WRITER IDENTIFICATION FOR DEVANAGARI SCRIPT

Thesis Submitted in Partial Fulfillment of Requirements for the Award of the Degree of

Master of Technology in

COMPUTER SCIENCE & ENGINEERING

SUBMITTED BY

TARUN NABIYAL

(2K15/CSE/19)

UNDER THE GUIDANCE

OF

MANOJ KUMAR

ASSOCIATE PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

BAWANA ROAD, DELHI-110042

(2015-2017)

CERTIFICATE

This is to certify that **Mr. TARUN NABIYAL** (2K15/CSE/17) has carried out the major project titled **"Writer identification for Devanagari Script"** as a partial requirement for the award of **Master of Technology** degree in **Computer Science & Engineering** by **Delhi Technological University, Delhi**. The Major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2015-2017. The Matter contained in this thesis has not been submitted elsewhere for the award of any other degree.

Date:

(Project guide)
Mr. Manoj Kumar
Associate Professor
Department of Computer
Science and Engineering
Delhi Technological University

ACKNOWLEDGEMENT

I express my gratitude to my major project guide **Mr. Manoj Kumar, Associate Professor in Department of Computer Science and Engineering at Delhi Technological University, Delhi** for the valuable support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his constructive criticism, interminable encouragement and valuable insight without which the project would not have shaped as it has. I humbly extend my word of gratitude to Dr. Kapil Sharma, Head of Department and all the other faculty members and staff of department for providing their valuable help, time and facilities at the need of hour.

Tarun Nabiyal

Roll No.: 2K15/CSE/19

ABSTRACT

Writer identification is the task of formative the person whose handwritten sample is available in a set of writings, collected from multitude of writers. This has useful applications in many areas, conspicuously in forensic analysis. The task of writer identification is quite difficult due to marginal variations found in different handwritten samples from same person/writer. Several identification algorithms have been recommended so far which are mostly for non-Indic writings.

Research into writer identification has been focused on two streams, off-line and on-line writer identification. Generally it is believed that text-independent writer identification is more thought-provoking than text-dependent writer identification. Text-independent Offline writer recognition is more stimulating than online writer recognition. Here we propose a system which extracts the simple writer specific features from the scanned handwritten documents by different writers and use them to recognize the writer. Based on the idea that has been presented in the previous studies, here we assume handwriting as texture image and a set of features which are based on multi-channel Gabor filters and GLCM (Gray Level Co-occurrence Matrix) are extracted from preprocessed image of documents. First order statistical features are also extracted from the documentation. Substantially, the property of proposed method is using of the bank of Gabor filters which is appropriate for structure of Devanagari handwritten texts and vision system.

As there is no predefined dataset of Devanagari Handwritten reports, so for this we first made our own particular database. The database comprises of pictures of composing of 45 authors. Each author composes five same constituent in five pages, it makes an aggregate of 225 records and every archive contains 102 words. Fifty percent of the aggregate authors of the database are female and the remaining authors are male with the age gathering changing from 21 years to 60 years. Out of 225 pictures, 180 content pictures are utilized for training and rest is utilized for testing. We have evaluated features from the 2-D Gabor filter, GLCM and first order statistical method and the classifier used for identification of writer is k-nearest neighbor (k-NN). The above approach is tested for our database and experimental results are encouraging.

CONTENTS

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
FDE	Forensic Document Examination
GLCM	Gray Level Co-occurrence Matrix
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
IDM	Inverse Difference Moment
K-NN	K- Nearest Neighbour
LCDF	Local Contour Distribution Function
PDA	Personal Digital Assistant
PDF	Probability Distribution Function
PNN	Probabilistic Neural Network
RGB	Red Green and Blue
SMR	Smart Meeting Room
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Notwithstanding the improvement of electronic document and forecasts of a paperless world, the significance of manually written reports has held its place and the issues of recognition and certification of the authors has been a dynamic zone of explore in the course of recent years. Contrasted with the electronic or impressed content, the manually written content conveys extra data about the identity of the individual who has composed. There exists a specific level of steadiness in the composition manner of a person which makes it conceivable to recognize the author for which one has as of now observed a composed content.

The demand to recognize the creator of a document is an intermittent issue that emerges regularly in the court of equity where the genuineness of an archive (e.g. a will) must be finished up [47]. It is postured in the domain of medical too where the medicine needs to originate from an approved individual [48] and in banks for the confirmation of signature [50]. It can likewise be utilized for the examination of old records because it has the capacity to do their ordering and recovery. We can utilize it for the acknowledgment of hand written message and exploiting the principle of adaptation of the scheme to the type of author [46]. Here, we are showing a framework for offline recognition of penmanship styles. The goal is to discover a list of likeness between the handwriting of an obscure author and the handwriting in the reference base, whose author is known. This is not quite the same as the verification framework where given 2 penmanship tests are s1 and s2, we have to decide whether the two specimens were composed by a similar individual or by two distinct individuals.

1.2 THE NATURE OF HANDWRITING

Hand composing is an aptitude that is particular to individuals. Key attributes of penmanship are three-overlay. It comprises of artificial graphical checks on a surface; its motivation is to impart something; this intent is fulfilled by merit of the mark‘s formal relation to language. Writing is believed to have made imaginable a lot of culture and human progress. Every script has an arrangement of symbols, which are known as characters or letters, which have sure fundamental patterns. There are principles for joining letters to speak to states of more elevated amount semantic units. For instance there are principles for consolidating the states of single letters in order to frame cursively composed words in Latin letter set.

1.3 ENDURANCE OF HANDWRITING

Copybooks and different written work techniques, similar to the Palmer strategy, penmanship examination, and signature gathering, are words that invoke a lost world in which individuals hoped to penmanship as both an example in congruity and a charm of the person. The reason that hand-composing perseveres in the age of the computerized PC is the comfort of paper and pen when contrasted with consoles for various everyday circumstances. Penmanship was created quite a while prior as a way to grow human memory and to encourage correspondence.

