Chapter 2
Optical Character Recognition Systems

Abstract Optical character recognition (OCR) is process of classification of optical patterns contained in a digital image. The character recognition is achieved through segmentation, feature extraction and classification. This chapter presents the basic ideas of OCR needed for a better understanding of the book. The chapter starts with a brief background and history of OCR systems. Then the different techniques of OCR systems such as optical scanning, location segmentation, pre-processing, segmentation, representation, feature extraction, training and recognition and post-processing. The different applications of OCR systems are highlighted next followed by the current status of the OCR systems. Finally, the future of the OCR systems is presented.

Keywords OCR · Segmentation · Feature extraction · Classification

2.1 Introduction

Optical character recognition (OCR) [2, 7] is process of classification of optical patterns contained in a digital image corresponding to alphanumeric or other characters. The character recognition is achieved through important steps of segmentation, feature extraction and classification [12]. OCR has gained increasing attention in both academic research and in industry. In this chapter we have collected together the basic ideas of OCR needed for a better understanding of the book. It has been man's ancient dream to develop machines which replicate human functions. One such replication of human functions is reading of documents encompassing different forms of text. Over the last few decades machine reading has grown from dream to reality through the development of sophisticated and robust Optical character recognition (OCR) systems. OCR technology enables us to convert different types of documents such as scanned paper documents, pdf files or images captured by a digital camera into editable and searchable data. OCR systems have become one of the most successful applications of technology

A. Chaudhuri et al., *Optical Character Recognition Systems for Different Languages with Soft Computing*, Studies in Fuzziness and Soft Computing 352,
DOI 10.1007/978-3-319-50252-6_2

in pattern recognition and artificial intelligence fields. Though many commercial systems for performing OCR exist for a wide variety of applications, the available machines are still not able to compete with human reading capabilities with desired accuracy levels.

OCR belongs to the family of machine recognition techniques performing automatic identification. Automatic identification is the process where the recognition system identifies objects automatically, collects data about them and enters data directly into computer systems i.e. without human involvement. The external data is captured through analysis of images, sounds or videos. To capture data, a transducer is employed that converts the actual image or sound into a digital file. The file is then stored and at a later time it can be analyzed by the computer.

We start with a review of currently available automatic identification techniques and define OCR's position among them. The traditional way of entering data in a computer is through the keyboard. However, this is not always the best or the most efficient way. The automatic identification may serve as an alternative in many cases. There exist various techniques for automatic identification which cover the needs for different application areas. Some notable technologies and their applications worth mentioning apart from OCR are speech recognition, radio frequency, vision systems, magnetic stripe, bar code, magnetic ink and optical mark reading. These technologies have been actively used in past decades [10]. Here we introduce these technologies briefly from application point of view. Interested readers can refer [3, 5, 6, 8, 9, 11] for more elaborate discussion on these technologies:

(a) In speech recognition systems spoken input from a predefined library of words are recognized. Such systems are speaker independent and are generally used for reservations or telephonic ordering of goods. Another kind of such systems are those which are used to recognize speaker rather than words for identification.
(b) The radio frequency identification is often used 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 required both to send and to read the information. The information is inaccessible to humans.
(c) The vision systems are enforced through the usage TV camera where the objects are identified by their shape or size. This approach is generally used in automatons for recirculation of bottles. The type of bottle must be recognized first as the amount reimbursed for a bottle depends on its type.
(d) The 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 cannot be read by humans.
(e) 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 glass window by a focused laser beam of weak intensity which is swept across glass

window in a specially designed scanning pattern. The reflected light is measured and analysed by computer. Due to early standardization bar codes are today widely used and constitute a major share of the total market for automatic identification. The bar code represents a unique number that identifies the product and a price lookup is necessary to retrieve information about the price. The binary pattern representing the barcode takes up much space considering the small amount of information it actually contains. The barcodes are not readable to humans. Hence, they are only useful when the information is printed elsewhere in a human readable form or when human readability is not required.

(f) The printing in magnetic ink is mainly used within bank applications. The characters are written in ink that contains finely ground magnetic material. 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, reading is dependent on characters being printed with magnetic ink.

(g) The optical mark reading technology is used to register location of marks. It is used to read forms where the information is given by marking predefined alternatives. Such forms are also readable to humans. This approach is efficient when input is constrained. It is predefined with fixed number of alternatives.

OCR tries to address several issues of abovementioned techniques for automatic identification. They are required when the information is readable both to humans and machines. OCR systems have carved a niche place in pattern recognition. Their uniqueness lies in the fact that it does not require control of process that produces information. OCR deals with the problem of recognizing optically processed characters. Optical recognition is performed offline after the writing or printing has been completed whereas the online recognition is achieved where 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 input documents. The more constrained the input is, better is the performance of OCR system. But when it comes to totally unconstrained handwriting performance of OCR machines is still questionable. The Fig. 2.1 shows the schematic representation of different areas of character recognition.

This chapter is organized as follows. A brief historical background of OCR systems is Sect. 2.2. In Sect. 2.3 a discussion of different techniques of OCR is highlighted. This is followed by the applications of OCR systems in Sect. 2.4. In Sect. 2.5 we present the status of the OCR systems. Finally in Sect. 2.6 the future of OCR systems is given.

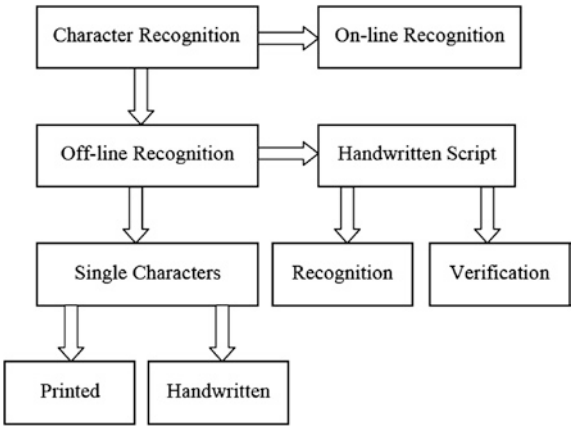

Fig. 2.1 The different areas of character recognition

2.2 Optical Character Recognition Systems: Background and History

Character recognition is a subset of pattern recognition area. Several concepts and techniques in OCR are borrowed from pattern recognition and image processing. However, it was character recognition that provided impetus for making pattern recognition and image analysis as matured fields of science and engineering.

Writing which has been the most natural mode of collecting, storing and transmitting information through the centuries now serves not only for communication among humans but also serves for communication of humans and machines. The intensive research effort in the field of OCR was not only because of its challenge on simulation of human reading but also because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents. To replicate human functions by machines and making the machine perform common tasks like reading is an ancient dream. The origins of character recognition dates back to 1870 when C.R. Carey of Boston Massachusetts [3, 7, 8] invented retina scanner which was an image transmission system using a mosaic of photocells. Early versions needed to be trained with images of each character and worked on one font at a time.

The history of OCR can be traced as early as 1900, when the Russian scientist Tyuring attempted to develop an aid for the visually handicapped [6]. The first character recognizers appeared in the middle of the 1940s with the development of digital computers [3]. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon a small set of well distinguished handwritten text or symbols. Machine printed OCR systems in this period generally used template matching in which an image is compared to a library of images. For handwritten text, low level image processing techniques have been used on the binary image to extract feature vectors which are

then fed to statistical classifiers. Successful but constrained algorithms have been implemented mostly for Latin characters and numerals. However, some studies on Japanese, Chinese, Hebrew, Indian, Cyrillic, Greek, and Arabic characters and numerals in both machine-printed and handwritten cases were also initiated [3].

Two decades later Nipkow [11] invented sequential scanner which was a major breakthrough both for modern television and reading machines. During the first few decades of 19th century several attempts were made to develop devices to aid the blind through experiments with OCR [8]. However, the modern version of OCR did not appear till the mid 1940s when digital computer came into force. The motivation for development of OCR systems started from then onwards when people thought for possible business and commercial applications.

By 1950 the technological revolution [3, 6] was moving forward at high speed and electronic data processing was becoming an upcoming and important field. The commercial character recognizers available in 1950s where electronic tablets captured the x–y coordinate data of pen tip movement was first introduced. This innovation enabled the researchers to work on the online handwriting recognition problem [3]. The data entry was performed through punched cards. A cost effective way of handling the increasing amount of data was then required. At the same time the technology for machine reading was becoming sufficiently mature for application. By mid 1950s OCR machines became commercially available [8]. The first OCR reading machine [9] was installed at Reader's Digest in 1954. This equipment was used to convert typewritten sales reports into punched cards for input into the computer.

