and the context in which each character appears has usually not been exploited. However, in advanced optical text-recognition problems, a system consisting only of single-character recognition will not be sufficient. Even the best recognition systems will not give 100% percent correct identification of all characters, but some of these errors may be detected or even corrected by the use of context.

There are two main approaches, where the first utilizes the possibility of sequences of characters appearing together. This may be done by the use of rules defining the syntax of

Matching.

Matching covers the groups of techniques based on similarity measures where the distance between the feature vector, describing the extracted character and the description of each class is calculated. Different measures may be used, but the common is the Euclidean distance. This minimum distance classifier works well when the classes are well separated, that is when the distance between the means is large compared to the spread of each class.

When the entire character is used as input to the classification, and no features are extracted (template-matching), a correlation approach is used. Here the distance between the character image and prototype images representing each character class is computed.

Optimum statistical classifiers.

In statistical classification a probabilistic approach to recognition is applied. The idea is to use a classification scheme that is optimal in the sense that, on average, its use gives the lowest probability of making classification errors.

A classifier that minimizes the total average loss is called the Bayes' classifier. Given an unknown symbol described by its feature vector, the probability that the symbol belongs to class c is computed for all classes $c=1...N$. The symbol is then assigned the class which gives the maximum probability.

For this scheme to be optimal, the probability density functions of the symbols of each class must be known, along with the probability of occurrence of each class. The latter is usually solved by assuming that all classes are equally probable. The density function is usually assumed to be normally distributed, and the closer this assumption is to reality, the closer the Bayes' classifier comes to optimal behaviour.

The minimum distance classifier described above is specified completely by the mean vector of each class, and the Bayes classifier for Gaussian classes is specified completely by the mean vector and covariance matrix of each class. These parameters specifying the classifiers are obtained through a training process. During this process, training patterns of each class is used to compute these parameters and descriptions of each class are obtained.

Neural networks.

Recently, the use of neural networks to recognize characters (and other types of patterns) has resurfaced. Considering a back-propagation network, this network is composed of several layers of interconnected elements. A feature vector enters the network at the input layer. Each element of the layer computes a weighted sum of its input and transforms it into an output by a nonlinear function. During training the weights at each connection are adjusted until a desired output is obtained. A problem of neural networks in OCR may be their limited predictability and generality, while an advantage is their adaptive nature.

the character like unintended gaps in the contour. In table 2 these features are therefore characterized as having a low tolerance to noise. However, they are tolerant to noise affecting the inside of the character and to distortions.

Structural analysis.

During structural analysis, features that describe the geometric and topological structures of a symbol are extracted. By these features one attempts to describe the physical make-up of the character, and some of the commonly used features are strokes, bays, end-points, intersections between lines and loops. Compared to other techniques the structural analysis gives features with high tolerance to noise and style variations. However, the features are only moderately tolerant to rotation and translation. Unfortunately, the extraction of these features is not trivial, and to some extent still an area of research.
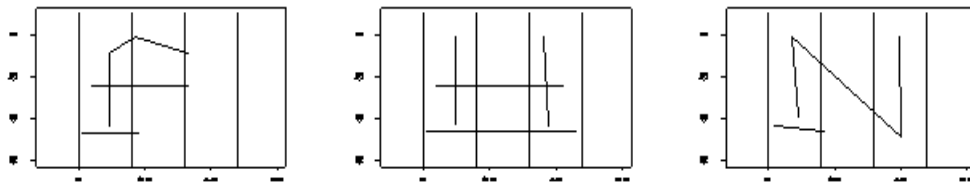
Figure 9 :   Strokes extracted from the capital letters F, H and N.

3.1.5   Classification.

The classification is the process of identifying each character and assigning to it the correct character class. In the following sections two different approaches for classification in character recognition are discussed. First decision-theoretic recognition is treated. These methods are used when the description of the character can be numerically represented in a feature vector.

We may also have pattern characteristics derived from the physical structure of the character which are not as easily quantified. In these cases the relationship between the characteristics may be of importance when deciding on class membership. For instance, if we know that a character consists of one vertical and one horizontal stroke, it may be either an "L" or a "T", and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed.

3.1.5.1   Decision-theoretic methods.

The principal approaches to decision-theoretic recognition are minimum distance classifiers, statistical classifiers and neural networks. Each of these classification techniques are briefly described below.

Characteristic loci.