Toward the start of the new thousand years, innovation has at the end of the day conveyed penmanship to a junction, Nowadays, there are various approaches to extend human memory and additionally to encourage correspondence and in this point of view, one may ask: Will penmanship be debilitated with elimination, or will it enter a time of real development?

Penmanship has changed hugely after some time and, up until this point, every innovation push has added to its development. The printing press and opened up the world to designed reports, expanding the quantity of perusers that, in him, diagramd out how to compose and to convey. PC and correspondence innovations, for example, word processors, fax machines, and email are affecting proficiency and penmanship. More up to date advances, for example, individual computerized collaborators (PDAs) and advanced mobile phones will likewise have an effect.

Every one of these innovations have prompted the adjusting and reinterpreting of the part of penmanship and manually written messages. Each time, the specialty possessed by penmanship has turned out to be all the more unmistakably characterized and promoted. When in doubt, it appears that as the length of transcribed messages diminishes, the quantity of individuals utilizing penmanship increments. Across the board acknowledgment of advanced PCs apparently challenges the eventual fate of penmanship. In any case, in various circumstances, a pen together with paper or a little scratch pad is a great deal more helpful than a console. For instance, understudies in a school are still not writing on a note pad PC. They collect dialect, conditions, and diagrams with a pen. This average worldview has prompted the idea of pen registering where the console is a costly and non-ergonomic segment to be supplanted by a pen tip position delicate surface superimposed on a realistic show that produces electronic ink. A definitive penmanship PC should handle electronic penmanship in an unconstrained situation, manage many composition styles and dialects, work with discretionary client characterized letters in order, and see any written by hand message by any author.

1.4 WHAT ARE HANDWRITING RECOGNITION, INTERPRETATION, AND RECOGNITION?

A few sorts of examination, acknowledgment, and understanding can be related with penmanship. Penmanship acknowledgment is the undertaking of changing a dialect spoken to in its spatial type of graphical imprints into its typical portrayal. The characters of most composed dialects of the world are spoken to today as 16-bit Unicode. Penmanship understanding is the assignment of deciding the significance of an assortment of penmanship, e.g., a transcribed address. Hand-composing recognizable proof is the errand of deciding the author of an example from an arrangement of authors, accepting that every individual's penmanship is individualistic. Mark check is the errand of deciding if the mark is that of a given individual. Recognizable proof and confirmation which have applications in criminological investigation, are forms that decide the exceptional idea of the written work of a particular author, while penmanship acknowledgment and understanding are forms whose destinations are to sift through the varieties to decide the message. The errand of perusing penmanship is one including particular human aptitudes. Information of the subject domain is basic as, on account of the infamous doctor's remedy, where a drug specialist utilizes learning of medications.

1.5 HANDWRITING AS A BIOMETRIC TOOL

The distinguishing proof of a man on the premise of examined pictures of penmanship is a valuable biometric methodology with application in criminological and memorable archive examination and constitutes an excellent investigation territory inside the exploration domain of behavioral biometrics. Physiological biometrics (e.g., iris, unique mark, hand geometry, retinal veins, DNA) are solid modalities for individual distinguishing proof because of the diminished fluctuation and high unpredictability of the biometric formats utilized. Be that as it may, these physiological modalities are generally more intrusive and require collaborating subjects. In actuality, behavioral biometrics (e.g., voice, stride, eye stroke flow, signature, penmanship) are less intrusive, yet the achievable ID exactness is less amazing because of the substantial changeability of the conduct determined biometric formats. Author recognizable proof relates to the classification of behavioral biometrics and has relevance in the criminological and noteworthy record examination domains. Author distinguishing proof is established in the more seasoned and more extensive space of programmed penmanship acknowledgment. For programmed penmanship acknowledgment, invariant portrayals are looked for which are fit for taking out varieties between various penmanship so as to order the states of characters and words powerfully. The issue of author distinguishing proof, in actuality, requires a particular upgrade of these varieties, which are trademark to an author's hand. Penmanship acknowledgment and author distinguishing proof accordingly speak to two restricting aspects of penmanship examination. It is essential, notwithstanding, to likewise say author recognizable proof could help the acknowledgment procedure if data on the author's general composition propensities and peculiarities is accessible to the penmanship acknowledgment framework.

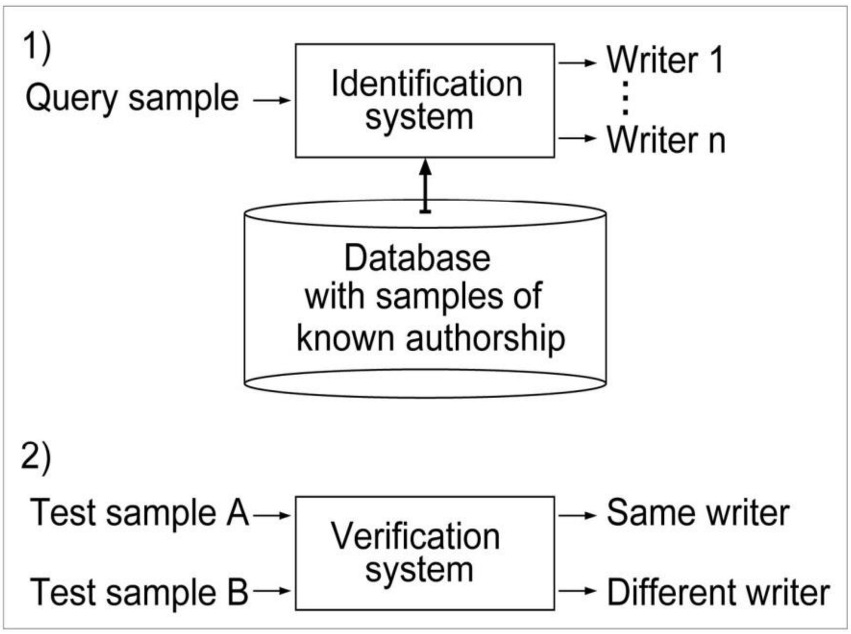


Fig 1.1 Writer Recognition and Verification Model [8]

