Feature extraction is one of the most significant steps in any character recognition system because the classifiers are trained using features. Sample of every character can be expressed as set of quantifiable attributes known as features [1]. During feature extraction the most relevant information is extracted from raw data drawn from character image. The features are considered which minimizes the within-class pattern variability and enhance between-class pattern variability [2]. A range of feature extraction methods can be found in literature for various scripts. Each feature extraction technique has its own merits and demerits. To achieve the higher recognition accuracy, first a range of suitable features are needed which should be unique and invariant with respect to the shape variations in a character. Specifically, in handwritten character recognition variation in writing styles and character shapes are universal type of complexity for all scripts. In such case the features must be unique and invariant against all possible variations in handwriting. [3]
6.1 **Feature Extraction Methods**

The key objective of feature extraction is extracting unique feature set, which maximizes the recognition rate with the least amount of elements. All Feature extraction methods can be grouped into following three categories, in other words, all these methods are based on following three types of features [4,5].

- Structural
- Statistical
- Global transformations and moments

In literature many researchers [6] categorized feature extraction methods into two types i.e. statistical and structural. Figure 6.1 illustrate hierarchy of various types of features.

![Figure 6.1 Categories of Feature Extraction Methods](image)

6.1.1 **Structural Features:**

Structural features are nothing but various topological and geometrical properties of the character for example loop, cross points, aspect ratio, strokes and their directions, branch points, horizontal curves at top or bottom, inflection between two points, end points, strokes and bays in various directions, loops and stroke relations, intersections of line segments etc. It captures information related with shape of the characters and should be selected by considering all the shape variations. Also the selected structural
description should not affect feature set at large scale. Different global as well as local features of characters image can be represented through geometrical and topological features. These features are high tolerance characteristic against distortions and style variations. This feature extraction method may also require prior knowledge regarding structure of character. Hundreds of topological and geometrical representations can be grouped into Extracting and Counting Topological Structures, Measuring and Approximating the Geometrical Properties, Coding, Graphs and Trees[5].

6.1.2 **Statistical Features:**

The Statistical feature extraction method basically uses some quantitative measurement. Features based on statistical distribution of points also take care of style variations to some extent. Even though this method does not allow the reconstruction of the original image again, still it is used for reducing the dimension of the feature set providing high speed and low complexity. The major statistical features used for character representation are:

- Zoning
- Projections and profiles
- Crossings and distances

6.1.3 **Global transformations and moments:**

In an overview of offline character recognition Nafiz et. al. [5] enlightens the global transformation and moments based representation and clarify that “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/and series expansion. Deformations like translation and rotations are invariant under global transformation and series expansion”. Common transform and series expansion methods used in the CR are:

- Fourier Transforms
- Gabor Transform
- Wavelets
- Moments
- Karhunen–Loeve Expansion
6.2 Handwritten Urdu Text Feature Extraction - Challenges

Review of various Urdu character recognition theories is given in third chapter i.e. State of Art in Urdu OCR. Most of the available feature extraction strategies were discussed for both printed and handwritten Urdu text from literature. Feature extraction of handwritten text is more complex as compared with printed Urdu text. The reason is printed text may consist constrained structural features. For example in a font and size constrained environment feature extraction is easy, while in handwritten text the probability of variation and uncertainty in appearance increases.

There are different type of variations in writing Urdu Characters specially while drawing secondary components; mostly in drawing two dots and three dots [7]. As shown in Figure 6.2 Samples A1, B1, and F4 secondary three dots are written with three variations. Broken characters is another challenges in character recognition, as shown in E5 and F5. 'Do-Chasm-Hey' (⿷) is written with two loops and with one loop respectively. Also in A2 and A3 characters 'Wow' (٫) is written with loop and without loop respectively which changes the structural features and hence, feature extraction become tricky.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>❋</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>❋</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.2 Urdu handwriting variation problems*

The presence of dots and diacritic marks in the Urdu characters also creates a lot of confusion. The dots often get merged with each other or touch the ligatures or adjacent diacritic marks, which results in distorted shapes. As the letters are usually written in thick style and the special characters are usually written in small size, many times the holes of the special characters or normal characters get filled or the parallel lines get merged, resulting in distorted shapes, as shown in Figure 6.3.
6.3 **Ghost Character Theory:**

The Characters of Arabic script based language (like Urdu) are usually written using two types of strokes primary strokes and secondary strokes. The primary strokes are the main body of the characters and secondary strokes further describe the characters. Figure 6.4 shows primary and secondary components of character Pay (١). In which  is the primary stroke or main body of the characters and three dots (٠) are the secondary components of character.