The commercial OCR systems appearing from 1960 to 1965 were often referred to as first generation OCR [3, 8]. The OCR machines of this generation were mainly characterized by constrained letter shapes. The symbols were specially designed for machine reading. When multi-font machines started to appear, they could read up to several different fonts. The number of fonts were limited by pattern recognition method applied and template matching which compares the character image with library of prototype images for each character of each font.

In mid 1960s and early 1970s the reading machines of second generation appeared [3, 8]. 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 as well as few letters and symbols. The first and famous system of this kind was IBM 1287 in 1965. During this period Toshiba developed the first automatic letter sorting machine for postal code numbers. Hitachi also 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 as OCR–A shown in Fig. 2.2. This font was highly stylized and designed to facilitate optical recognition although still readable to humans. A European font was also designed as OCR–B shown in Fig. 2.3 which had more natural fonts than American standard. Attempts were made to merge two fonts in one standard through machines which could read both standards.

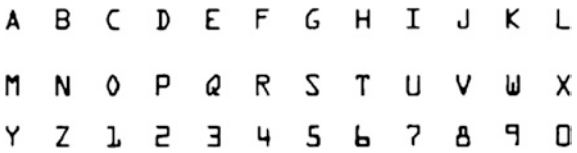

Fig. 2.2 OCR–A font

Fig. 2.3 OCR–B font

In mid 1970s the third generation of OCR systems appeared [3, 8]. The challenge was handing documents of poor quality and large printed and hand written character sets. The low cost and high performance objectives were achieved through dramatic advances in hardware technology. This resulted in the growth of sophisticated OCR machines for users. In the period before 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 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 at reduced equipment costs.

Although OCR machines became commercially available already in the 1950s, only few thousand systems were sold till 1986 [6]. The main reason for this was the cost of systems. However, as hardware prices went down and OCR systems started to become available as software packages, the sale increased considerably. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components. Today few millions of OCR systems are sold every week. The cost of omnifont OCR has dropped with a factor of ten every other year for the last few decades.

The studies up until 1980 suffered from the lack of powerful computer hardware and data acquisition devices. With the explosion of information technology, the previously developed methodologies found a very fertile environment for rapid growth in many application areas, as well as OCR system development [3]. The structural approaches were initiated in many systems in addition to the statistical methods [3]. The OCR research was focused basically on the shape recognition techniques without using any semantic information. This led to an upper limit in the recognition rate which was not sufficient in many practical applications.

The real progress on OCR systems achieved during 1990s using the new development tools and methodologies which are empowered by the continuously

growing information technologies. In the early 1990s, image processing and pattern recognition techniques were efficiently combined with artificial intelligence methodologies. Researchers developed complex OCR algorithms, which receive high-resolution input data and require extensive number crunching in the implementation phase. Nowadays, in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras, and electronic tablets, we have efficient, modern use of methodologies such as artificial neural networks (ANNs), hidden Markov models (HMMs), fuzzy set reasoning and natural language processing. The recent systems for the machine printed offline [1, 3] and limited vocabulary, user dependent online handwritten characters [1] are quite satisfactory for restricted applications. However, still a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading especially for unconstrained online and offline handwriting.

2.3 Techniques of Optical Character Recognition Systems

The main concept in automatic recognition of patterns is first to teach the machine which class of patterns that may occur and what they look like [3, 4]. In OCR patterns are letters, numbers and some special symbols like commas, question marks as well as different characters. The teaching of machine is performed by showing machine examples of characters of all different classes. Based on these examples the machine builds prototype or description of each class of characters. During recognition the unknown characters are compared to previously obtained descriptions and assigned to class that gives the best match. In most commercial systems for character recognition training process is performed in advance. Some systems however include facilities for training in the case of inclusion of new classes of characters.

A typical OCR system consists of several components as shown in Fig. 2.4 [3, 7]. The first step is to digitize analog document using an optical scanner. When regions containing text are located each symbol is extracted through segmentation process. The extracted symbols are pre-processed, eliminating noise to facilitate feature extraction. The identity of each symbol is found by comparing extracted features with descriptions of symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct words and numbers of the original text. These steps are briefly presented here. Interested readers can refer [11] for more elaborate discussion of OCR system components.

2.3.1 Optical Scanning

The first component in OCR is optical scanning. Through scanning process digital image of original document is captured. In OCR optical scanners are used which

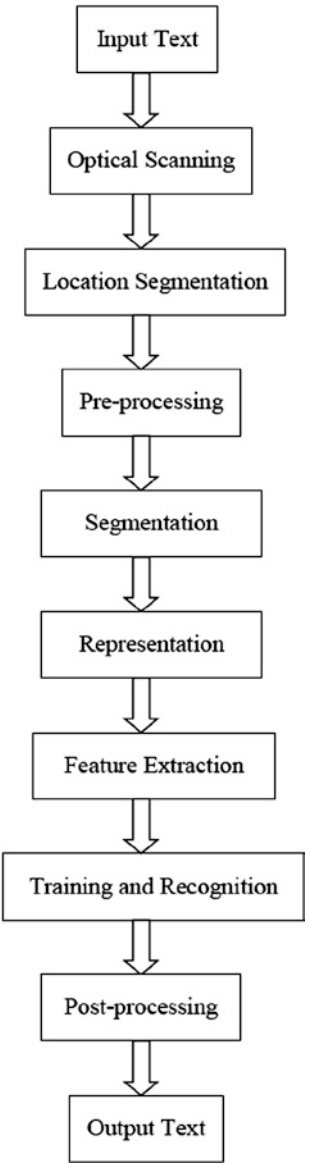

Fig. 2.4 The components of an OCR system

consist of transport mechanism and sensing device that converts light intensity into grey levels. Printed documents consist of black print on white background. When performing OCR multilevel image is converted into bi-level black and white image. This process known as thresholding is performed on scanner to save memory space and computational effort. The thresholding process is important as the results of recognition are totally dependent on quality of bi-level image. A fixed

threshold is used where gray levels below this threshold are black and levels above are white. For high contrast document with uniform background a pre-chosen fixed threshold can be sufficient. However, documents encountered in practice have rather large range. In these cases more sophisticated methods for thresholding are required to obtain good results. The best thresholding methods vary threshold adapting to local properties of document such as contrast and brightness. However, such methods usually depend on multilevel scanning of document which requires more memory and computational capacity.

2.3.2 Location Segmentation

The next OCR component is location segmentation. Segmentation determines constituents of an image. It is necessary to locate regions of document which have printed data and are distinguished from figures and graphics. For example, when performing automatic mail sorting through envelopes address must be located and separated from other prints like stamps and company logos, prior to recognition. When applied to text, segmentation is isolation of characters or words. Most of OCR algorithms segment words into isolated characters which are recognized individually. Usually segmentation is performed by isolating each connected component. This technique is easy to implement but problems arise if characters touch or they are fragmented and consist of several parts. The main problems in segmentation are: (a) extraction of touching and fragmented characters (b) distinguishing noise from text (c) misinterpreting graphics and geometry with text and vice versa. For interested readers further details are available in [11].

2.3.3 Pre-processing

The third OCR component is pre-processing. The raw data depending on the data acquisition type is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. The image resulting from scanning process may contain certain amount of noise. Depending on the scanner resolution and the inherent thresholding, the characters may be smeared or broken. Some of these defects which may cause poor recognition rates and are eliminated through pre-processor by smoothing digitized characters. Smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in digitized characters while thinning reduces width of line. The most common technique for smoothing moves a window across binary image of character and applies certain rules to the contents of window. Pre-processing also includes normalization alongwith smoothing. The normalization is applied to obtain characters of uniform size, slant and rotation. The correct rotation is found through its angle. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew.

The pre-processing component thus aims to produce data that are easy for the OCR systems to operate accurately. It is an important activity to be performed before the actual data analysis. The main objectives of pre-processing can be pointed as [1, 3]: (a) noise reduction (b) normalization of the data and (c) compression in the amount of information to be retained. In rest of this subsection the aforementioned objectives of pre-processing objectives are discussed with the corresponding techniques.

(a) Noise reduction: The noise introduced by the optical scanning device or the writing instrument causes disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortion including local variations, rounding of corners, dilation and erosion is a potential problem. It is necessary to eliminate these imperfections prior to actual processing of the data. The noise reduction techniques can be categorized in three major groups [1, 3]: (i) filtering (ii) morphological operations and (iii) noise modeling.