Explore in author recognizable proof and confirmation has gotten huge enthusiasm for late years because of its scientific pertinence (e.g., the instance of the Bacillus anthrax letters). An author distinguishing proof framework plays out a one-to-many pursuit in a substantial database with penmanship tests of known origin and returns a conceivable rundown of competitors (diagram 1.1). This speaks to an exceptional instance of picture recovery, where the recovery procedure depends on highlights catching penmanship uniqueness. The hit list is additionally examined by the scientific master who settles on an ultimate conclusion in regards to the personality of the creator of the addressed specimen. Author recognition is in this way conceivable just if there exists past examples of penmanship by that individual enlisted in the measurable database. Author check includes a coordinated examination with a choice in the matter of regardless of whether the two examples are composed by a similar individual. The decidability of this issue gives understanding into the idea of penmanship uniqueness. Author confirmation has potential materialness in a situation in which a particular author must be naturally recognized in a flood of manually written reports. As opposed to different types of biometric individual recognizable proof utilized as a part of scientific labs, programmed author recognition frequently takes into consideration deciding character in conjunction with the purposeful parts of a wrongdoing, for example, on account of danger or payoff letters. This is an essential contrast from other biometric techniques, where the connection between the proof material and the subtle elements of an offense can be very remote.

1.6 APPLICATION OF AUTHOR RECOGNITION AND VERIFICATION

Author recognition and verification have applications in various domains. Especially the domain of biometric recognition has recently gained increased interest due to threats of terrorist attacks with the need to identify sure individuals. Furthermore, as more and more devices have built-in support for pen input, applications in this domain are likely to increase.

Forensic Document Examination: Forensic Document Examination (FDE) in-tends to clarify the author of a questioned handwritten document. Applications include the verification of signatures, e.g., the genuineness of a signature on a bank check, or the recognition of the author of a questioned document, e.g., the author of a threat or a ransom letter [49]. FDE has been largely based on manual examination by human experts and only recently automated methods have been introduced. This process has been aided by several rulings in the United States that question the admissibility of handwriting as evidence because of the lack of a scientific basis. In this case a semi- or full-automatic author recognition and verification system can provide an objective measurement independent of a human's judgement.

Biometric Recognition: Biometric acknowledgment means the programmed acknowledgment of people in view of their physiological or behavioral qualities. Physiological qualities depend on a physical characteristic of the human body (e.g., confront, unique mark, iris, DNA, and so forth.). Behavioral qualities utilize singular attributes of a man's conduct for recognizable proof (e.g., voice, walk, keystroke progression, signature, penmanship, and so on.).Because these characteristics are assumed to be unique to each person they are more reliable and more distinctive than knowledge-based.

Digital Libraries: Computerized libraries are an arrangement of electronic assets and related specialized abilities for making, looking, and utilizing data. In this sense they are an expansion and improvement of data stockpiling and recovery frameworks that control computerized information in any medium. The substance of computerized libraries incorporates information and metadata that describe various aspects of the data stored in a library. An interesting aspect of digital libraries is to preserve rare and ancient handwritten documents. The precious cultural heritage is preserved by digitization and then made available in digital form. Possible applications of author recognition and verification in this context are to retrieve historical documents who have not yet been assigned to one author or to validate the authorship of a document.

Smart Meeting Rooms: The point of a Smart Meeting Room (SMR) is to auto-mate standard errands for the most part performed by people in a meeting and to develop techniques that help to obtain required information from a meeting. To record a meeting, a SMR is equipped with synchronized audio and visual recording devices. Explore on smart meeting rooms seeks to develop methods to capture, store, structure, query, and browse the data acquired. One imperative undertaking in a SMR is to catch the penmanship rendered on a whiteboard amid a meeting. This assignment additionally incorporates distinguishing the author of a content composed on a whiteboard. Taking care of this issue empowers its clients to mark the penmanship with the author's personality.

Ambient Intelligence: Encompassing insight intends to make conditions that are touchy and receptive to the nearness of individuals. The thought is to enhance individuals' personal satisfaction by making the coveted atmosphere and by providing the desired functionality through intelligent and personalized systems and services of everyday life. Two key concepts of ambient intelligence are context awareness and personalization [49].

1.7 HANDWRITING ANALYSIS

Off-line handwriting data: Firstly, two best in class frameworks to address the errands of author ID and check are introduced and tentatively assessed. While the primary framework utilizes Hidden Markov Models (HMMs), the second framework depends on Gaussian Mixture Models (GMMs) to demonstrate a man's penmanship. To the best of the author's information neither HMMs nor GMMs have already been connected to disconnected author distinguishing proof and confirmation. Also, different component choice strategies are connected to enhance the execution of a current disconnected author distinguishing proof framework. The performance of the system is significantly improved by selecting a good sub-set of the original feature set. The methods discussed can potentially be applied to any set of features and are thus not restricted to off-line hand-writing. The diagram 1.2 shows the analysis of handwriting.

On-line handwriting data: In the on-line case, a GMM-based framework to distinguish the author of on-line written by hand notes on a whiteboard is introduced and experimentally evaluated. Furthermore, another way to deal with enhance the execution of an author recognizable proof framework by combining non-concurrent highlight streams is displayed. To the best of the writer information, this is the main framework to distinguish the author of on-line whiteboard information.