![Main Body (Ghost)](image)

![Secondary Components](image)

*Figure 6.4 Urdu character with primary and secondary strokes*

If Secondary strokes of all the characters are removed then, remaining primary strokes along with single strokes characters can be called as ghost characters. Attash Durrani [8, 9, 10] is the theorist of Ghost Character of Arabic set. According to his proposed theory ‘All Arabic script based languages like Urdu, Persian, Sindhi, Punjabi, Pashto, Blochi, etc can be written with only 44 ghost characters. The Ghost character consists of 22 basic shapes known as Kashti and 22 dot or diacritical marks. Figure 6.5 shows the basic ghost characters used in Arabic script based languages. The primary shapes are usually known as ghost character whereas secondary strokes related with each character are known as diacritical marks. These 44 ghost characters are common can be used for all Arabic script based language therefore no need of different font for different languages’.

![Ghost characters useful for all Arabic script based languages](image)

*Figure 6.5 Ghost characters useful for all Arabic script based languages [11]*
This theory is also preferred by many researchers for large Multilanguage recognition of Arabic script based languages. Significant attainment is that Ghost Characters theory was partially included in UNICODE i.e. the dot less character set was completed by including dot less Bay, Fey and Quaff in the UNICODE Version 3.1. Later Ghost Characters theory has accepted and all the proposed characters were given 08 places on UNICODE. In any Arabic script based languages primary stroke contains zero or more diacritical. Besides the character contains up to four secondary stroke i.e. diacritical marks and can be increased up to six by adding zer, zaber etc. As mentioned earlier that these secondary strokes can be present below, middle or above of the primary ghost characters. [12]
The current research work is based on ghost-like character theory where Primary and Secondary strokes i.e. ghost characters and secondary strokes are separated using connected component analysis method. The diacritical marks are cropped from character image and ghost characters are extracted. The main body was identified through largest connected component region. As primary stroke is usually the largest component so connected component with largest region area property is cropped as primary or ghost character. As Urdu character shapes changes as per the change in their position i.e. isolated, initial, middle, end. That’s why, in current research work 135 Urdu characters are considered which includes all possible four shapes of characters. For all 135 characters recognition and classification process may be complex task. It needs extra efforts and processing capacity.

To normalize the recognition process and improve system accuracy, in current research work, ghost like characters (i.e. only primary strokes) are preferred. As mentioned earlier every ghost-like character is extracted by removing the secondary components. Surprisingly these 135 characters are reduced up to 60 ghost like characters (Primary Strokes) consists distinct main body shapes. Figure 6.6 show the derivation of 60 characters from 135 Urdu characters. It saves required efforts during classification and recognition process and also reduces complexity and increase recognition rate Urdu characters. For Example if any character from is to be recognized then it can be easily recognized with the ghost character i.e. . To achieve this objective first remove secondary stroke i.e. diacritical marks i.e. . After that only ghost character i.e. will remain in image. Figure 6.7 illustrate it in detail. Using this method five characters can be clustered or grouped based on ghost character i.e. .
Ghost character recognition theory can be implemented using four basic steps

- Initially separate the primary strokes or diacritical marks from the characters as shown in Figure 6.7. Afterwards ghost character and secondary strokes with their details like (position, type of secondary stroke) should be stored separately.
- Ghost character (Primary stroke) is recognized using feature extraction and classification.
- Similarly diacritical marks (Primary stroke) are recognized using feature extraction and classification method.
- Use the stored information about secondary character (type and positions above, below or middle) to recognize the actual Urdu character shown in Figure 6.8.

For every character, once the ghost character is recognized using classifiers, stored information of secondary component like type and position of secondary stroke is used to recognize the actual character. Figure 6.8 illustrate this process, where type of secondary strokes \( \text{\&} \) and their position (above/below) are used in combination to recognize the characters \( \text{\&} \) and \( \text{\&} \). This method saves the time and efforts required for recognition and improve accuracy of system.
6.4 Moment Invariant

6.4.1 Invariants:

Recognition of patterns and objects became significant attention of researchers during recent decades. Jan Flusser [13] explained following the three basic approaches to this problem

- Brute force
- Image normalization
- Invariant features

Brute Force:

In such approaches, parametric space of every possible image-degradation is searched. Here the training set for every class not only contain all class representatives but also their various versions are considered, for example their scaled, rotated, deformed and blurred versions. This approach is extremely complex and requires huge time, and thus it is considered as practically inapplicable by researchers.