   (i) Filtering aims to remove noise and diminish spurious points usually introduced by uneven writing surface and poor sampling rate of the data acquisition device. Various spatial and frequency domain filters have been designed for this purpose. The basic idea is to convolute a predefined mask with the image to assign a value to a pixel as a function of the gray values of its neighboring pixels. Several filters have been designed for smoothing, sharpening, thresholding, removing slightly textured or coloured background and contrast adjustment purposes [1, 3].
   (ii) The basic idea behind the morphological operations is to filter the character image replacing the convolution operation by the logical operations. Various morphological operations have been designed to connect the broken strokes, decompose the connected strokes, smooth the contours, prune wild points, thin the characters and extract the boundaries [1, 3]. The morphological operations can be successfully used to remove noise on the character images due to low quality of paper and ink as well as erratic hand movement.
   (iii) Noise can generally be removed by calibration techniques if it would have been possible to model it. However, noise modeling is not possible in most of the applications. There exists some available literature on noise modeling introduced by optical distortion such as speckle, skew and blur. It is also possible to assess the quality of the character images and remove the noise to a certain degree [1, 3].

(b) Normalization: The normalization methods aim to remove the variations of the writing and obtain standardized data. Some of the commonly used methods for normalization are [1, 3]: (i) skew normalization and baseline extraction (ii) slant normalization (iii) size normalization and (iv) contour smoothing.

   (i) Skew normalization and baseline extraction: Due to inaccuracies in the scanning process and writing style the writing may be slightly tilted or curved within the image. This can hurt the effectiveness of the algorithms and thus should be detected and corrected. Additionally, some

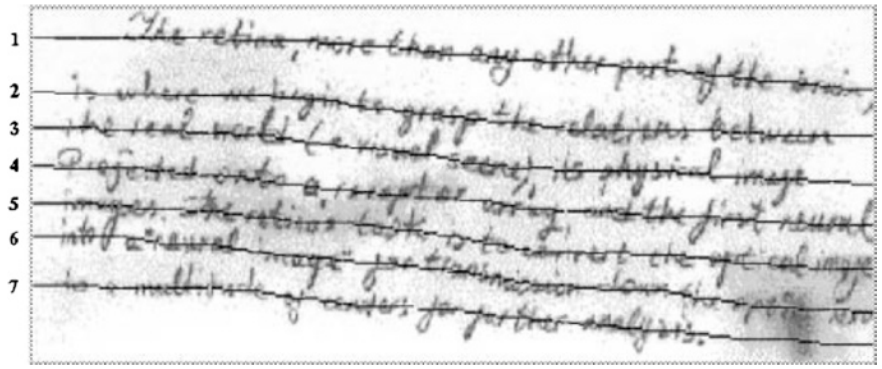

Fig. 2.5 The baseline extraction using attractive and repulsive network

characters are distinguished according to the relative position with respect to the baseline, such as 9 and g. The methods of baseline extraction include using the projection profile of the image, nearest neighbor clustering, cross correlation method between lines and Hough transform [1, 3]. An attractive repulsive nearest neighbor is used for extracting the baseline of complicated handwriting in heavy noise [1, 3] as shown in Fig. 2.5. After skew detection the character or word is translated to the origin, rotated or stretched until the baseline is horizontal and retranslated back into the display screen space.

(ii) Slant normalization: One of the measurable factors of different handwriting styles is the slant angle between longest stroke in a word and the vertical direction. Slant normalization is used to normalize all characters to a standard form. The most common method for slant estimation is the calculation of the average angle of near vertical elements as shown in Fig. 2.6a, b. The vertical line elements from contours are extracted by tracing chain code components using a pair of one dimensional filters [1, 3]. The coordinates of the start and end points of each line element provide the slant angle. The projection profiles are computed for a number of angles away from the vertical direction [1, 3]. The angle corresponding to the projection with the greatest positive derivative is used to detect the least amount of overlap between vertical strokes and the dominant slant angle. The slant detection is performed by dividing the image into vertical and horizontal windows [1, 3]. The slant is estimated based on the center of gravity of the upper and lower half of each window averaged over all the windows. A variant of the Hough transform is used by scanning left to right across the image and calculating projections in the direction of 21 different slants [1, 3]. The top three projections for any slant are added and the slant with the largest count is taken as the slant value. In some cases the recognition systems do not use slant correction and compensate it during training stage [1, 3].

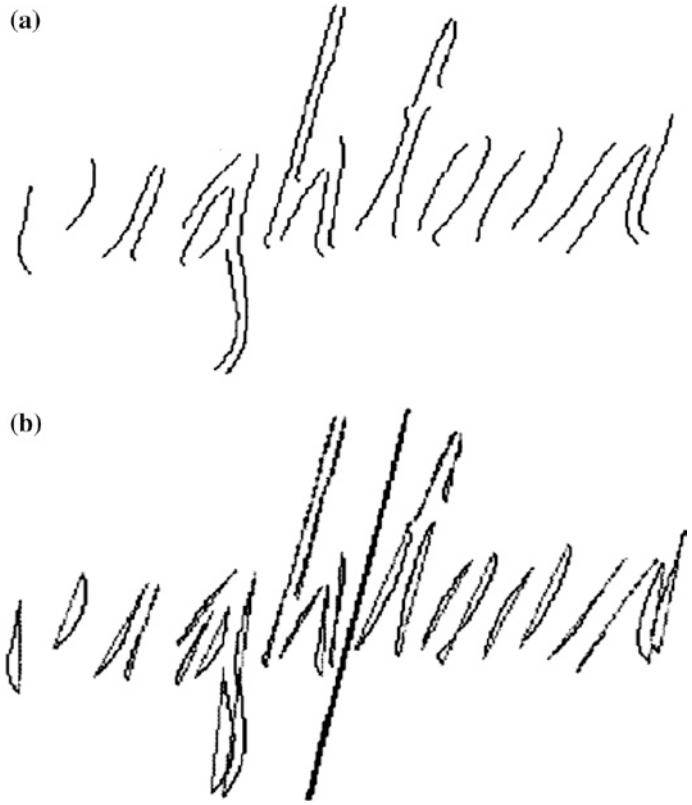

Fig. 2.6 **a** Slant angle estimation near vertical elements. **b** Slant angle estimation average slant angle

(iii) Size Normalization is used to adjust the character size to a certain standard. The OCR methods may apply for both horizontal and vertical size normalizations. The character is divided into number of zones and each of these zones is separately scaled [1, 3]. The size normalization can also be performed as a part of the training stage and the size parameters are estimated separately for each particular training data [1, 3]. In Fig. 2.7 two sample characters are gradually shrunk to the optimal size which maximize the recognition rate in the training data. The word recognition preserves large intra class differences in the length of words so they may also assist in recognition; it tends to only involve vertical height normalization or bases the horizontal size normalization on the scale factor calculated for vertical normalization [1, 3].

(iv) Contour smoothing eliminates the errors due to the erratic hand motion during the writing. It generally reduces the number of sample points needed to represent the script and thus improves efficiency in remaining pre-processing steps [1, 3].

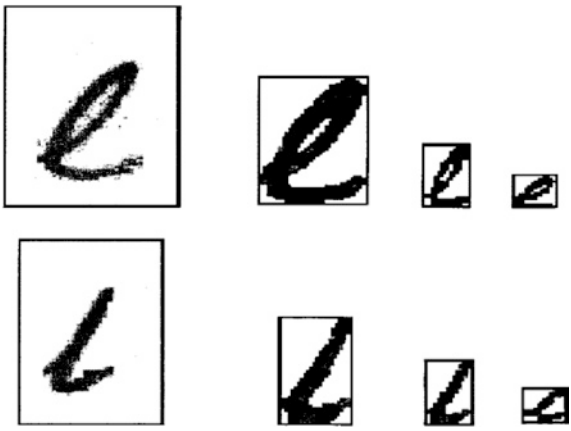

Fig. 2.7 The normalization of characters

(c) Compression: It is well known that classical image compression techniques transform the image from the space domain to domains which are not suitable for recognition. The compression for OCR requires space domain techniques for preserving the shape information. The two popular compression techniques used are: (i) thresholding and (ii) thinning.

(i) Thresholding: In order to reduce storage requirements and to increase processing speed it is often desirable to represent gray scale or color images as binary images by picking a threshold value. The two important categories of thresholding are viz global and local. The global thresholding picks one threshold value for the entire character image which is often based on an estimation of the background level from the intensity histogram of the image [1, 3]. The local or adaptive thresholding use different values for each pixel according to the local area information [1, 3]. A comparison of common global and local thresholding techniques is given by using an evaluation criterion that is goal directed keeping in view of the desired accuracy of the OCR system [1, 3]. It has been shown that Niblack's locally adaptive method [1, 3] produces the best result. An adaptive logical method is developed [1, 3] by analyzing the clustering and connection characteristics of the characters in degraded images.