For each point in the background of the character, vertical and horizontal vectors are generated. The number of times the line segments describing the character are intersected by these vectors are used as features.
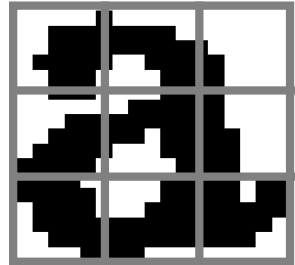

Figure 7 :   Zoning

Transformations and series expansions.

These techniques help to reduce the dimensionality of the feature vector and the extracted features can be made invariant to global deformations like translation and rotation. The transformations used may be Fourier, Walsh, Haar, Hadamard, Karhunen-Loeve, Hough, principal axis transform etc.
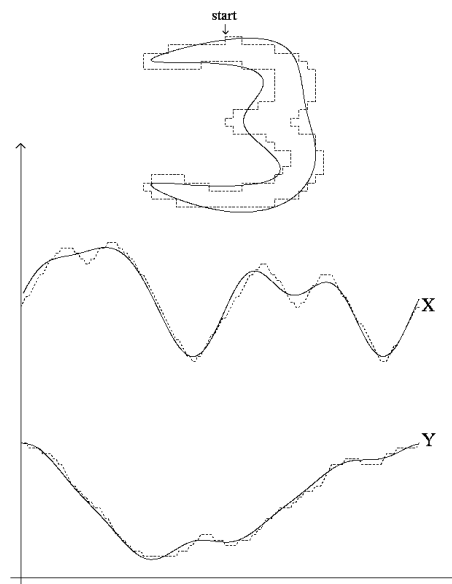

Figure 8 :   Elliptical Fourier descriptors

Many of these transformations are based on the curve describing the contour of the characters. This means that these features are very sensitive to noise affecting the contour of

3.1.4.1 Template-matching and correlation techniques.

These techniques are different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern.

The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines. However, this technique is sensitive to noise and style variations and has no way of handling rotated characters.

3.1.4.2 Feature based techniques.

In these methods, significant measurements are calculated and extracted from a character and compared to descriptions of the character classes obtained during a training phase. The description that matches most closely provides recognition. The features are given as numbers in a feature vector, and this feature vector is used to represent the symbol.

Distribution of points.

This category covers techniques that extracts features based on the statistical distribution of points. These features are usually tolerant to distortions and style variations. Some of the typical techniques within this area are listed below.

Zoning.
The rectangle circumscribing the character is divided into several overlapping, or non-overlapping, regions and the densities of black points within these regions are computed and used as features.

Moments.
The moments of black points about a chosen centre, for example the centre of gravity, or a chosen coordinate system, are used as features.

Crossings and distances.
In the crossing technique features are found from the number of times the character shape is crossed by vectors along certain directions. This technique is often used by commercial systems because it can be performed at high speed and requires low complexity.

When using the distance technique certain lengths along the vectors crossing the character shape are measured. For instance the length of the vectors within the boundary of the character.

n-tuples.
The relative joint occurrence of black and white points (foreground and background) in certain specified orderings, are used as features.

- **Robustness.**

  1) *Noise.*
  Sensitivity to disconnected line segments, bumps, gaps, filled loops etc.

  2) *Distortions.*
  Sensitivity to local variations like rounded corners, improper protrusions, dilations and shrinkage.

  3) *Style variation.*
  Sensitivity to variation in style like the use of different shapes to represent the same character or the use of serifs, slants etc.

  4) *Translation.*
  Sensitivity to movement of the whole character or its components.

  5) *Rotation.*
  Sensitivity to change in orientation of the characters.

- **Practical use.**

  1) *Speed of recognition.*

  2) *Complexity of implementation.*

  3) *Independence.*
  The need of supplementary techniques.

Each of the techniques evaluated in table2 are described in the next sections.

Feature extraction technique	Robustness 1 2 3 4 5					Practical use 1 2 3		
Template matching	◐	◐	○	○	○	○	●	○
Transformations	○	●	●	●	●	○	○	◐
Distribution of points: Zoning	○	◐	○	○	◐	●	●	○
Moments	◐	◐	○	●	●	○	◐	○
n-tuple	◐	○	◐	○	◐	●	●	◐
Characteristic loci	○	●	●	●	◐	●	●	○
Crossings	○	●	●	●	◐	●	●	○
Structural features	○	●	●	●	◐	●	○	●

● High or easy   ◐ Medium   ○ Low or difficult

Table 2 :   Evaluation of feature extraction techniques.