Figure 6.8 Secondary stroke based decision Making (recognition) after feature extraction

6.4 Moment Invariant
**Image normalization:**

In this approach the objects or images to be recognized are transformed into a standard position and normalized prior to classify. This method works better than brute force and efficient for classification stage. But major difficulty is normalization of object against various ill-conditioned objects.

**Invariant features:**

It is extensively used approach and most promising for invariant feature extraction. The basic concept behind this method is to describe the objects through a set of measurable quantities known as invariants. These invariants are insensitive to specific deformations and also provide enough discrimination power for distinguishing an object which belongs to various classes. From a mathematical point of view, invariant I is a functional defined on the space of all admissible image functions that does not change its value under degradation operator D, i.e. that satisfies the condition $I(f) = I(D(f))$ for any image function $f$. This property is called invariance. In practice, in order to accommodate the influence of imperfect segmentation, intra-class variability and noise, we usually formulate this requirement as a weaker constraint: $I(f)$ should not be significantly different from $I(D(f))$. Another desirable property of I, as important as invariance, is discriminability. For objects belonging to different classes, I must have significantly different values. Clearly, these two requirements are antagonistic – the broader the invariance, the less discrimination power and vice versa. Choosing a proper tradeoff between invariance and discrimination power is a very important task in feature based object recognition (see Figure 6.9 for an example of a desired situation). Usually, one invariant does not provide enough discrimination power and several invariants $I_1 \ldots I_n$ must be used simultaneously. Then, we speak about an invariant vector. In this way, each object is represented by a point in an $n$-dimensional metric space called feature space or invariant space.
Figure 6.9 Two-dimensional feature space with two classes, almost an ideal example. Each class forms a compact cluster (the features are invariant) and the clusters are well separated (the features are discriminative)[13].

6.4.2 Categories of invariant

An invariant features which can be used to describe the 2D objects can be grouped according to different points of view. First simple method for categorization is according to the type of invariance. Where translation, rotation, elastic geometric, affine, projective, and scaling invariants are recognize. The Radiometric invariants exist with respect to linear contrast stretching, nonlinear intensity transforms, and to convolution. Categorization according to the mathematical tools used may be as follows:

• Simple shape descriptors – compactness, convexity, elongation, etc.
• transform coefficient features are calculated from a certain transform of the image – Fourier descriptors, Hadamard descriptors, Radon transform coefficients, and Wavelet-based features;
• Point set invariants use positions of dominant points;
• Differential invariants employ derivatives of the object boundary;
• Moment invariants are special functions of image moments.
One more viewpoint reflects what part of the object is needed to calculate the invariant. According to this invariants can be categorized into Global invariants, Local Invariants and Semi local invariants. The Global invariants can be calculated using the whole image including background in case where segmentation is not performed. On the other hands the Local invariants are calculated through a certain neighborhood of dominant points only. Differential invariants are typical representatives of this category. The Semilocal invariants attempt to retain the positive properties of the two groups above and to avoid the negative ones. They divide the object into stable parts and describe each part by some kind of global invariant. The whole object is then characterized by a string of vectors of invariants and recognition under occlusion is performed by maximum substring matching.

6.4.3 \textbf{Moments:}

The moments can be defined as scalar quantities used to characterize a function and to capture its significant features. The moments have been frequently used by researchers for in statistics for description of the shape of a probability density function and in classic rigid-body mechanics to measure the mass distribution.
Feature Extraction

of a body. Mathematically the moments are “projections” of a function onto a polynomial. Utilization of moments intended to object characterization in both invariant and non-invariant tasks has received significant attention during recent decades. The principal techniques explored include Moment Invariants, Geometric Moments, Rotational Moments, Orthogonal Moments, and Complex Moments. Various forms of moment descriptors have been extensively employed as pattern features in scene recognition, registration, object matching as well as data compression [14].

6.4.4 Moment Invariants:

Moment invariants often used as an important feature group in image processing, shape recognition, remote sensing, etc. The Moments can give characteristics of an object which are unique shape features. The Invariant shape recognition can be achieved through classification in the multidimensional moment invariant feature space. Hu was the pioneer who first set out the mathematical foundation for 2-dimensional moment invariants also demonstrated their use in shape recognition. These values are invariant to translation, scale and rotation of the shape. Hu defines seven of these shape descriptor values computed from central moments through order three that are independent to object translation, scale and orientation. Translation invariance is done through computing moments which are normalized with respect to the centre of gravity so that the centre of mass of the distribution is at the origin i.e. central moments. Size invariant moments are derived from algebraic invariants but these can be shown to be the result of simple size normalization. From the second and third order values of the normalized central moments a set of seven invariant moments can be computed which are independent of rotation [15].