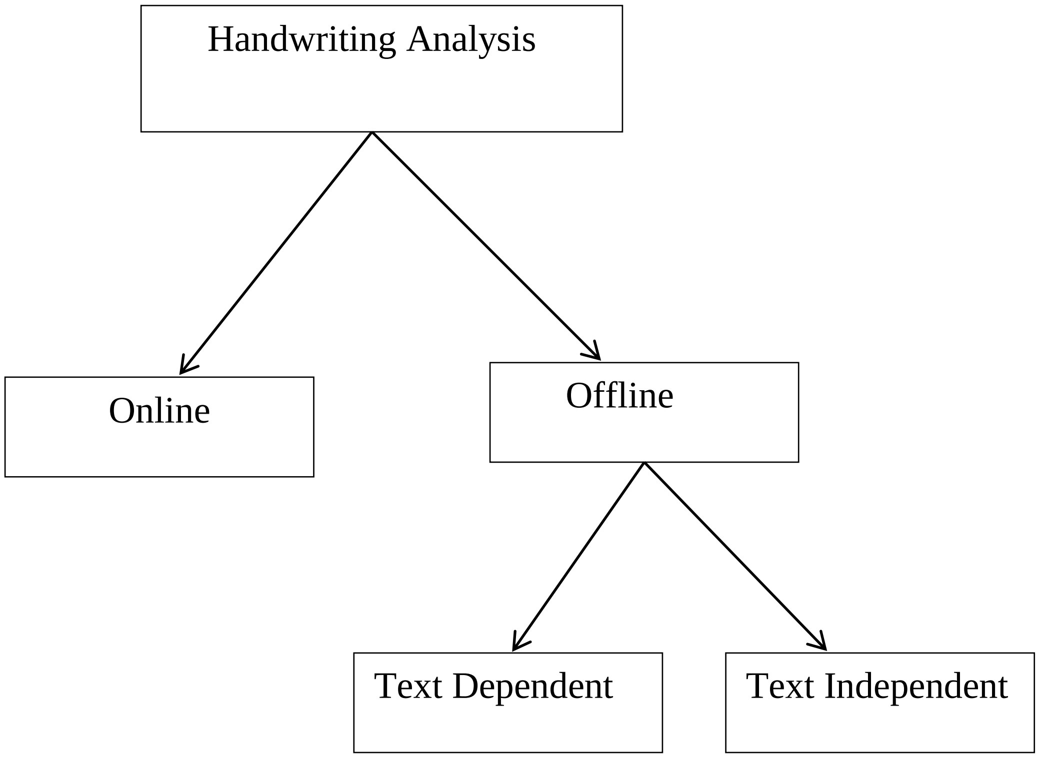


Fig. 1.2 Analysis of handwriting

Text dependent system: A text dependent system is tied to a pre-defined text that is used both for training and recognition of the author.

Text independent system: A text independent author recognition system has no prior knowledge of testing data.

Author recognition and confirmation are just conceivable to the degree that the variety in penmanship style between various journalists surpasses the varieties natural for each and every author considered in segregation. The outcomes revealed in this paper at last speak to a factual examination of the relationship contradicting the between-author changeability and the inside author inconstancy in include space. The present investigation accept that the penmanship was created utilizing a characteristic written work state of mind. Produced or camouflaged penmanship is not tended to in our technique. The counterfeiter tries to change the penmanship style, as a principle by changing the inclination and additionally the picked letter patterns. Utilizing nitty gritty manual examination, criminological specialists are now and then ready to accurately recognize a fashioned transcribed example. Then again, our proposed calculations work on the examined penmanship reliably, considering every graphical shape experienced in the picture under the start that they are made by the constant and normal script style of the author.

1.8 AUTHOR RECOGNITION ARCHITECTURE

Author recognition involves many steps to completely recognize the actual author of the documents. These stages are named as: Pre-preparing, Segmentation, classification and Feature extraction. The design of these stages is appeared in diagram 1.3 and these stages are recorded underneath with brief portrayal. The significant goal of the framework immature, is that it ought to be similarly material to all dialects. The elements gathered are content autonomous. We have utilized neural system for author distinguishing proof. In the accompanying segments, we would clarify in detail, each of the squares in the above framework.

Data Acquisition: First of all there is no standard/public datasets available for Devanagari Handwritten documents, so it is necessary to build our own datasets. Here we have taken documents of 45 author of different age group. It is necessary to build a datasets before designing the system. This process involves digitization of paper document collected from different sources using scanner.

Pre-processing: The pre-handling stage regularly incorporates numerous strategies connected for binarization, commotion expulsion, skew recognition, incline redress, standardization, form making, cushioning and skeletonization like procedures to make character picture simple to remove applicable elements and proficient acknowledgment.

Feature Extraction: Highlight extraction is utilized to extricate applicable elements for acknowledgment of characters in light of these elements. To begin with highlights are registered and extricated and after that most important elements are chosen to build include vector which is utilized in the end for acknowledgment and distinguishing proof. The calculation of elements depends on auxiliary, factual, directional, minute, change like methodologies**.** The features evaluated for author recognition of devanagari scripts are Gabor filter, GLCM features and first order statistical [33].

Classification: Each example having highlight vector is ordered in predefined classes utilizing classifiers. Classifiers are first prepared by a preparation set of example tests to set up a model which is later used to perceive the test tests. The preparation information should comprise of wide assortments of tests to perceive every single conceivable example amid testing. A few cases of for the most part honed classifiers are- Probabilistic Neural Network (PNN) [36],Artificial Neural Network (ANN) [24], K-Nearest Neighbor (K-NN) [37] Support Vector Machine (SVM) [21].

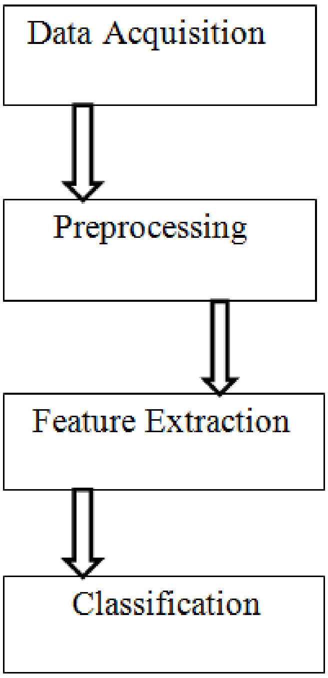


Figure 1.3 Architecture of Author Recognition

1.9 ABOUT DEVANAGARI SCRIPT

Even though a descendent of the Brahmi script, Devanagari has evolved into a highly cursive script. Many languages in India, such as Hindi and Sanskrit, use Devanagari and many more languages throughout India use local variants of this script.

