2.1 Historical evaluation of OCR:

1900-1960:

To mimic the human function of reading by machines is the base of the OCR system, and making it able to read the text like a human being. The beginning of character recognition system can be found in 1870 when C.R Carey of Boston Massachusetts invented retina scanner. It was an image transmission system in which a mosaic of photocells was used. After two decades the sequential scanner of such type was invented by Polish P. Nipkow. This invention played a vital role in the development of modern television and reading machines. [1]. In May, 1929 Gustav Tauschek received a patent with the name ‘Reading Machine’ in Germany. It was the first idea of a machine which would recognize the pre-written text [2, 3]. The machine used a simple-minded template matching scheme. The major problem to tackle the lack of computational resources and limit was what could actually create. He used a mechanical character mask and placed characters to be recognized under this mask and applied light through those masks. Under the input character, a photo-
detector was placed, and if the input character matched the character mask, then the light was blocked and therefore did not reach the photo detector and character could be recognized [4]. In the middle of 1940, the developments in computational resources facilitated the researcher to design optical character recognition systems. It made the development of modern character recognition system possible. Electronic data processing was attracting every sector. The punched cards were used for data entry purpose which was expensive. Due to increasing data entry requirement, a cost-effective way was needed for data entry purpose. In the beginning of 1950, scientists tried to capture the images of characters and text, first by mechanical and optical means of rotating disks an photomultiplier, flying spot scanner with a cathode ray tube lens, followed by photocells and array of them. The character recognizers were commercially available by the middle of 1950. Electronic Data processing was introduced and a new way of data entry, in 1954, first true OCR reading machine was installed at Reader’s Digest which converts typewritten sales report into punched cards for input to the computers [2]

1960-1975:

The reading machines were successfully implemented in sectors like hospital, bank, and postal services etc. The purpose of reading machine was to recognize machine printed characters. The variation in font type and size were still problematic and accuracy of the system effected. To tackle these difficulties companies together and make a decision on standardization for OCR system, which include specific printed font, paper, ink etc. Finally the basic requirements of an OCR system were studied and a standard font called OCR-A was introduced in 1966. It was globally known standard font, used for character recognition systems only. The font consists of specific features which help to be recognized by any character recognition system. The confusions in recognition of various similar looking characters were solved due to this standardization. Implementation of standardized OCR-A increased the accuracy in recognition rate greater than before. Further requirements of character recognition changed and ECMA has developed a new font known as OCR-B in 1968, it has more natural appearance to eye than previous font. Besides in few systems both fonts were merged into one standard, to make possible to a machine to read both standards. Figure 2.1 shows both OCR-A and OCR-B fonts.
1975-1990:

During this period most of the Character recognition systems were based on techniques like template matching, stroke detection, simple line and geometric features, and the extraction of their derivatives. These were not sufficient to process the specific handwritten documents like forms [1]. The difficulties for developing and handling robust character recognition system were Lack of computational power, high cost and unsatisfactory results. Limited computer processing capacity restricted researchers to implement new algorithms for feature extraction and recognition in OCR. Still the semantic information of text has not covered; most of the research work during this period was based on shape recognition techniques.

After 1990:

With high computation power researchers were able to implement new algorithms which required high processing capacity. Previously the major problem for developing and handling robust character recognition system was lack of computational power, high cost and unsatisfactory results. Commercial OCR system become available as packages and implemented in multiple sectors. For researcher new development tools and environments made available by companies. These assisted researchers to apply new algorithms and investigate the best methods for improving feature extraction and recognition rate. Advanced programming environments, high quality image capturing devices were introduced in early nineties. The image processing, pattern recognition and Artificial Intelligence methods were applied jointly by researchers. It enhanced the quality of character recognition system by improving feature extraction and classification. For preprocessing, feature extraction and classification advanced techniques were implemented. Methods like
HMM, ANN, Fuzzy Logic etc. improved the system accuracy. With these advancements the Bilingual or Multilingual Character recognition systems are easily available and implemented in various processes like automatic mail sorting, form processing, banking, examination, reservation counter etc. The progress of character recognition system is still continue, many languages and scripts are waiting to be recognized by online or offline method.