(ii) Thinning: While it provides a tremendous reduction in data size, thinning extracts the shape information of the characters. Thinning can be considered as conversion of offline handwriting to almost online like data with spurious branches and artifacts. The two basic approaches for thinning are based on pixel wise and non-pixel wise thinning [1, 3]. The pixel wise thinning methods locally and iteratively process the image until one pixel wide skeleton remains. They are very sensitive to noise and deforms the shape of the character. The non-pixel wise methods use

some global information about the character during the thinning. They produce a certain median or center line of the pattern directly without examining all the individual pixels [1, 3]. The clustering based thinning method [1, 3] defines the skeleton of character as the cluster centers. Some thinning algorithms identify the singular points of the characters such as end points, cross points and loops [1, 3]. These points are the source of problems. In a non-pixel wise thinning they are handled with global approaches [1, 3]. The iterations for thinning can be performed either in sequential or parallel algorithms. The sequential algorithms examine the contour points by raster scan or contour following [1, 3]. The parallel algorithms are superior to sequential ones since they examine all the pixels simultaneously using the same set of conditions for deletion [1, 3]. They can be efficiently implemented in parallel hardware [1, 3].

It is to be noted that the above techniques affect the data and may introduce unexpected distortions to the character image. As a result these techniques may cause the loss of important information about writing and thus should be applied with care.

2.3.4 Segmentation

The pre-processing stage yields a clean character image in the sense that a sufficient amount of shape information, high compression, and low noise on a normalized image is obtained. The next OCR component is segmentation. Here the character image is segmented into its subcomponents. Segmentation is important because the extent one can reach in separation of the various lines in the characters directly affects the recognition rate. Internal segmentation is used here which isolates lines and curves in the cursively written characters. Though several remarkable methods have developed in the past and a variety of techniques have emerged, the segmentation of cursive characters is an unsolved problem. The character segmentation strategies are divided into three categories [1, 3]: (a) explicit segmentation (b) implicit segmentation and (c) mixed strategies.

(a) In explicit segmentation the segments are identified based on character like properties. The process of cutting up the character image into meaningful components is achieved through dissection. Dissection analyzes the character image without using a specific class of shape information. The criterion for good segmentation is the agreement of general properties of the segments with those expected for valid characters. The available methods based on the dissection of the character image use white space and pitch, vertical projection analysis, connected component analysis and landmarks. The explicit segmentation can be subjected to evaluation using the linguistic context [1, 3].

(b) The implicit segmentation strategy is based on recognition. It searches the image for components that matches the predefined classes. The segmentation is performed by using the recognition confidence including syntactic or semantic correctness of the overall result. In this approach two classes of methods are employed viz (i) methods that make some search process and (ii) methods that segment a feature representation of the image [1, 3]. The first class attempts to segment characters into units without use of feature based dissection algorithms. The image is divided systematically into many overlapping pieces without regard to content. These methods originate from schemes developed for the recognition of machine printed words [1, 3]. The basic principle is to use a mobile window of variable width to provide sequences of tentative segmentations which are confirmed by OCR. The second class of methods segments the image implicitly by classification of subsets of spatial features collected from the image as a whole. This can be done either through hidden markov chains or non markov based approaches. The non markov approach stem from the concepts used in machine vision for recognition of occluded object [1, 3]. This recognition based approach uses probabilistic relaxation, the concept of regularities and singularities and backward matching [1, 3].

(c) The mixed strategies combine explicit and implicit segmentation in a hybrid way. A dissection algorithm is applied to the character image, but the intent is to over segment i.e. to cut the image in sufficiently many places such that the correct segmentation boundaries are included among the cuts made. Once this is assured, the optimal segmentation is sought by evaluation of subsets of the cuts made. Each subset implies a segmentation hypothesis and classification is brought to bear to evaluate the different hypothesis and choose the most promising segmentation [1, 3]. The segmentation problem is formulated [1, 3] as finding the shortest path of a graph formed by binary and gray level document image. The hidden markov chain probabilities obtained from the characters of a dissection algorithm are used to form a graph [1, 3]. The optimum path of this graph improves the result of the segmentation by dissection and hidden markov chain recognition. The mixed strategies yield better results compared to explicit and implicit segmentation methods. The error detection and correction mechanisms are often embedded into the systems. The wise usage of context and classifier confidence generally leads to improved accuracy [1, 3].

2.3.5 Representation

The fifth OCR component is representation. The image representation plays one of the most important roles in any recognition system. In the simplest case, gray level or binary images are fed to a recognizer. However, in most of the recognition systems in order to avoid extra complexity and to increase the accuracy of

the algorithms, a more compact and characteristic representation is required. For this purpose, a set of features is extracted for each class that helps distinguish it from other classes while remaining invariant to characteristic differences within the class [1, 3]. The character image representation methods are generally categorized into three major groups: (a) global transformation and series expansion (b) statistical representation and (c) geometrical and topological representation.

(a) Global transformation and series expansion: A continuous signal generally contains more information than needs to be represented for the purpose of classification. This may be true for discrete approximations of continuous signals as well. One way to represent a signal is by a linear combination of a series of simpler well defined functions. The coefficients of the linear combination provide a compact encoding known as transformation or series expansion. Deformations like translation and rotations are invariant under global transformation and series expansion. Some common transform and series expansion methods used in OCR are: (i) fourier transform (ii) gabor transform (iii) wavelets (iv) moments and (v) karhunen loeve expansion.

   (i) Fourier transform: The general procedure is to choose magnitude spectrum of the measurement vector as the features in an n-dimensional euclidean space. One of the most attractive properties of the fourier transform is the ability to recognize the position shifted characters when it observes the magnitude spectrum and ignores the phase.
   (ii) Gabor transform: This is a variation of the windowed fourier transform. In this case, the window used is not a discrete size but is defined by a gaussian function [1, 3].
   (iii) Wavelet transformation is a series expansion technique that allows us to represent the signal at different levels of resolution. The segments of character image correspond to the units of the character and are represented by wavelet coefficients corresponding to various levels of resolution. These coefficients are then fed to a classifier for recognition [1, 3]. The representation in multi resolution analysis with low resolution absorbs the local variation in handwriting as opposed to the high resolution. However, the representation in low resolution may cause the important details for the recognition stage to be lost.
   (iv) Moments such as central moments, legendre moments and zernike moments form a compact representation of the original character image that make the process of recognizing an object scale, translation and rotation invariant [1, 3]. The moments are considered as series expansion representation since the original character image can be completely reconstructed from the moment coefficients.
   (v) Karhunen loeve expansion is an eigenvector analysis which attempts to reduce the dimension of the feature set by creating new features that are linear combinations of the original ones. It is the only optimal transform in terms of information compression. Karhunen loeve expansion is used in several pattern recognition problems such as face recognition.

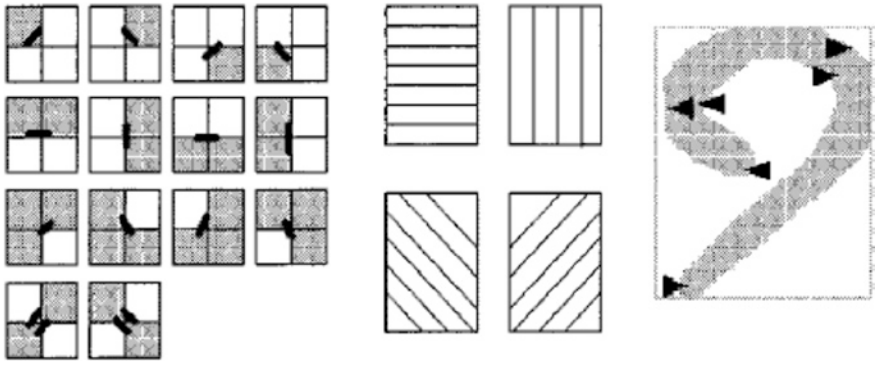

Fig. 2.8 The contour direction and bending point features with zoning

It is also used in the National Institute of Standards and Technology (NIST) OCR system for form based handprint recognition [1, 3]. Since it requires computationally complex algorithms, the use of karhunen loeve features in OCR problems is not widespread. However, by the continuous increase of the computational power, it has become a realistic feature for the current OCR systems [1, 3].

(b) Statistical representation: The representation of a character image by statistical distribution of points takes care of style variations to some extent. Although this type of representation does not allow the reconstruction of the original image, it is used for reducing the dimension of the feature set providing high speed and low complexity. Some of the major statistical features used for character representation are: (i) zoning (ii) crossings and distances and (iii) projections.

(i) Zoning: The frame containing the character is divided into several overlapping or non-overlapping zones. The densities of the points or some features in different regions are analyzed and form the representation. For example, contour direction features measure the direction of the contour of the character [1, 3] which are generated by dividing the image array into rectangular and diagonal zones and computing histograms of chain codes in these zones. Another example is the bending point features which represent high curvature points, terminal points and fork points [1, 3]. The Fig. 2.8 indicates contour direction and bending point features.