Hindu scriptures are written in Devanagari, a fact illustrated by the etymology of the name. "Devanagari" is a compound word with two roots: *deva* means "deity", and *nagari* means "city". Together it implies a script that is both religious as well as urbane or sophisticated. A few examples of manually written Devanagari archives are appeared in Figure 1.4.

Prominent Features

- Case of composing framework: alphasyllabary/abugida.

- way of composing: horizontally left to right.
- Consonant letters convey an inborn vowel which can be adjusted or quieted by methods for diacritics or matra.
- When consonants happen together in bundles, one of a kind conjunct letters are applied.
- The request for of the letters relies on upon articulatory phonetics

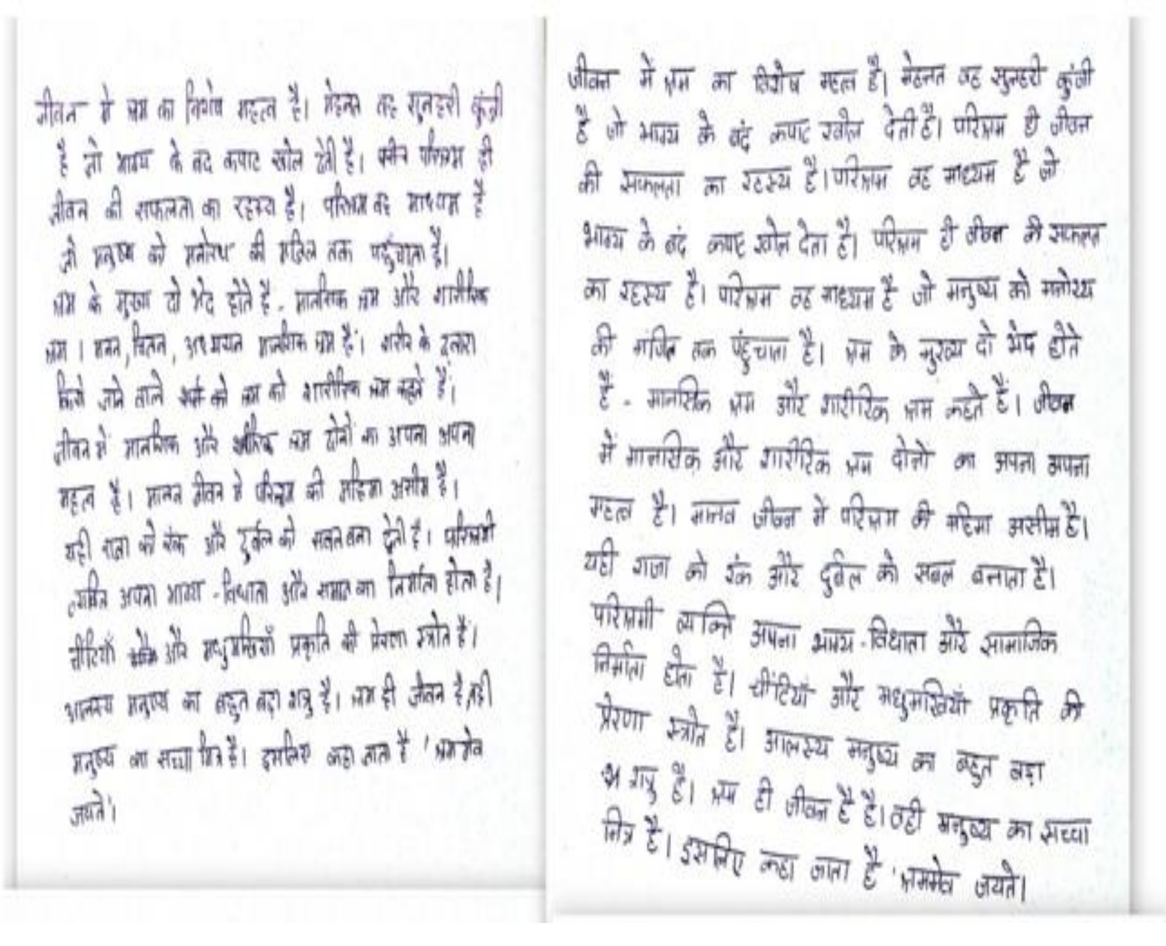


Fig. 1.4 Handwritten Devanagari text image

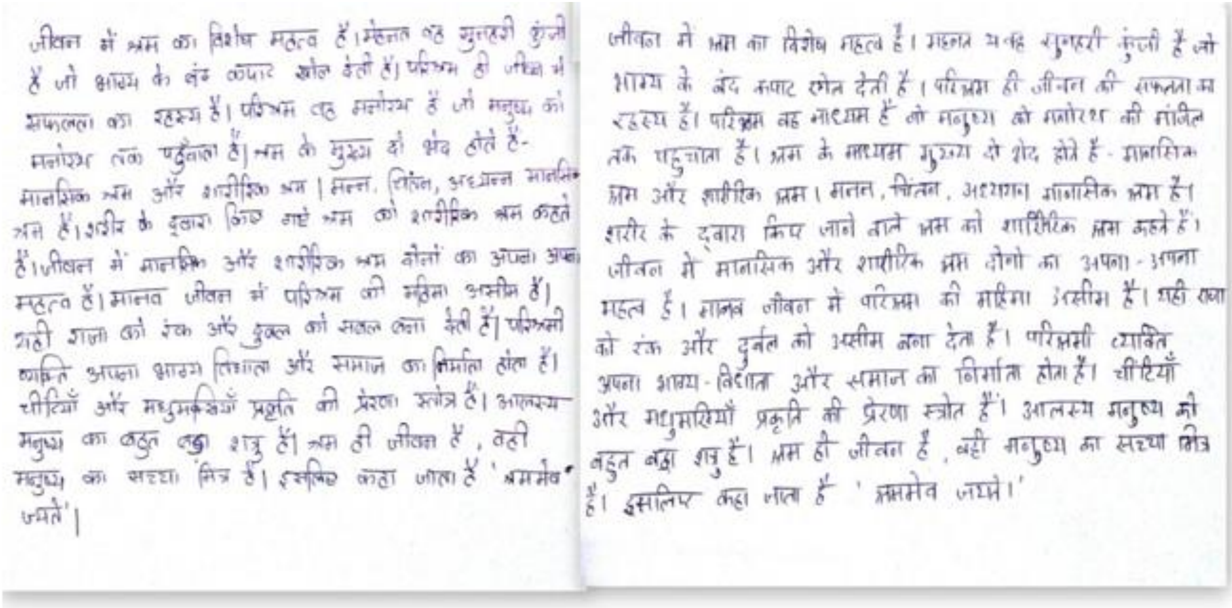


Figure 1.4 Handwritten Devanagari text image

- Vowels can be created as self-sufficient letters, or by using a various diacritical engravings which are formed above, underneath, prior or after the consonant they have a place with. This part is essential to most by far of the letters arranged by South and South East Asia.

1.10 PROPOSED WORK

In our proposed work we have identified the unknown author by evaluating the features from handwritten text. Here we evaluate features like first order statistical, Gabor filter and gray level co-occurrence matrix (GLCM) features from the whole document of the image. We divide the text document for testing and training. There are five documents of same content of one author is present. We have taken four text documents for training and one for testing. There are total 45 author’s documents available. The classifier used for classification is k-nearest neighbor (K-NN). The result we have found is quite good.

1.11 OUTLINES OF THESIS REPORT

This is an introductory chapter which gives an overview of generalized author recognition system, architecture of author recognition, used features and the overview of Devanagari script. Followed by outline of the proposed work and organization of the text in the respective order. Remaining text engineered as follows:

Chapter 2 Covers the literature survey related to earlier approaches adopted for author recognition, and the motivation behind the proposed work.

Chapter 3 Discuss the detail of our proposed work, step by step process of author recognition.

Chapter 4 Shows experimental results and analysis of proposed work.

Chapter 5 Covers conclusion and future scope.

CHAPTER 2

LITERATURE SURVEY

2.1 OVERVIEW

In this chapter literature survey regarding the previous work and related approaches about author recognition and verification is presented. The most of the related works on author recognition have done on English, Chinese and Arabic like scripts and languages. Author recognition and verification on Devanagari script is recent trends. The trends of Author Recognition on Devanagari script are recent. As work on related to Character Recognition is very famous in Devanagari script.

In our writing study we have considered many explore moved toward honed on numerous dialects and scripts like English, Chinese, Arabic and Bengali and Devanagari. Our accentuation is to ponder and break down the element extraction approaches, watched or announced outcomes and numerous other pertinent issues impressive to new explore chip away at Author Recognition. After our point by point writing study we could find the subject of our proposed work incorporating procedures we joined in different stages, to execute it and to assess our outcomes and conclusion.