3.1.3   Preprocessing

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a preprocessor to smooth the digitized characters.

The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. The most common techniques for smoothing, moves a window across the binary image of the character, applying certain rules to the contents of the window.

In addition to smoothing, preprocessing usually includes normalization. The normalization is applied to obtain characters of uniform size, slant and rotation. To be able to correct for rotation, the angle of rotation must be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew. However, to find the rotation angle of a single symbol is not possible until after the symbol has been recognized.

Figure 6 :   Normalization and smoothing of a symbol.

3.1.4   Feature extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

The different groups of features may be evaluated according to their sensitivity to noise and deformation and the ease of implementation and use. The results of such a comparison are shown in table 1. The criteria used in this evaluation are the following:

such methods usually depend upon a multilevel scanning of the document which requires more memory and computational capacity. Therefore such techniques are seldom used in connection with OCR systems, although they result in better images.

3.1.2   Location and segmentation.

Segmentation is a process that determines the constituents of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition.

Applied to text, segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component, that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The main problems in segmentation may be divided into four groups:

- *Extraction of touching and fragmented characters.*
  Such distortions may lead to several joint characters being interpreted as one single character, or that a piece of a character is believed to be an entire symbol. Joints will occur if the document is a dark photocopy or if it is scanned at a low threshold. Also joints are common if the fonts are serifed. The characters may be split if the document stems from a light photocopy or is scanned at a high threshold.

- *Distinguishing noise from text.*
  Dots and accents may be mistaken for noise, and vice versa.

- *Mistaking graphics or geometry for text.*
  This leads to nontext being sent to recognition.

- *Mistaking text for graphics or geometry.*
  In this case the text will not be passed to the recognition stage. This often happens if characters are connected to graphics.
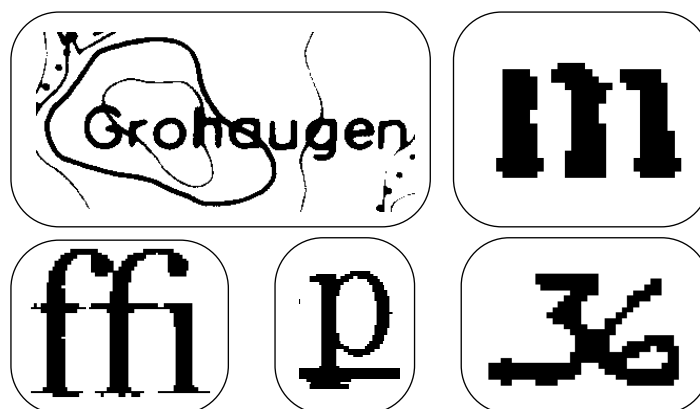

Figure 5 :   Degraded symbols.

The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text. In the next sections these steps and some of the methods involved are described in more detail.

3.1.1   Optical scanning.

Through the scanning process a digital image of the original document is captured. In OCR optical scanners are used, which generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing OCR, it is common practice to convert the multilevel image into a bilevel image of black and white. Often this process, known as thresholding, is performed on the scanner to save memory space and computational effort.

The thresholding process is important as the results of the following recognition is totally dependent of the quality of the bilevel image. Still, the thresholding performed on the scanner is usually very simple. A fixed threshold is used, where gray-levels below this threshold is said to be black and levels above are said to be white. For a high-contrast document with uniform background, a prechosen fixed threshold can be sufficient. However, a lot of documents encountered in practice have a rather large range in contrast. In these cases more sophisticated methods for thresholding are required to obtain a good result.
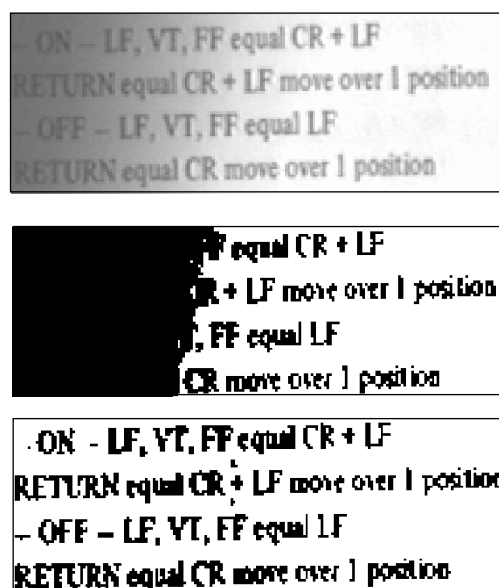

Figure 4 :   Problems in thresholding: *Top*: Original greylevel image, *Middle*: Image thresholded with global method, *Bottom*: Image thresholded with an adaptive method.