(ii) Crossings and distances: A popular statistical feature is the number of crossing of a contour by a line segment in a specified direction. The character frame is partitioned into a set of regions in various directions and then the black runs in each region are coded by the powers of two [1, 3]. Another study encodes the location and number of transitions from background to foreground pixels along vertical lines through the

word [1, 3]. Also, the distance of line segments from a given boundary such as the upper and lower portion of the frame, they can be used as statistical features [1, 3]. These features imply that a horizontal threshold is established above, below and through the center of the normalized character. The number of times the character crosses a threshold becomes the value of that feature. The obvious intent is to catch the ascending and descending portions of the character.

(iii) Projections: The characters can be represented by projecting the pixel gray values onto lines in various directions. This representation creates a one dimensional signal from a two dimensional character image which can be used to represent the character image [1, 3].

(c) Geometrical and topological representation: The various global and local properties of characters can be represented by geometrical and topological features with high tolerance to distortions and style variations. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object. The topological and geometrical representations can be grouped into: (i) extracting and counting topological structures (ii) measuring and approximating the geometrical properties (iii) coding and (iv) graphs and trees.

(i) Extracting and counting topological structures: In this representation group, a predefined structure is searched in a character. The number or relative position of these structures within the character forms a descriptive representation. The common primitive structures are the strokes which make up a character. These primitives can be as simple as lines and arcs which are the main strokes of Latin characters and can be as complex as curves and splines making up Arabic or Chinese characters. In online OCR, a stroke is also defined as a line segment from pen down to pen up [1, 3]. The characters can be successfully represented by extracting and counting many topological features such as the extreme points, maxima and minima, cusps above and below a threshold, openings to the right, left, up, and down, cross points, branch points, line ends, loops, direction of a stroke from a special point, inflection between two points, isolated dots, a bend between two points, symmetry of character, horizontal curves at top or bottom, straight strokes between two points, ascending, descending, and middle strokes and relations among the stroke that make up a character [1, 3]. The Fig. 2.9 indicates some of the topological features.

(ii) Measuring and approximating the geometrical properties: In many studies [1, 3] the characters are represented by the measurement of the geometrical quantities such as the ratio between width and height of the bounding box of a character, the relative distance between the last point and the last y-min, the relative horizontal and vertical distances between first and last points, distance between two points, comparative lengths

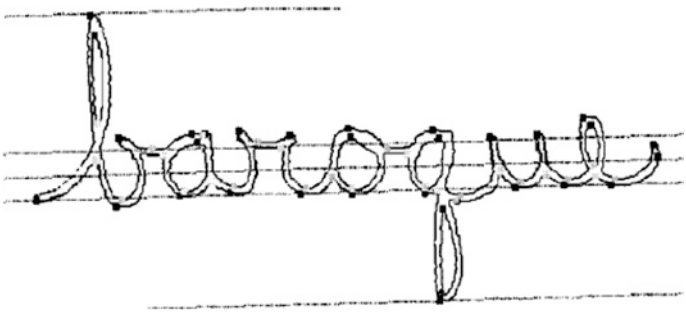

Fig. 2.9 The topological features: Maxima and minima on the exterior and interior contours, reference lines, ascenders and descenders

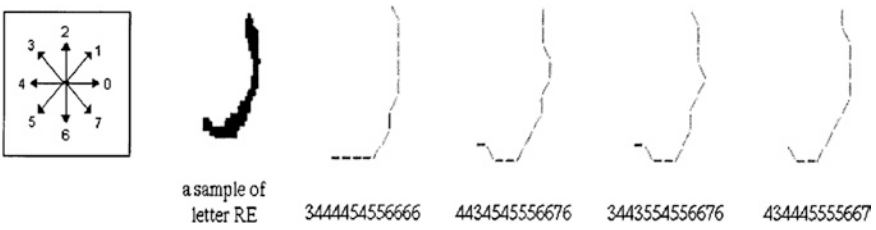

Fig. 2.10 Sample arabic character and the chain codes of its skeleton

between two strokes and width of a stroke. A very important characteristic measure is the curvature or change in the curvature [1, 3]. Among many methods for measuring the curvature information, the suggestion measures local stroke direction distribution for directional decomposition of the character image. The measured geometrical quantities can be approximated by a more convenient and compact geometrical set of features. A class of methods includes polygonal approximation of a thinned character [1, 3]. A more precise and expensive version of the polygonal approximation is the cubic spline representation [1, 3].

(iii) Coding: One of the most popular coding schema is freeman's chain code. This coding is essentially obtained by mapping the strokes of a character into a two dimensional parameter space which is made up of codes as shown in Fig. 2.10. There are many versions of chain coding. As an example, the character frame is divided to left right sliding window and each region is coded by the chain code.

(iv) Graphs and trees: The characters are first partitioned into a set of topological primitives such as strokes, loops, cross points etc. Then these primitives are represented using attributed or relational graphs [1, 3]. There are two kinds of image representation by graphs. The first kind uses the coordinates of the character shape [1, 3]. The second kind is

an abstract representation with nodes corresponding to the strokes and edges corresponding to the relationships between the strokes [1, 3]. The trees can also be used to represent the characters with a set of features which have a hierarchical relation [1, 3]. The feature extraction process is performed mostly on binary images. However, binarization of a gray level image may remove important topological information from characters. In order to avoid this problem some studies attempt to extract features directly from grayscale character images [1, 3].

In conclusion, the major goal of representation is to extract and select a set of features which maximizes the recognition rate with the least amount of elements. The feature extraction and selection is defined [1, 3] as extracting the most representative information from the raw data which minimizes the within class pattern variability while enhancing the between class pattern variability.

2.3.6 Feature Extraction

The sixth OCR component is feature extraction. The objective of feature extraction is to capture essential characteristics of symbols. Feature extraction is accepted as one of the most difficult problems of pattern recognition. The most straight forward way of describing character is by actual raster image. Another approach is to extract certain features that characterize symbols but leaves the unimportant attributes. The techniques for extraction of such features are divided into three groups' viz. (a) distribution of points (b) transformations and series expansions and (c) structural analysis. The different groups of features are evaluated according to their noise sensitivity, deformation, ease of implementation and use. The criteria used in this evaluation are: (a) robustness in terms of noise, distortions, style variation, translation and rotation and (b) practical usage in terms of recognition speed, implementation complexity and independence. Some of the commonly used feature extraction techniques are template matching and correlation, transformations, distribution of points and structural analysis. For interested readers further details are available in [11].

Another important task associated with feature extraction is classification. Classification is the process of identifying each character and assigning to it correct character class. The two important categories of classification approaches for OCR are decision theoretic and structural methods. In decision theoretic recognition character description is numerically represented in feature vector. There may also be pattern characteristics derived from physical structure of character which are not as easily quantified. Here relationship between the characteristics may important when deciding on class membership. For example, if we know that a character consists of one vertical and one horizontal stroke it may be either 'L' or 'T'. The relationship between two strokes is required to distinguish characters. The principal approaches to decision theoretic recognition are minimum distance

classifiers, statistical classifiers and neural networks. In structural recognition syntactic methods are the most prevalent approaches. A detailed discussion of these approaches is available in [3, 11].

2.3.7 Training and Recognition

The seventh OCR component is training and recognition. OCR systems extensively use the methodologies of pattern recognition which assigns an unknown sample into a predefined class. The OCR are investigated in four general approaches of pattern recognition as suggested in [1, 3]: (a) template matching (b) statistical techniques (c) structural techniques and (d) ANNs. These approaches are neither necessarily independent nor disjointed from each other. Occasionally, an OCR technique in one approach can also be considered to be a member of other approaches. In all of the above approaches, OCR techniques use either holistic or analytic strategies for the training and recognition stages. The holistic strategy employs top down approaches for recognizing the full character eliminating the segmentation problem. The price for this computational saving is to constrain the problem of OCR to limited vocabulary. Also, due to the complexity introduced by the representation of a single character or stroke the recognition accuracy is decreased. On the other hand, the analytic strategies employ bottom up approach starting from stroke or character level and going toward producing a meaningful text. The explicit or implicit segmentation algorithms are required for this strategy, not only adding extra complexity to the problem but also introducing segmentation error to the system. However, with the cooperation of segmentation stage, the problem is reduced to the recognition of simple isolated characters or strokes, which can be handled for unlimited vocabulary with high recognition rates.

(a) Template matching: The OCR techniques vary widely according to the feature set selected from the long list of features described in the previous section for image representation. The features can be as simple as the gray-level image frames with individual characters complicated as graph representation of character primitives. The simplest way of OCR is based on matching the stored prototypes against the character to be recognized. Generally speaking, matching operation determines the degree of similarity between two vectors such as group of pixels, shapes, curvature etc. in the feature space. The matching techniques can be classified in three classes: (i) direct matching (ii) deformable templates and elastic matching and (iii) relaxation matching.