2.2 WORK RELATED TO AUTHOR RECOGNITION

Srihari et al. [1] tested penmanship tests of 1500 people, illustrative of the U.S. populace as for sexual orientation, age, ethnic gatherings, and so on., were acquired. Breaking down contrasts in penmanship was finished by utilizing PC calculations for separating highlights from examined pictures of penmanship. Qualities normal for the penmanship were acquired, e.g., line partition, incline, character patterns, and so forth. These traits, which are a subset of characteristics utilized by criminological archive analysts (FDEs), were utilized to quantitatively set up uniqueness by utilizing machine learning approaches. Utilizing worldwide traits of penmanship and not very many characters in the composition, the capacity to decide the author with a high level of surety was built up. The work is a stage towards giving logical support to conceding penmanship prove in court. The numerical approach and the subsequent programming additionally have the guarantee of supporting the FDE.

Bulacu et al.[2] proposed edge-based directional likelihood appropriations as components in author recognizable proof in contrast with various non-rakish elements. It is noticed that the joint likelihood dissemination of the point mix of two "pivoted" edge sections beats all other individual components. Consolidating components may enhance the execution. Confinements of the strategy relate to the measure of manually written material required so as to get dependable circulation gauges. The worldwide elements treated in this examination are touchy to real style variety (upper-versus bring down case), incline, and manufactured styles, which requires the utilization of different elements in sensible scientific author distinguishing proof methodology.

Schomaker & Bulacu [3] implemented a new strategy for disconnected author recognizable proof is introduced, utilizing associated segment forms (COCOCOs or CO3s) in capitalized written by hand tests. In this model, the author is believed to be depicted by a stochastic case generator, making a gathering of related portions for the promoted character set. Using a codebook of CO3s from a free planning set of 100 authors, the probability thickness work (PDF) of CO3s was diagramd for a self-ruling test set containing 150 disguised exploreers. Results revealed a high-affectability of the CO3 PDF for perceiving particular exploreers on the introduce of a single sentence of promoted characters. The proposed customized approach traverses any hindrance between picture bits of knowledge approaches toward one side and physically measured allograph components of individual characters on the other side. Joining the CO3 PDF with a free edge-based presentation and shape PDF yielded high right ID rates.

Bulacu & Schomaker [4] proposed various new and extremely viable elements for programmed author distinguishing proof and confirmation. They are likelihood dissemination capacities (PDFs) separated from the penmanship pictures and describe author independence freely of the literary substance of the composed examples. In this paper, they play out a broad investigation of highlight mixes. In their combination conspire, the last exceptional separation between two transcribed examples is registered as the normal of the separations because of the individual components taking an interest in the mix. Acquired on a vast dataset containing 900 authors, our outcomes demonstrate that combining different components (directional, grapheme, run-length PDFs) yields expanded author recognizable proof and confirmation execution.