The best methods for thresholding are usually those which are able to vary the threshold over the document adapting to the local properties as contrast and brightness. However,

Chapter 3

Methods of OCR

The main principle in automatic recognition of patterns, is first to teach the machine which classes of patterns that may occur and what they look like. In OCR the patterns are letters, numbers and some special symbols like commas, question marks etc., while the different classes correspond to the different characters. The teaching of the machine is performed by showing the machine examples of characters of all the different classes. Based on these examples the machine builds a prototype or a description of each class of characters. Then, during recognition, the unknown characters are compared to the previously obtained descriptions, and assigned the class that gives the best match.

In most commercial systems for character recognition, the training process has been performed in advance. Some systems do however, include facilities for training in the case of inclusion of new classes of characters.

3.1   Components of an OCR system

A typical OCR system consists of several components. In figure 3 a common setup is illustrated. The first step in the process is to digitize the analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process. The extracted symbols may then be preprocessed, eliminating noise, to facilitate the extraction of features in the next step.
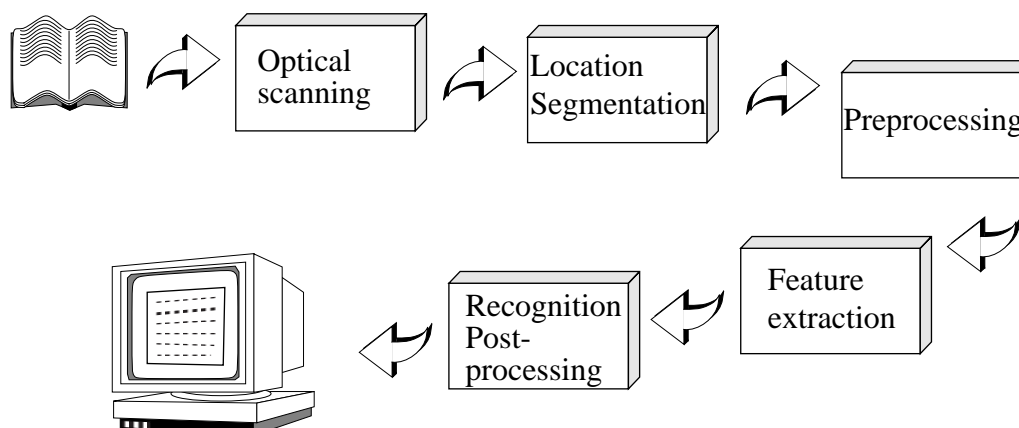
Figure 3 :   Components of an OCR-system

2.5   Third generation OCR.

For the third generation of OCR systems, appearing in the middle of the 1970's, the challenge was documents of poor quality and large printed and hand-written character sets. Low cost and high performance were also important objectives, which were helped by the dramatic advances in hardware technology.

Although more sophisticated OCR-machines started to appear at the market simple OCR devices were still very useful. In the period before the personal computers and laser printers started to dominate the area of text production, typing was a special niche for OCR. The uniform print spacing and small number of fonts made simply designed OCR devices very useful. Rough drafts could be created on ordinary typewriters and fed into the computer through an OCR device for final editing. In this way word processors, which were an expensive resource at this time, could support several people and the costs for equipment could be cut.

2.6   OCR today.

Although, OCR machines became commercially available already in the 1950's, only a few thousand systems had been sold world wide up to 1986. The main reason for this was the cost of the systems. However, as hardware was getting cheaper, and OCR systems started to become available as software packages, the sale increased considerably. Today a few thousand is the number of systems sold every week, and the cost of an omnifont OCR has dropped with a factor of ten every other year for the last 6 years.

1870	The very first attempts
1940	The modern version of OCR.
1950	The first OCR machines appear
1960 - 1965	First generation OCR
1965 - 1975	Second generation OCR
1975 - 1985	Third generation OCR
1986 ->	OCR to the people

Table 1 :   A short OCR chronology.

2.3 First generation OCR.

The commercial OCR systems appearing in the period from 1960 to 1965 may be called the first generation of OCR. This generation of OCR machines were mainly characterized by the constrained letter shapes read. The symbols were specially designed for machine reading, and the first ones did not even look very natural. With time multifont machines started to appear, which could read up to ten different fonts. The number of fonts were limited by the pattern recognition method applied, template matching, which compares the character image with a library of prototype images for each character of each font.