    (i) Direct matching: A gray level or binary input character is directly compared to a standard set of stored prototypes. According to the similarity measures such as euclidean, mahalanobis, jaccard or yule, a prototype matching is done for recognition. The matching techniques can be as simple as one-to-one comparison or as complex as decision tree analysis in which only selected pixels are tested. A template matcher can

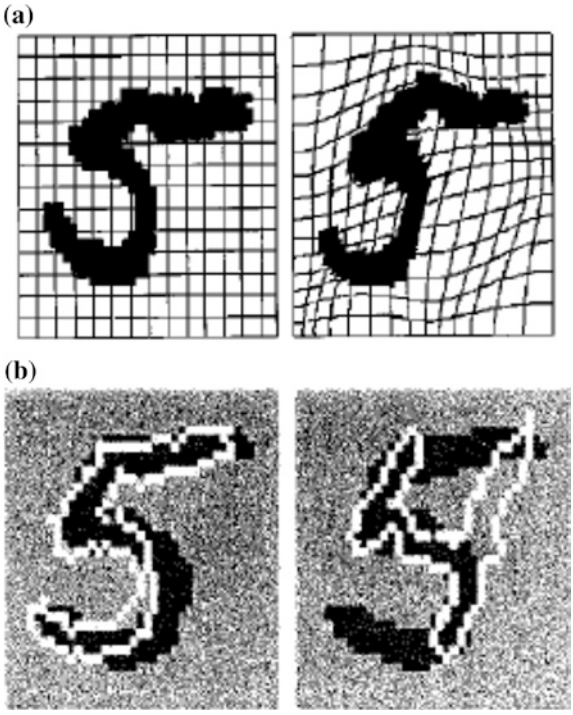

Fig. 2.11 **a** The deformable templates: deformations of a sample digit. **b** The deformable templates: deformed template superimposed on target image with dissimilarity measures

combine multiple information sources including match strength and *k*-nearest neighbor measurements from different metrics [1, 3]. Although direct matching method is intuitive and has a solid mathematical background, the recognition rate of this method is very sensitive to noise.

(ii) Deformable templates and elastic matching: An alternative method is the use of deformable templates where an image deformation is used to match an unknown image against a database of known images. The two characters are matched by deforming the contour of one to fit the edge strengths of the other [1, 3]. A dissimilarity measure is derived from the amount of deformation needed, the goodness of fit of the edges and the interior overlap between the deformed shapes as shown in Fig. 2.11a, b. The basic idea of elastic matching is to optimally match the unknown symbol against all possible elastic stretching and compression of each prototype. Once the feature space is formed, the unknown vector is matched using dynamic programming and a warping function [1, 3]. Since the curves obtained from the skeletonization of the characters could be distorted, elastic-matching methods cannot deal with topological correlation between two patterns in the OCR. In order to avoid this difficulty, a self-organization matching approach is proposed in [1, 3] for hand printed OCR using thick strokes. The elastic matching is also popular in on-line recognition systems [1, 3].

(iii) Relaxation matching: It is a symbolic level image matching technique that uses feature based description for the character image. First the matching regions are identified. Then based on some well-defined ratings of the assignments, the image elements are compared to the model. This procedure requires a search technique in a multidimensional space for finding the global maximum of some functions [1, 3]. A handwritten Chinese character system is proposed [1, 3] where a small number of critical structural features such as end points, hooks, T-shape, cross, and corner are used. The recognition is done by computing the matching probabilities between two features by a relaxation method. The matching techniques mentioned above are sometimes used individually or combined in many ways as part of the OCR schemes.

(b) Statistical techniques: The statistical decision theory is concerned with statistical decision functions and a set of optimality criteria which maximizes the probability of the observed pattern given the model of a certain class [1, 3]. The statistical techniques are mostly based on three major assumptions viz (i) The distribution of the feature set is Gaussian or in the worst case uniform (ii) There are sufficient statistics available for each class and (iii) Given ensemble of images $\{I\}$ one is able to extract a set of features $\{f_i\} \in F;\ i = \{1, \ldots, n\}$ which represents each distinct class of patterns. The measurements taken from n features of each character unit can be thought to represent an n-dimensional vector space and the vector whose coordinates correspond to the measurements taken represents the original character unit. The major statistical approaches applied in the CR field are: (i) nonparametric recognition (ii) parametric recognition (iii) clustering analysis (iv) hidden markov chains and (v) fuzzy set reasoning.

(i) Nonparametric recognition: This method is used to separate different pattern classes along hyperplanes defined in a given hyperspace. The best known method of nonparametric classification is ANN and is extensively used in OCR [1, 3]. It does not require apriori information about the data. An incoming pattern is classified using the cluster whose center is the minimum distance from the pattern over all the clusters.

(ii) Parametric recognition: Since apriori information is available about the characters in the training data, it is possible to obtain a parametric model for each character [1, 3]. Once the parameters of the model which are based on some probabilities are obtained, the characters are classified according to some decision rules such as maximum likelihood or bayes method.

(iii) Clustering analysis: The clusters of character features which represent distinct classes are analyzed by way of clustering methods. Clustering can be performed either by agglomerative or divisive algorithms. The agglomerative algorithms operate step-by-step merging of small clusters into larger ones by a distance criterion. On the other hand, the divisive methods split the character classes under certain rules for identifying the underlying character [1, 3].

(iv) Hidden markov chains: The hidden markov chains are the widely and successfully used technique for handwritten OCR problems [1, 3]. It is defined as a stochastic process generated by two interrelated mechanisms consisting of a markov chain having a finite number of states and a set of random functions each of which is associated with a state [1, 3]. At discrete instants of time the process is assumed to be in some state and an observation is generated by the random function corresponding to the current state. The underlying Markov chain then changes states according to its transitional probabilities. Here, the job is to build a model that explains and characterizes the occurrence of the observed symbols [1, 3]. The output corresponding to a single symbol can be characterized as discrete or continuous. The discrete outputs may be characters from a finite alphabet or quantized vectors from a codebook while continuous outputs are represented by samples from a continuous waveform. In generating a character, the system passes from one state to another each state emitting an output according to some probabilities until the entire character is obtained. There are two basic approaches to OCR systems using hidden markov chains such as model and path discriminant hidden markov chains [1, 3].

(v) Fuzzy set reasoning: The fuzzy set reasoning employs fuzzy set elements in describing the similarities between the features of the characters. The fuzzy set elements give more realistic results when there is no a priori knowledge about the data and therefore the probabilities cannot be calculated. The characters can be viewed as a collection of strokes which are compared to reference patterns by fuzzy similarity measures. Since the strokes under consideration are fuzzy in nature, the fuzziness is utilized in the similarity measure. In order to recognize a character, an unknown input character is matched with all the reference characters and is assigned to the class of the reference character with the highest score of similarity among all the reference characters. The fuzzy similarity [1, 3] measure is utilized to define the fuzzy entropy for handwritten Chinese characters. A handwritten OCR system is proposed [1, 3] using a fuzzy graph theoretic approach where each character is described by a fuzzy graph. A fuzzy graph matching algorithm is then used for recognition. An algorithm is presented which uses average values of membership for final decision [1, 3].

(c) Structural techniques: The recursive description of a complex pattern in terms of simpler patterns based on the shape of the object was the initial idea behind the creation of the structural pattern recognition. These patterns are used to describe and classify the characters in OCR systems. The characters are represented as the union of the structural primitives. It is assumed that the character primitives extracted from writing are quantifiable and one can find the relations among them. The structural methods are applied to the OCR problems are: (i) grammatical methods and (ii) graphical methods.

(i) Grammatical methods: The grammatical methods consider the rules of linguistics for analyzing the written characters. Later various orthographic, lexicographic and linguistic rules were applied to the recognition schemes. The grammatical methods create some production rules in order to form the characters from a set of primitives through formal grammars. These methods may combine any type of topological and statistical features under some syntactic and semantic rules [1, 3]. Formal tools like language theory allows to describe the admissible constructions and to extract the contextual information about the writing by using various types of grammars such as string grammars, graph grammars, stochastic grammars and picture description language [1, 3]. In grammatical methods, training is done by describing each character by a grammar G_i. In the recognition phase, the string, tree or graph of any character is analyzed in order to decide to which pattern grammar it belongs [1, 3]. The top down or bottom up parsing does syntax analysis. The grammatical methods in OCR area are applied in various character levels [1, 3]. In character level, picture description language is used to model each character in terms of a set of strokes and their relationship. This approach has been used for Indian OCR where Devanagari characters are presented by a picture description language [1, 3]. The system stores the structural description in terms of primitives and the relations. The recognition involves a search for the unknown character based on the stored description. The grammatical methods are mostly used in the post-processing stage for correcting the recognition errors [1, 3].