Moment invariant is one of the results promising method, numerous papers have been presented to various improvements, extensions and generalizations of moment invariants and also to their use in many areas of application. These seven features are most important and most frequently used shape descriptors. The output of moment invariant is independent from translation, scale, rotation or mirror image of a particular character. Figure 6.11 shows the various transformation forms of Urdu character ‘Cheem’, for all these images Hu’s seven functions equation set gives near about similar numerical results.
6.4.5 Theory

Conventionally, moment invariants are calculated using the information of both the shape boundary and its interior region [Hu, 1962]. The moments which are used for constructing moment invariants can be defined in the continuous and for practical implementation purpose they are computed in the discrete form. Given a function \( f(x,y) \), these regular moments are defined by:

\[
M_{pq} = \int \int x^p y^q f(x, y) \, dx \, dy \quad (1)
\]

Where \( M_{pq} \) is two-dimensional Geometric Moments of order \( p + q \) of density distribution \( f(x,y) \). \( p \) and \( q \) are both natural numbers. If an image is represented by a discrete function then the integrals are replaced by summations. In this way the equation (1) can be rewrite as follows,

\[
M_{pq} = \sum_x \sum_y x^p y^q f(x, y) \quad (2)
\]

To normalize for translation in the image plane, the image centroids are used to define the central moments. The central moments for a digital character image \( f(x,y) \) are inherently translation independent, and central moments can be defined in their discrete representation as:

\[
\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q \, f(x, y) \quad (3)
\]
Where \( \bar{x} = \frac{m_{10}}{m_{00}} \quad \bar{y} = \frac{m_{01}}{m_{00}} \)

As under the translation of coordinates, the central moments do not change

\[
x' = x + \alpha \quad , \quad y' = y + \beta \quad \text{where } \alpha \text{ and } \beta \text{ are constants}
\]

Central moment’s translation invariant theorem says that, ‘The central moments are invariants under translation’. The summary of central moment of order up to 3 is expressed using following equations.

\[
\begin{align*}
\mu_{00} &= m_{00} \\
\mu_{10} &= 0 \\
\mu_{01} &= 0 \\
\mu_{11} &= m_{11} - \bar{y}m_{10} \\
\mu_{20} &= m_{20} - \bar{x}m_{10} \\
\mu_{02} &= m_{02} - \bar{y}m_{01} \\
\mu_{30} &= m_{30} - 3\bar{x}m_{20} + \bar{x}^2m_{10} \\
\mu_{03} &= m_{03} - 3\bar{y}m_{02} + 2\bar{y}^2m_{01} \\
\mu_{21} &= m_{21} - 2\bar{x}m_{11} - \bar{y}m_{20} + 2\bar{x}^2m_{01} \\
\mu_{12} &= m_{12} - 2\bar{y}m_{11} - \bar{x}m_{02} + 2\bar{y}^2m_{10}
\end{align*}
\]

Under the similitude transformation, i.e. the change of size,

\[
\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \quad \alpha \text{ – Constant} \quad (5)
\]

The normalized central moments, denoted \( \eta_{pq} \), are defined as

\[
\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}}
\]

Where the normalization factor \( \gamma = \frac{p+q}{2} + 1 \). For \( p + q = 2,3,... \)

Hu defines the a set of following seven moment invariants functions (\( \phi_1 - \phi_7 \)), computed from central moments through order three, that are invariant with respect to object scale, translation and rotation.

---

125
In which six are absolute orthogonal invariants \((\phi_1 - \phi_7)\) and one is skew orthogonal invariants \(\phi_7\) [16]. The functions \(\phi_1\) through \(\phi_6\) are invariant with respect to rotation and reflection, while \(\phi_7\) changes sign under reflection. The skew invariant \(\phi_7\) is useful in distinguishing mirror images.

### 6.5 Our approach:

In this research work Moment Invariants (MI) are used to evaluate seven distributed parameters for handwritten Urdu characters. In any character recognition systems the characters are proposed to extract features that uniquely represent properties of the character [17]. The MIs are well known to be invariant under translation, rotation, scaling and reflection. They are measures of the pixel distribution around the center of gravity of the character and allow capturing the global character shape information. In the present work, the moment invariants are evaluated using central moment of the image function \(f(x,y)\) up to third order[18]. Urdu Characters are grouped into single component and multi component characters.