Bulacu & Schomaker [5] grown new and extremely compelling systems for programmed author recognition and confirmation that utilization probability distribution functions (PDFs) separated from the penmanship pictures to portray author singularity. A characterizing property of these techniques is that they are intended to be free of the printed substance of the written by hand tests. These strategies work at two levels of examination: the surface level and the character-shape (allograph) level. At the surface level, utilizes form based joint directional PDFs that encode introduction and ebb and flow data to give a cozy portrayal of individual penmanship style. In this investigation at the allograph level, the author is thought to be portrayed by a stochastic example generator of ink-follow sections, or graphemes. The PDF of these straightforward patterns in a given penmanship test is trademark for the author and is processed utilizing a typical shape codebook acquired by grapheme bunching. Consolidating numerous components (directional, grapheme, and run-length PDFs) yields expanded author distinguishing proof and confirmation execution. The proposed strategies are pertinent to free-form penmanship (both cursive and secluded) and have reasonable achievability, under the presumption that a couple of content lines of written by hand material are accessible keeping in mind the end goal to get dependable likelihood gauges.

Schomaker et al.[6] proposed a new algorithm for measurable or verifiable author ID, utilizing the forms of divided associated parts in free-form penmanship. The author is thought to be portrayed by a stochastic example generator, creating a group of character parts (fraglets). Utilizing a codebook of such fraglets from a free preparing set, the likelihood dispersion of fraglet forms was registered for an autonomous test set. Results uncovered a high affectability of the fraglet histogram in distinguishing singular authors on the premise of a passage of content. Expansive scale investigates the ideal size of Kohonen maps of fraglet forms were performed, demonstrating usable grouping rates inside a non-basic scope of Kohonen delineate. The proposed programmed approach overcomes any issues between picture measurements approaches and simply information based manual character-based techniques.

Siddiqi & Vincent [7] built up a nearby approach, in view of the extraction of qualities that are particular to an author. To misuse the presence of repetitive examples inside a penmanship, the written work is partitioned into a substantial number of little sub-pictures, and the sub-pictures that are morphologically comparable are assembled together in similar classes. The examples, which happen oftentimes for an author, are in this way removed. The creator of the obscure report is then recognized by a Bayesian classifier. The framework prepared and tried on 50 records of a similar number of creators, detailed a recognizable proof rate of 94%.

Yan et al. [8] implemented a phantom element extraction strategy in light of quick Fourier change for author distinguishing proof is exhibited. As per the constructability of the surface picture of penmanship, advanced an estimation strategy for numerical desire estimation of surface picture's unearthly elements. This strategy takes out the arbitrariness of ghastly components and gets steady phantom elements. The trial comes about demonstrate that this approach upgrades the recognizable proof precision to an expansive degree utilizing informational collections with substantial penmanship tests.

Shahabi & Rahmati [9] proposed another strategy for disconnected author distinguishing proof which depends on Farsi penmanship and content autonomous. In view of the possibility that has been displayed in the past examinations, here accept penmanship as surface picture and an arrangement of elements which depend on multi-channel Gabor channels are extricated from preprocessed picture of records. Generously, the property of proposed strategy is utilizing of the bank of Gabor channels which is fitting for structure of Farsi written by hand messages and vision framework. Additionally, another element extraction technique is proposed which depends on Gabor-vitality and minutes. For the main, we review distinctive techniques for highlight extraction from yield of Gabor channels. These strategies with co-event framework and Said strategy are executed and exploratory outcomes on penmanship of 40 people groups exhibit that the proposed technique accomplishes better execution on Farsi written by hand reports.

Garain & Paquet [10] proposed a two-dimensional (2D) autoregressive (AR) demonstrating method. Every author is spoken to by an arrangement of 2D AR show coefficients. A strategy to evaluate AR show coefficients is proposed. This strategy is connected to a picture of content composed by a particular author so that AR coefficients are gotten to portray the author. For a given example, AR coefficients are processed and its L2 separate with each of the put away (author) models distinguishes the author for the specimen. The technique has been tried on datasets of two unique scripts, to be specific RIMES containing 382 French journalists and ISI comprising of tests from 40 Bengali scholars. Demonstrating of composing styles utilizing distinctive setting designs at various picture determination has been examined. Test comes about demonstrate that the system accomplishes comes about similar with that of the past methodologies.

Chanda et al. [11] proposed a framework to experience such unfavorable circumstance with regards to Bengali script. Tests with discrete directional element and inclination include are accounted for here, alongside Support Vector Machine (SVM) as classifier. We got promising aftereffects of 95.19% author recognizable proof exactness at first top decision and 99.03% while considering initial three top decisions.

Biswas & Das [12] use two distinct arrangements of segments (basically sections of characters); in particular piece set-An and part set-B. Components are separated from every component of these two sets to distinguish the written work style of a specific individual. The elements are diagramd in view of Radon change projection profile. The proposed approach utilizes lesser measure of data from the transcribed examples; along these lines sparing calculation time and in addition memory prerequisite. The condition to verify that the author is obscure (i.e., there is no manually written specimen from that author in reference base) is additionally proposed. The approach is tried on a gathered dataset of Bangla compositions and the test comes about are empowering.

Ding et al. [13] uses neighborhood form appropriation highlights are proposed for author recognizable proof. The Local Contour Distribution Feature (LCDF) is removed from the pieces which are parts of the shape in sliding windows. Keeping in mind the end goal to decrease the effect of stroke weight, the parts which don't specifically associate the middle point are overlooked in the element deliberation method. The edge point dispersions of the parts are tallied and standardized into LCDFs. Finally, the weighted Manhattan remove is utilized as comparability estimation. The trials on ICDAR 2011 author distinguishing proof database demonstrate that the execution of the proposed technique reach or surpass those of existing condition of-craftsmanship strategies.

Halder et al. [14] gather 5 duplicates of transcribed characters to invalidate intra-composing variety, from 50 unique individuals primarily understudies. In the wake of preprocessing and character extraction, 64-dimensional element is diagramd in light of angle of the pictures. Some manual handling is required on the grounds that a few clamors are excessively troublesome, making it impossible to evacuate consequently as they are considerably nearer to the characters. They utilized LIBLINEAR and LIBSVM classifiers of WEKA condition to get the singularity of characters. We have done the author distinguishing proof with every one of the characters and got 99.12 % precision for LIBLINEAR with all scholars.