(ii) Graphical methods: The characters can also be represented by trees, graphs, digraphs or attributed graphs. The character primitives such as strokes are selected by a structural approach irrespective of how the final decision making is made in the recognition [1, 3]. For each class, a graph or tree is formed in the training stage to represent strokes. The recognition stage assigns the unknown graph to one of the classes by using a graph similarity measure. There are a variety of approaches that use the graphical methods. The hierarchical graph representation approach is used for handwritten Chinese OCR [1, 3]. Simon have proposed [1, 3] an off-line cursive script recognition scheme. The features are regularities which are defined as uninformative parts and singularities which are defined as informative strokes about the characters. The stroke trees are obtained after skeletonization. The goal is to match the trees of singularities. Although it is computationally expensive, relaxation matching is a popular method in graphical approaches to the OCR systems [1, 3].

(d) ANNs: The ANN possess a massively parallel architecture such that it performs computation at a higher rate compared to the classical techniques. It adapts to the changes in data and learns the characteristics of input signal. ANN contains many nodes. The output from one node is fed to another one in the network and the final decision depends on the complex interaction

of all nodes. In spite of the different underlying principles, it can be shown that most of the ANN architectures are equivalent to statistical pattern recognition methods [1, 3]. Several approaches exist for training of ANNs [1, 3]. These include the error correction, boltzman, hebbian and competitive learning. They cover binary and continuous valued input as well as supervised and unsupervised learning. The ANN architectures are classified into two major groups viz feedforward and feedback (recurrent) networks. The most common ANNs used in the OCR systems are the multilayer perceptron of the feedforward networks and the kohonen's self-organizing map of the feedback networks. The multilayer perceptron proposed by Rosenblatt [1, 3] and elaborated by Minsky and Papert [1, 3] has been applied in OCR. An example is the feature recognition network proposed by Hussain and Kabuka [1, 3] which has a two-level detection scheme. The first level is for detection of sub-patterns and the second level is for detection of the characters. The neocognitron of Fukushima [1, 3] is a hierarchical network consisting of several layers of alternating neuron-like cells. S-cells are used for feature extracting and C-cells allow for positional errors in the features. The last layer is the recognition layer. Some of the connections are variable and are be modified by learning. Each layer of S and C cells are called cell planes. Here training patterns useful for deformation-invariant recognition of a large number of characters are selected. The feedforward ANN approach to the machine printed OCR problem has proven to be successful [1, 3] where ANN is trained with a database of 94 characters and tested in 300,000 characters generated by a postscript laser printer with 12 common fonts in varying size. Here Garland et al. propose a two-layer ANN trained by a centroid dithering process. The modular ANN architecture is used for unconstrained handwritten numeral recognition in [1, 3]. The whole classifier is composed of subnetworks. A subnetwork which contains three layers is responsible for a class among ten classes. Most of the recent developments on handwritten OCR research are concentrated on Kohonen's self-organizing map (SOM) [1, 3]. SOM integrates the feature extraction and recognition steps in a large training set of characters. It can be shown that it is analogous to *k*-means clustering algorithm. An example of SOM on OCR systems is the study by [1, 3] which presents a self-organization matching approach to accomplish the recognition of handwritten characters drawn with thick strokes. In [1, 3] a combination of modified SOM and learning vector quantization is proposed to define a three-dimensional ANN model for handwritten numeral recognition. Higher recognition rates are reported with shorter training time than other SOMs.

2.3.8 Post-processing

The eighth OCR component is post-processing. Some of the commonly used post-processing activities include grouping and error detection and correction. In

grouping symbols in text are associated with strings. The result of plain symbol recognition in text is a set of individual symbols. However, these symbols do not usually contain enough information. These individual symbols are associated with each other making up words and numbers. The grouping of symbols into strings is based on symbols' location in document. The symbols which are sufficiently close are grouped together. For fonts with fixed pitch grouping process is easy as position of each character is known. For typeset characters distance between characters are variable. The distance between words are significantly large than distance between characters and grouping is therefore possible. The problems occur for handwritten characters when text is skewed. Until grouping each character is treated separately, the context in which each character appears has not been exploited. However, in advanced optical text recognition problems, system consisting only of single character recognition is not sufficient. Even best recognition systems will not give 100% correct identification of all characters [3, 4, 7]. Only some of these errors are detected or corrected by the use of context. There are two main approaches. The first utilizes the possibility of sequences of characters appearing together. This is done by using rules defining syntax of word. For different languages the probabilities of two or more characters appearing together in sequence can be computed and is utilized to detect errors. For example, in English language probability of k appearing after h in a word is zero and if such a combination is detected an error is assumed. Another approach is dictionaries usage which is most efficient error detection and correction method. Given a word in which an error is present and the word is looked up in dictionary. If the word is not in dictionary an error is detected and is corrected by changing word into most similar word. The probabilities obtained from classification helps to identify character erroneously classified. The error transforms word from one legal word to another and such errors are undetectable by this procedure. The disadvantage of dictionary methods is that searches and comparisons are time consuming.

2.4 Applications of Optical Character Recognition Systems

The last few decades have seen a widespread appearance of commercial OCR products satisfying requirements of different users. In this section we highlight some notable application areas of OCR. The major application areas are often distinguished as data entry, text entry and process automation. Interested readers can refer [3, 8, 11] for different OCR application areas.

The data entry area [7] covers technologies for entering large amounts of restricted data. Initially such machines were used for banking applications. The systems are characterized by reading only limited set of printed characters usually numerals and few special symbols. They are designed to read data like account numbers, customer's 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 up to 150 documents per hour. Single character error and reject rates are 0.0001 and 0.01% respectively. Due to limited character set these readers are usually tolerant to bad printing quality. These systems are specially designed for their applications and prices are therefore high.

The text entry reading machines [7] are used as page readers 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 word processing environment. These page readers are in strong competition with direct key-input and electronic exchange of data. As character set read by these machines is rather large, the performance is extremely dependent on quality of the printing. However, under controlled conditions single character error and reject rates are about 0.01 and 0.1% respectively. The reading speed is few hundred characters per second.

In process automation [7] major concern is not to read what is printed but rather to control some particular process. This is actually automatic address reading technology for mail sorting. Hence, the goal is to direct each letter into appropriate bin regardless of whether each character was correctly recognized or not. The general approach is to read all information available and use postcode as redundancy check. The acceptance rate of these systems is obviously dependent on properties of mail. This rate therefore varies with percentage of handwritten mail. Although rejection rate for mail sorting may be large, miss rate is usually close to zero. The sorting speed is typically about 30 letters per hour.

The abovementioned application areas are those in which OCR has been successful and widely used. However, many other areas of applications exist and some of which are [3, 7]:

(a) Aid for blind: In the early days before digital computers and requirement for input of large amounts of data emerged this was an imagined application area for reading machines. Along with speech synthesis system such reader enables blind to understand printed documents.
(b) Automatic number plate readers: A few systems for automatic reading of number plates of cars exist. As opposed to other OCR applications, input image is not natural bilevel image and must be captured by very fast camera. This creates special problems and difficulties although character set is limited and syntax restricted.
(c) Automatic cartography: The character recognition from maps presents special problems within character recognition. The symbols are intermixed with graphics, text is printed at different angles and characters are of several fonts or even handwritten.
(d) Form readers: Such systems are able to read specially designed forms. In such forms all irrelevant information to reading machine is printed in colour invisible to scanning device. The fields and boxes indicating where to enter text is printed in this invisible colour. The characters are in printed or hand written upper case letters or numerals in specified boxes. The instructions are often printed on form as how to write each character or numeral. The processing

speed is dependent on amount of data on each form but may be few hundred forms per minute. The recognition rates are seldom given for such systems.

(e) Signature verification and identification: This application is useful for banking environment. Such system establishes the identity of writer without attempting to read handwriting. The signature is simply considered as pattern which is matched with signatures stored in reference database.

2.5 Status of Optical Character Recognition Systems

A wide variety of OCR systems are currently commercially available [7]. In this section we explore the capabilities of OCR systems and the main problems encountered therein. We also take a step forward in discussing the evaluation performance of an OCR system.