2.4 Second generation OCR.

The reading machines of the second generation appeared in the middle of the 1960's and early 1970's. These systems were able to recognize regular machine printed characters and also had hand-printed character recognition capabilities. When hand-printed characters were considered, the character set was constrained to numerals and a few letters and symbols.

The first and famous system of this kind was the IBM 1287, which was exhibited at the World Fair in New York in 1965. Also, in this period Toshiba developed the first automatic letter sorting machine for postal code numbers and Hitachi made the first OCR machine for high performance and low cost.

In this period significant work was done in the area of standardization. In 1966, a thorough study of OCR requirements was completed and an American standard OCR character set was defined; OCR-A. This font was highly stylized and designed to facilitate optical recognition, although still readable to humans. A European font was also designed, OCR-B, which had more natural fonts than the American standard. Some attempts were made to merge the two fonts into one standard, but instead machines being able to read both standards appeared.

```
A  B  C  D  E  F  G  H  I  J  K  L

M  N  O  P  Q  R  S  T  U  V  W  X

Y  Z  1  2  3  4  5  6  7  8  9  0


A  B  C  D  E  F  G  H  I  J  K  L

M  N  O  P  Q  R  S  T  U  V  W  X

Y  Z  1  2  3  4  5  6  7  8  9  0
```

Figure 2 : OCR-A (top), OCR-B (bottom).

Chapter 2

The History of OCR

Methodically, character recognition is a subset of the pattern recognition area. However, it was character recognition that gave the incentives for making pattern recognition and image analysis matured fields of science.

2.1   The very first attempts.

To replicate the human functions by machines, making the machine able to perform tasks like reading, is an ancient dream. The origins of character recognition can actually be found back in 1870. This was the year that C.R.Carey of Boston Massachusetts invented the retina scanner which was an image transmission system using a mosaic of photocells. Two decades later the Polish P. Nipkow invented the sequential scanner which was a major breakthrough both for modern television and reading machines.

During the first decades of the 19'th century several attempts were made to develop devices to aid the blind through experiments with OCR. However, the modern version of OCR did not appear until the middle of the 1940's with the development of the digital computer. The motivation for development from then on, was the possible applications within the business world.

2.2   The start of OCR.

By 1950 the technological revolution was moving forward at a high speed, and electronic data processing was becoming an important field. Data entry was performed through punched cards and a cost-effective way of handling the increasing amount of data was needed. At the same time the technology for machine reading was becoming sufficiently mature for application, and by the middle of the 1950's OCR machines became commercially available.

The first true OCR reading machine was installed at Reader's Digest in 1954. This equipment was used to convert typewritten sales reports into punched cards for input to the computer.

Optical Character Recognition.

Optical character recognition is needed when the information should be readable both to humans and to a machine and alternative inputs can not be predefined. In comparison with the other techniques for automatic identification, optical character recognition is unique in that it does not require control of the process that produces the information.

1.2   Optical Character Recognition

Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents.
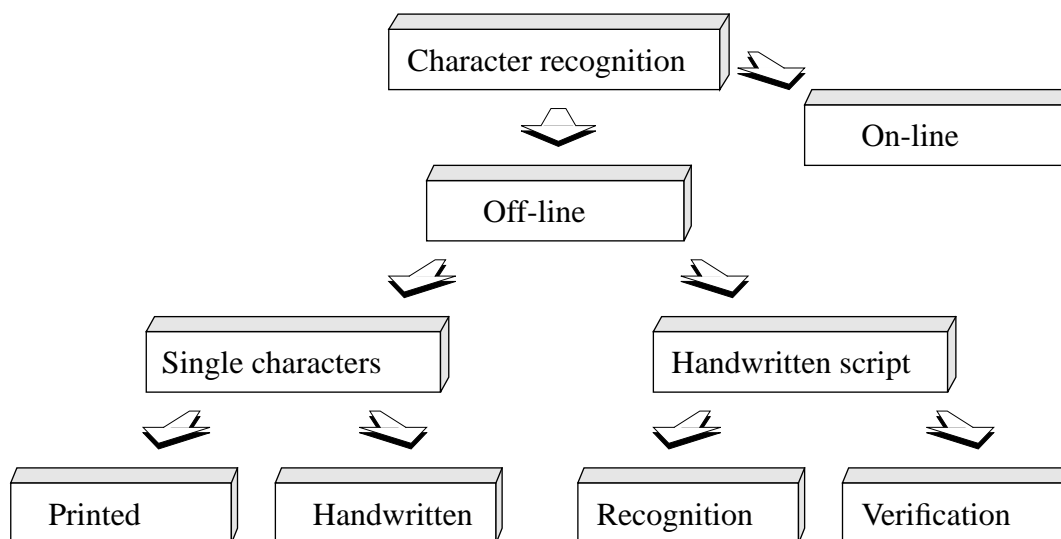