2.2 **OCR System: General Approach**

Character Recognition (OCR) is an automation of converting scanned images of machine printed or handwritten text into the computer editable format. CR system accelerate the process of preserving document in digital and editable form, to skip manual entry of text thus improving productivity and reducing typing mistakes of human error. The fundamental objective of character recognition system is converting non electronic documents into electronic documents which can be manipulated by word processing software. Simply this technology allows a machine to recognize characters automatically with the help of features of characters. Human beings recognize numerous objects based on specific features of the object in similar manner machine can understand the characters based on their unique features.

2.2.1 **Online OCR**

A character recognition system can be classifies into two basic classes, online character recognition and offline character recognition system as shown in Figure 2.2. In case of online character recognition system, characters get recognized at real time when it is written. Efficiency and recognition of online systems achieve better results than offline recognition as we get information of variation of pen moves, up and down of pen tip because online systems acquire the position of the pen as a function of time straight from the interface. This is typically done through pen-based interfaces in which the writer writes with a special pen on a digital pad or electronic tablet.

2.2.2 **Offline OCR**

A fruitful summary about character recognition is available at [5] in which a bird eye view put on this domain. It is illustrate that, “Offline Optical character recognition (OCR) is the process of converting an image of text, such as a scanned
paper document or electronic fax file, into computer-editable text. The text in an image is not editable: the letters/characters are made of tiny dots (pixels) that together form a picture of text. During OCR, the software analyzes an image and converts the pictures of the characters to editable text based on the patterns of the pixels in the image. After OCR, you can export the converted text and use it with a variety of word-processing, page layout and spreadsheet applications”.

Offline character recognition can be further classified to printed characters and handwritten character recognition. In offline character recognition, the typewritten/handwritten character is typically scanned in form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm.

![Hierarchy of Optical Character Recognition System](image)

*Figure 2.2  Hierarchy of Optical Character Recognition System*

Offline character recognition become more challenging due to shape of characters, great variation of character symbol and document quality. Therefore offline character recognition is considered as a more challenging task then its online counterpart. Handwritten character recognition is even more complex then printed character recognition as handwriting of persons are differs in various attributes like shape, size, style etc. Technologist faces challenges in developing offline character recognition. Firstly, for reading a page in unknown language, anyone may be unable to recognize the various characters. But on the same page, numerical statements can be easily interpreted because the symbols for numbers are universally used. This explains why many OCR systems recognize numbers only, while relatively few
understand the full alphanumeric character range. For example, OCR system developed for applications like reading postal code from the post-cards, reading amount from the cheque or demand draft etc. Second, there is similarity between many numerical and alphabetical symbol shapes. For example, In English character recognition, there is very little visible difference between a capital letter “O” and the numeral “0”. Humans can re-read the sentence or entire paragraph to determine the accurate meaning [5].

2.2.3 MICR - Magnetic Ink Character Recognition:

MICR Magnetic Ink Character Recognition is an offline OCR, in which the characters are printed in special typestyle with a magnetic ink or toner, usually containing iron oxide. A specific recognition system is used for MICR; each character produces a unique waveform when it passes through MICR reader head and magnetic character gets recognized by the system. Basically this system is used by banks for processing the cheques. Figure 2.3 shows Magnetic Ink Characters. A set of rules used in magnetic ink character recognition called CMC i.e. code for magnetic characters.

![Figure 2.3 Example of Magnetic ink characters used for MICR](image)

2.3 A Typical offline OCR System:

As in Offline Character Recognition System, document image is acquired using scanning device and converted into editable ASCII strings. Researchers proposed various techniques to achieve this objective. Such system can be developed by using any of the following two methods.

- Segmentation free Method (Holistic)
- Segmentation Based Method (Analytical)

In holistic approach a large lexicon i.e. word class is build for recognition of word. In this segmentation step is skipped and directly complete words are matched from the lexicons to recognize it. The words are not segmented into
characters. Instead of segmented characters, the features of words/string are observed. But this technique is suitable for class of small and fixed lexicons. Holistic approach fails if new words are to be recognized which are not present in lexicon. Usually HMM methods are preferred for holistic approach which gives superior recognition results.