Initially it is verified whether character consists of single component or more than one component. If character is single component then image is normalized into 60 X 60 and divided into three horizontal zones for features extraction as shown in Figure 6.12. From each zone 7 Moment Invariant features and from whole image 7 Moment Invariant features were computed, in this manner 28 Moment Invariant features were determined extracted as shown in Figure 6.13. Afterwards SVM is used
for classification and character is put into appropriate class and finally recognized the single stroke component character.

Figure 6.12 Single-stroke Character ['Meem'] directly divided into three zones

![Diagram](image)

Figure 6.13 Training Phase of Single Stroke (Components) Characters like ['Meem'].

Apart from single stroke character, Urdu script consists of the characters which need more than one stroke or components to be completed. For example Urdu character [ڏ] is a multistroke character, to write this character four strokes or components are required. The first stroke is called primary stroke i.e. ¯ and remaining three dots are known as secondary stroke i.e. ۔. If Character belongs to multi component then initially primary and secondary component are separated using connected component analysis method and both components are stored on different locations as shown in Figure 6.14.
Once the primary and secondary components are separated using connected component analysis method then primary stroke is divided into three horizontal zones. Similarly secondary components are divided into two horizontal zones. Figure 6.15 shows the zoning of primary into three zone and secondary components into two horizontal zones. Usually zoning method is used to improve accuracy of the system. In literature many researchers preferred zoning method to improve the efficiency by extracting more unique features. If we are passing whole image it will extraction limited features of the characters. But if we divide the image into multiple zones then extra unique features can be extracted for each zone. For this purpose the separated image of primary stroke is normalized into 60 X 60 by maintain height and width ratio [19] and further divided into 3 zones. The secondary component is normalized into 22 X 22 and divided into 2 horizontal zones as shown in Figure 6.15. Unlike whole character image, each zone may have unique features.

After zoning of primary components into three zones 7 Moment Invariant of complete image is extracted. Also 7 Moment Invariant features for each of three zones are extracted individually. In this way total 28 features of primary stroke are
extracted. And system is trained using Support Vector Machine as shown in Figure 6.16.

Figure 6.16 Training Phase of Primary Components (e.g. example of character cheem 🧘‍♂️)

Figure 6.17 Training Phase of Secondary Components (e.g. example 🧘‍♀️ of character Cheem 🧘‍♂️)

Similarly separated secondary strokes or components are zoned into two horizontal zones and 7 Moment Invariant of complete image are extracted. Beside 7
features for each of two zones of secondary stroke are extracted. Using this method total 21 features of secondary stroke are extracted system is trained using Support vector machine as illustrated in Figure 6.17. In this way instead of recognizing complete character as whole, both primary and secondary strokes are recognized individually by separating them.

After separating character into primary and secondary we have two alternatives. First is counting number of secondary strokes and their types like (aviors) with one dot below, pay (aviors) with three dots below, say (aviors) three dot above and Tate (aviors) one toe above. But it may give incorrect results in case of handwritten character recognition due to variation in writing secondary components. For example characters Pay (aviors) should consists of a primary component (i.e.aviors) and secondary components (i.e. three dots below). But in case of handwritten characters the structure of secondary components varies from writer to writer. The variation in writing Character Pay (aviors) is shown in Figure 6.18 – D4, E4 and F4 are which should have three dots below primary component (aviors) but E4 consist one dot below to a zigzag, and F4 have (aviors) as secondary component.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.18 Handwriting variations in Urdu Script

In handwritten Urdu character secondary strokes can have high variations for same shape Figure 6.19 shows possible shapes of secondary strokes. Counting number of secondary strokes and their types may give unsatisfactory results in case of handwritten characters. That why Moment Invariant of secondary component are considered as a feature instead of counting number of secondary components and their types. Instead of that segmented secondary components were normalized into 22 X 22. Then normalized image is divided into two horizontal zones and from each 7 Moment Invariant features and from whole image 7 Moment Invariant features were
computed, in this manner 21 features were determined and SVM is used to classify secondary stroke.

Figure 6.19 Possible Primary Components

Figure 6.20 System Architecture for Feature Extraction of Handwritten Urdu Character

The position of secondary component is one of the important features in Urdu character recognition. In case primary stroke of many characters are similar, then type and position of secondary character helps to recognize them. For example the in primary stroke of all these character is same i.e. ( ). In this case type of secondary stroke and their positions above or below are useful to distinguish them from each other. Due to this importance of position of secondary strokes, with
Invariant Moment of Secondary component we have also considered their positions i.e. *Above, Below or middle*. Based on 28 features of primary component and 21 features of secondary component they were recognized. But final decision making is conceded with the help information regarding position of secondary components and the actual character is recognized. The overall system architecture of our system is shown in Figure 6.20.

**Reference:**