Figure 1 :   The different areas of character recognition.

The more constrained the input is, the better will the performance of the OCR system be. However, when it comes to totally unconstrained handwriting, OCR machines are still a long way from reading as well as humans. However, the computer reads fast and technical advances are continually bringing the technology closer to its ideal.

Magnetic stripe.

Information contained in magnetic stripes are widely used on credit cards etc. Quite a large amount of information can be stored on the magnetic stripe, but specially designed readers are required and the information can not be read by humans.

Bar code.

The bar code consists of several dark and light lines representing a binary code for an eleven-digit number, ten of which identify the particular product. The bar code is read optically, when the product moves over a glass window, by a focused laser beam of weak intensity which is swept across the glass window in a specially designed scanning pattern. The reflected light is measured and analysed by a computer. Due to early standardization, bar codes are today widely used and constitute about 60 % of the total market for automatic identification.

The bar code represents a unique number that identifies the product, and a price look-up (PLU) is necessary to retrieve information about price etc. The binary pattern representing the barcode takes up much space considering the small amount of information it actually contains. Also, the barcodes are not readable to humans. Hence, they are only useful when the information can be printed elsewhere in a human readable form or when human readability is not required. Laser-scanning of barcodes is therefore only in a few cases an alternative to optical character recognition.

Magnetic ink.

Printing in magnetic ink is mainly used within bank applications. The characters are written in ink that contains finely ground magnetic material and they are written in stylized fonts which are specifically designed for the application. Before the characters are read, the ink is exposed to a magnetic field. This process accentuates each character and helps simplify the detection. The characters are read by interpreting the waveform obtained when scanning the characters horizontally. Each character is designed to have its own specific waveform. Although designed for machine reading, the characters are still readable to humans. However, the reading is dependent on the characters being printed with magnetic ink.

Optical Mark Reading.

This technology is used to register location of marks. It may be used to read forms where the information is given by marking predefined alternatives. Such forms will also be readable to humans and this approach may be efficient when the input is constrained and may be predefined and there is a fixed number of alternatives.

Chapter 1

Introduction to OCR

Optical character recognition belongs to the family of techniques performing automatic identification. Below we discuss these different techniques and define OCR's position among them.

1.1 Automatic Identification

The traditional way of entering data into a computer is through the keyboard. However, this is not always the best nor the most efficient solution. In many cases automatic identification may be an alternative. Various technologies for automatic identification exist, and they cover needs for different areas of application. Below a brief overview of the different technologies and their applications is given.

Speech recognition.

In systems for speech recognition, spoken input from a predefined library of words are recognized. Such systems should be speaker-independent and may be used for instance for reservations or ordering of goods by telephone. Another kind of such systems are those used to recognize the speaker, rather than the words, for identification.

Radio frequency.

This kind of identification is used for instance in connection with toll roads for identification of cars. Special equipment on the car emits the information. The identification is efficient, but special equipment is needed both to send and to read the information. The information is also inaccessible to humans.

Vision systems.

By the use of a TV-camera objects may be identified by their shape or size. This approach may for instance be used in automatons for recirculation of bottles. The type of bottle must be recognized, as the amount reimbursed for a bottle depends on it's type.

Introduction

Machine replication of human functions, like reading, is an ancient dream. However, over the last five decades, machine reading has grown from a dream to reality. Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to compete with human reading capabilities.

In the first chapter of this documents, we discuss different technologies for automatic identification and establish OCR's position among these techniques. The next chapter gives a brief overview of the historical background and development of character recognition. We also present the different steps, from a methodical point of view, which have been employed in OCR. An account of the wide area of applications for OCR is given in chapter 4, and the following chapter looks into the current status of OCR. In the final chapter we discuss the future of OCR.

Table of Contents

OCR

Optical Character Recognition

Line Eikvil

December 1993