In case of segmentation based approach the words are segmented into individual characters and then characters are recognized. The strings are formed by concatenating the recognized characters. And finally words are recognized using linguistic post-processing lexicon-driven and lexicon-free techniques. A typical segmentation based Character Recognition system is shown in Figure 2.4. Detail study and fruitful survey of segmentation based offline character recognition is given in [6], where the development of character recognition is divided into five major stages:

A. Preprocessing  
B. Segmentation  
C. Feature Extraction  
D. Recognition  
E. Post Processing

![Diagram of Character Recognition System](image)

*Figure 2.4 A typical Character Recognition System*
2.3.1 Preprocessing

To improve the readability and the automatic recognition of handwritten document images, preprocessing step is imperative. This step consists of sub-steps of noise removal and filtering includes text normalization such as baseline correction, slant normalization and skew correction. Performing proper preprocessing makes the feature extraction process more reliable and effective [7]. Scanned document images come across in practical OCR applications often gives poor recognition results due to many problems that can be improved with appropriate preprocessing. The inputs of OCR systems are usually scanned documents. Raw scanned documents cannot be used for character recognition and need preprocessing. For example, the input of character recognition phase is usually binary image, while the scanned documents are usually gray scale or color images. So the image must be converted to binary image in preprocessing phase. Another problem in character recognition is removing non-text regions which come under document image processing. During page segmentation Removing of non-text region perform. Noise removal is another important step which must be done during preprocessing. The noises have different sources, such as noises produced while scanning and noises related to quality of paper such as noises of scanned newspapers [8]. So in case of offline character recognition system scanned documents need to be preprocessed before forwarding to other steps of OCR systems and the preprocessing has great effect on the results of other phases [9]. Two major objectives of preprocessing are Noise reduction and Normalization

Noise Reduction:

The overall accuracy of the system is highly depends upon the noise removal from the scanned document image. Due to its importance great number of works has been done for noise removal. There are some special methods available to remove noise from documents with heavy noise [10]. For Urdu Character document image noise removal, general noise removal methods cannot be used simply since the dots are very significant and they are similar to noises. In literature of character recognition system many noise removing techniques are available few of them are underneath.

- Filtering
- Morphological operations
- Noise Modeling
Normalization:

In handwritten words and characters numerous writing variations encountered, for good outcome from OCR system invariant data is necessitate which improve the recognition rate. The normalizations process is usually expected to remove the handwriting variations and get standardized data. Normalization can be ranged as below

- Skew Normalization
- Base line Extraction
- Slat Normalization
- Size Normalization

2.3.2 Segmentation

In preprocessing stage data get cleaned and unwanted noise pixels get removed from the document image. After preprocessing normalized image is obtained and it consists of sufficient amount of shape information, high compression, and low noise. The after that segmenting stage come which split the document into its subcomponents. Segmentation is a significant stage because it is responsible for obtaining separate words, lines, or characters and it directly affects the recognition rate of the script. The Character segmentation operation decomposes an image of printed of handwritten string into sub-images of individual symbols. Its decision, that a pattern isolated from the image is that of a character (or some other identifiable unit), can be right or wrong. It is wrong sufficiently often to make a major contribution to the error rate of the system. [11]. In case of offline handwritten character recognition segmentation is tidies as complex task. Since cursive handwriting segmentation is a difficult task in an off-line environment, the use of segmentation free or also called as word-level recognition approach in place of common character-level approach has been suggested to avoid the segmentation problem altogether. While word-level recognition strategy does avoid the difficult segmentation issue, the approach is limited in its discrimination capability and is suitable only for limited vocabulary applications. As described by [6] in segmentation based approach there are two types of segmentation: external segmentation and internal segmentation. External Segmentation, which segments paragraphs, sentences or words. And in case of internal segmentation words are segmented into letters. Various available Segmentation methods can be summaries as
• The classical approach, in which segments are identified based on ‘character like’ properties. This process of cutting up the image into meaningful components is named ‘dissection’, referring to the decomposition of the image into a sequence of sub-images using general features.

• Recognition based segmentation, in which the system searches the image for components that match classes in its alphabet.

• Holistic methods (or global approach), in which the system seeks to recognize whole words, avoiding the need to segment them into characters.