OCR systems are generally divided into two classes [3]. The first class includes special purpose machines dedicated to specific recognition problems. The second class covers systems that are based on PC and low cost scanner. The first recognition machines are all hardwired devices. As these hardware were expensive, throughput rates were high to justify cost and parallelism was exploited. Today such systems are used in specific applications where speed is of high importance. For example, within areas of mail sorting and check reading. The cost of these machines are still high up to few million dollars and they recognize wide range of fonts. The advancements in computer technology has made it possible to fully implement recognition part of OCR in software packages which work on personal computers. The present PC systems are comparable to large scaled computers of early days and their cost of such systems are low. However, there are some limitations in such OCR software especially when it comes to speed and character sets read. The hand held scanners for reading do also exist. These are usually limited to reading of numbers and few additional letters or symbols of fixed fonts. They often read a line at a time and transmits it to application programs. A wide array of commercial software products are available over the years. The speed of these systems have grown over years. The sophistication of OCR system depends on type and number of fonts recognized [3]. Based on the OCR systems' capability to recognize different character sets, five different classes of systems are recognized viz. fixedfont, multifont, omnifont, constraint handwriting and scripts [7].

The fixedfont OCR machines [7] deal with recognition of one specific typewritten font. These fonts are characterized by fixed spacing between each character. In several standards fonts are specially designed for OCR, where each character has a unique shape to avoid ambiguity with other similar characters. Using these character sets it is quite common for commercial OCR machines to achieve recognition rate as high as 99.99% with high reading speed. The systems of first generation OCR were fixed font machines and the methods applied were based on template matching and correlation.

The multifont OCR machines [7] recognize more than one font. However, fonts recognized by these machines are usually of same type as those recognized by fixed font system. These machines appeared after fixedfont machines. They are able to read up to about ten fonts. The limit in number of fonts is due to pattern recognition algorithm and 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 fonts in library are selected with care.

An omnifont OCR machine [7] can recognize mostly non-stylized fonts without having to maintain huge databases of specific font information. Usually omnifont technology is characterized by feature extraction usage. The database of an omnifont system contains description of each symbol class instead of symbols themselves. This gives flexibility in automatic recognition of variety of fonts. Although omnifont is common for OCR systems, this should not be understood that system is able to recognize all existing fonts. No OCR machine performs equally well on all the fonts used by modern typesetters.

The recognition of constrained handwriting through OCR machine deals with the unconnected normal handwritten characters' problem. The optical readers with such capabilities are common these days [3, 4] and they exist. These systems require well-written characters and most of them recognize digits unless certain standards for hand printed characters are followed. The characters should be printed as large as possible to retain good resolution and are entered in specified boxes. The writer is instructed to keep to certain models avoiding gaps and extra loops. Commercially intelligent character recognition is often used for systems that recognize hand printed characters.

Generally all methods for character recognition described here deal with isolated character recognition problem. However, to humans it would be more interesting if it were possible to recognize entire words consisting of cursively joined characters. The script recognition deals with the problem of recognizing unconstrained handwritten characters which may be connected or cursive. In signature verification and identification the objective is to establish identity of writer irrespective of handwritten contents. The identification establishes identity of writer by comparing specific attributes of pattern describing the signature with list of writers stored in a reference database. When performing signature verification the claimed identity of writer is known and the signature pattern is matched against the signature stored in database for the person. Many such systems of this type are commercially available [4]. A more challenging problem is script recognition where contents of handwriting must be recognized. The variations in shape of handwritten characters are infinite and depend on writing habit, style, education, mood, social environment and other conditions of writer. Even best trained optical readers and humans make about 4% errors when reading. The recognition of characters written without any constraint is available in some commercially available systems [3].

The accuracy of OCR systems directly depends upon the quality of input documents [2, 7]. The major problems encountered in different documents are classified in terms of (a) shape variations due to serifs and style variations (b)

deformations caused by broken characters, smudged characters and speckle (c) spacing variations due to subscripts, superscripts, skew and variable spacing and (d) a mixture of text and graphics. These imperfections affect and cause problems in different parts of the recognition process of OCR systems resulting in rejections or misclassifications.

The majority of errors in OCR systems are often due to problems in scanning process and the following segmentation which results in joined or broken characters [3]. The segmentation process errors also results in confusion between text and graphics or between text and noise. Even if a character is printed, scanned and segmented correctly it may be incorrectly classified. This happens if character shapes are similar and selected features are not sufficiently efficient in separating different classes or if the features are difficult to extract and has been computed incorrectly. The incorrect classification also happens due to poor classifier design. This occurs if the classifier has not been trained on sufficient test samples representing all the possible forms of each character. The errors may also creep in due to post processing when isolated symbols are associated to reconstruct the original words as characters which are incorrectly grouped. These problems occur if the text is skewed such that in some cases there is proportional spacing and symbols have subscripts or superscripts. As OCR devices employ wide range of approaches to character recognition all systems are not equally affected by the abovementioned complexities [7]. The different systems have their 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 weakest link of any OCR system.

Finally we conclude this section with some insights to the performance evaluation of OCR systems [3]. There exist no standardized test sets for character recognition. This is mainly because the performance of OCR system is highly dependent on the quality of input which makes it difficult to evaluate and compare different systems. The recognition rates are often given and usually presented as percentage of characters correctly classified. However, this does not say anything about the errors committed. Thus in evaluation of OCR system three different performance rates are investigated such as (a) recognition rate which is the proportion of correctly classified characters (b) rejection rate which is the proportion of characters which the system is unable to recognize and (c) error rate which is the proportion of characters erroneously classified. There is usually a trade-off between different recognition rates. A low error rate leads to higher rejection rate and a lower recognition rate. Because of the time required to detect and correct OCR errors, error rate is most important when evaluating the cost effectiveness of an OCR system. The rejection rate is less critical. As an example consider reading operation from barcode. Here a rejection while reading bar-coded price tag will only lead to rescanning of the code whereas a wrongly decoded price tag results in charging wrong amount to the customer. In barcode industry error rates are therefore as low as one in million labels whereas 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 recognition rates of a system. A correct recognition rate of 99% implies an error

rate of 1%. In case of text recognition on 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.

2.6 Future of Optical Character Recognition Systems

All through the years, the methods of OCR systems have improved from 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 [4] and also hand printed characters. The new methods for character recognition continue appear with development of computer technology and decrease in computational restrictions [3]. However, the greatest potential lies in exploiting existing methods by hybridizing technologies and making more use of context. The integration of segmentation and contextual analysis improves recognition of joined and split characters. Also higher level contextual analysis which looks at semantics of entire sentences are useful. Generally there is a potential in using context to larger extent than what is done today. In addition, a combination of multiple independent feature sets and classifiers where weakness of one method is compensated by the strength of another improves recognition of individual characters [2]. The research frontiers within character recognition continue to move towards recognition of sophisticated cursive script that is handwritten connected or calligraphic characters. Some promising techniques within this area deal with recognition of entire words instead of individual characters.

References

1. Arica, N., Vural, F. T. Y., An Overview of Character Recognition focused on Offline Handwriting, IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, 31(2), pp 216–233, 2001.
2. Bunke, H., Wang, P. S. P. (Editors), Handbook of Character Recognition and Document Image Analysis, World Scientific, 1997.
3. Chaudhuri, A., Some Experiments on Optical Character Recognition Systems for different Languages using Soft Computing Techniques, Technical Report, Birla Institute of Technology Mesra, Patna Campus, India, 2010.
4. Cheriet, M., Kharma, N., Liu, C. L., Suen, C. Y., Character Recognition Systems: A Guide for Students and Practitioners, John Wiley and Sons, 2007.
5. Dholakia, K., A Survey on Handwritten Character Recognition Techniques for various Indian Languages, International Journal of Computer Applications, 115(1), pp 17–21, 2015.
6. Mantas, J., An Overview of Character Recognition Methodologies, Pattern Recognition, 19(6), pp 425–430, 1986.
7. Rice, S. V., Nagy, G., Nartker, T. A., Optical Character Recognition: An Illustrated Guide to the Frontier, The Springer International Series in Engineering and Computer Science, Springer US, 1999.

8. Schantz, H. F., The History of OCR, Recognition Technology Users Association, Manchester Centre, VT, 1982.
9. Scurmann, J., Reading Machines, Proceedings of International Joint Conference on Pattern Recognition, Munich, pp 1031–1044, 1982.
10. Singh, S., Optical Character Recognition Techniques: A Survey, Journal of Emerging Trends in Computing and Information Sciences, 6 (4), pp 545–550, 2013.
11. Young, T. Y., Fu, K. S., Handbook of Pattern Recognition and Image Processing, Academic Press, 1986.
12. Yu, F. T. S., Jutamulia, S. (Editors), Optical Pattern Recognition, Cambridge University Press, 1998.

http://www.springer.com/978-3-319-50251-9

Optical Character Recognition Systems for Different Languages with Soft Computing

Chaudhuri, A.; Mandaviya, K.; Badelia, P.; K Ghosh, S.

2017, XIX, 248 p. 95 illus., Hardcover
ISBN: 978-3-319-50251-9