Segmentation is one of the most difficult tasks in Urdu character recognition system, development of robust Urdu OCR system falls behind due to word and character segmentation dilemma. Due to cursive in writing and space insertion and space deletion problems in handwritten Urdu text, the segmentation of words and characters become complex problem. Word segmentation is a non-trivial task, as there is not much difference between inter-word and intra-word vertical gap. It is very difficult to judge if the two adjacent ligatures belong to same or different words.

2.3.3 Feature Extraction

Features are the information passed to the recognizer, such as pixels, shape data, or mathematical properties. The extraction of the features, which are used for the recognition process, is a difficult task with two aims: first, to identify which features are relevant, and secondly, to find all of them. After testing several different features we finally chose the grey valued pixels of the normalized word image as features. Each letter consists of unique features and due to these exclusive properties it appears different from other characters. The goal of feature extraction is to extract accurate features from the characters, symbols, and words. Almost all the features can be grouped into two types, known as statistical and structural. Basically statistical features are derived from statistical distributions of points, such as moments, zoning, histograms, or projection features, on the other hand structural features focus on geometrical details of character which consists of holes, line ends, opening to (right, left, up or down), number of strokes, end points, joints, direction of stroke, extreme points branch points, etc. Many researchers investigated for different statistical and
structural features which are frequently preferred for handwritten numeral/character recognition [12]. Nafiz Arica and Fatos T. Yarman-Vural [6] grouped the features extraction techniques following categories

- Global Transformation & series expansion
  - Fourier transformation
  - Gabor Transformation
  - Wavelets
  - Moments
- Statistical representation
  - Zoning
  - Crossing & distances
  - Projections
- Geometrical & topological representation
  - Extracting & counting topological structure
  - Measuring & Approximating topological structure
  - Coding
  - Graph & Trees

2.3.4 Training and Recognition

After extracting the unique features of each character now the next step is to assign this character to particular class of character which has similar features. There are various techniques available for assigning the character into specific class of characters, this assign an unknown sample into predefined class. As summarized in review [6, 8] the recognition techniques can be grouped into

a) Template matching
b) Statistical techniques
c) Structural techniques
d) Neural networks

The matching technique determines the degree of similarity between two vectors like group of pixels, shapes, curvature etc. matching techniques can be categorized into Direct Matching, Deformable Templates, Elastic Matching and, Relational Matching. 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. Statistical techniques can be grouped into Non Parametric Recognition, Parametric Recognition, Clustering Analysis, Hidden Markov Modeling HMM, and Fuzzy Set Reasoning
In case of Structural Techniques unique structural features are considered, 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 structural pattern recognition. These patterns are used to describe and classify the characters in the CR 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. ANN is defined as a computing architecture that consists of a massively parallel interconnection of adaptive “neural” processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. An NN 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.

2.3.5 Post Processing

Post processing is the final stage of character recognition which uses the information of all above steps and converts it into output of the process which can be useful as an editable form. This step can improve the recognition accuracy by making use of additional tools such as spell-checkers. Urdu is derived from Arabic script and similarity is found in case of various letters, in case of Arabic, the pre-fixes and suffixes may have to be removed before the words are looked up in a lexicon [13]. Preprocessing cleans document in a specific way to remove the noise and it may remove significant information, since the context information is not available at this stage. Unavailability of context information during the segmentation stage may cause even more severe and irreversible errors since it yields meaningless segmentation boundaries. It is clear that if the semantic information were available to a certain extent, it would contribute a lot to the accuracy of the CR stages. On the other hand, the entire CR problem is for determining the context of the document image. Therefore, utilization of the context information in the CR problem creates a chicken and egg problem. The review of the recent CR research indicates minor improvements when only shape recognition of the character is considered. Therefore, the incorporation of context and shape information in all the stages of CR systems is necessary for meaningful improvements in recognition rates.
This is done in the post processing stage with a feedback to the early stages of CR. Utilization of a dictionary for correcting the minor mistakes of the CR systems. In some applications, the context information confirms the recognition results of the different parts in the document image. In automatic reading of bank checks, the inconsistencies between the legal and the courtesy amount can be detected, and the recognition errors can be potentially corrected.

Reference:
5. Introduction to optical character recognition available at web portal http://www.srimca.edu.in/Srijan/Srijanjan2011/IT/IntroductionToOCR.html