Kumar et al. [15] display a framework that uses those basic properties as elements that graphologists and master penmanship analyzers use for deciding the author's identity qualities and for making different appraisals. Each stroke trademark mirrors an identity characteristic. They have measured the viability of these elements on a subset of manually written Devanagari and Latin script datasets from the Center for Pattern Analysis and Recognition (CPAR-2012), which were composed by 100 individuals where every individual composed three specimens of the Devanagari and Latin content that is intended for tests. The trial yielded 100% right distinguishing proof on the preparation set. Be that as it may, we watched a 88% and 89% right recognizable proof rate when we tried different things with 200 preparing tests and 100 test tests on transcribed Devanagari and Latin content. The table 2.1 shows the features, accuracy and language of various author recognition systems.

Table 2.1 Different writer identification system

Author	Year	Title	Features and Classifier	Language	Accuracy
Srihari et al.	2002	Individuality of Handwriting	Features extracted from word, paragraph and documents level. Micro and Macro features are evaluated. The classifier used is Nearest neighbor [1].	English	N.R.
Bulacu, Schomaker, and Vuurpijl	2003	Writer Identification Using Edge-Based Directional Features	Evaluated feature is edge-based directional features and classifier used is nearest neighbor	Dutch	N.R.

			[2].		
Schomaker and Bulacu	2004	Automatic Writer Identification Using Connected Component Contours and Edge-Based Features of Uppercase Western Script	Connected-components contours are used for extraction of feature and KNN is used for classification of data [3].	English	N.R.
Schlapbach and Bunke	2006	Combining Multiple Features for Text-Independent Writer Identification and Verification	Probability distribution function is used for feature extraction part and chi square distance is used for similarity measure [5].	English and Dutch	N.R
Bulacu and Schomaker	2007	Using codebooks of fragmented connected-component contours in forensic and historic writer identification	Stochastic pattern generator is calculated for feature extraction and nearest neighbor is for classification [7].	Dutch	N.R.

Yan et al.	2009	Chinese Handwriting Identification Based on Stable Spectral Feature of Texture Images	Fast fourier transformation is used for feature extraction and weighted Euclidean distance classifier is used as classifier [9].	Chinese	N.R.
Shahabi and Rahmati	2009	A New Method for Writer Identification of Handwritten Farsi Documents	Features are Gabor filter, GLCM and k-NN classifier is used [10].	Farsi	N.R.
Biswas and Das	2012	Writer Identification of Bangla handwritings by Radon Transform Projection Profile	Extracted feature is radon transform projection profile and similarity measure is done by distance matrix [13].	Bangla	92.72%
Krishnammal	2013	Text Dependent Writer Identification using Support Vector Machine	Features that are extracted are edge hinge distribution, run length distribution, entropy etc. and applied classifier is SVM [14].	Telugu	94.27%

Ding et al.	2014	Writer Identification Based on Local Contour Distribution Feature	The used features are local distribution feature and used classifier is weighted Manhattan distance [15].	Chinese	N.R.
Kumar, Ravulakollu, and Bhat	2015	Fuzzy-Membership Based Writer Identification from Handwritten Devnagari Script	Extracted features are margins, interline spacing, inter-word spacing, and intra-word spacing and classifier used is KNN [17].	Devanagari	89%

2.3 GAPS IN LITERATURE

The domain author recognition is very popular for document verification. Work related in this domain in various languages like English, Chinese and Arabic is trending but for Devanagari and Indic script are very recent. There are possibility lots of explore in this sector. The previous work that has been done uses structural features and gradient features [16] [17]. The applied features used previously are margins, interline spacing, inter-word spacing, 64-dimensional features are computed based on gradient of the images but here we used Gabor and Statistical features for recognition of author. We have applied these features on whole text picture and characteristics are used for classification of author. Features we have applied do not require segmentation of line and word. So it is simple and efficient to apply features on whole documents and the result found is upright.

2.4 PROBLEM FORMULATION

There are unlike paths to evaluate the feature of the documents. Basically we can divide the features as Local and Global features. The Local features are calculated locally like for words, characters, or lines and global features are calculated for whole documents. Various features like edge-based directional probability, connected component contours, probability distribution function, HMM, GLCM, Gabor filters, Radon transform projection profile etc. are extracted for author recognition. The work related to Devanagari script is recent so we tried features like Gabor filters, GLCM and first order statistical features. The result found through k-NN classifier is virtuous.

2.5 MAIN OBJECTIVE

The fundamental target of this exploration work is to create a robust off-line author recognition for Devanagari script. During the process, explorations have been carried out following objectives:

- To create our database for author recognition, this is used during training and testing.
- Evaluate the features from documents using Gabor filter, GLCM filter and first order statistical features.
- Classify the documents using k-NN classifier.

CHAPTER 3

PROPOSED SYSTEM

3.1 OVERVIEW

In this chapter we have proposed the recognition of Devanagari handwritten text. The recognition of author of the given handwritten Devanagari text images is our main work and is presented in this chapter. For recognition of text, we have used gray level co-occurrence matrix (GLCM), first order statistical method, features alongside 2-D Gabor filters.

Our database consists of handwritten Devanagari text of 30 authors and each author has written 5 same documents of same content. The number of words in an image is 102 and thus it makes total of 510 words for a author and overall of 15060 words for all authors.

To recognize the author of given written by hand message. we have extracted different features set. The accuracy obtained is 90.90% with first order statistical method using k-NN classifiers. We also tried to evaluate the features from 2-D Gabor filters and we got an accuracy of 91.67% using k-NN classifiers and for GLCM features we got an accuracy of 92.30%. These different results are compared and analyzed.

3.2 DATASETS

The recognition of author in Devanagari script is recent, so there is no public or standard datasets available. Therefore we made our own database that consists of 45 authors of different age group and gender. The authors were asked to write the text document on A-4 size paper and the content of texts are same for all the authors. Each author wrote 5 documents of same content consisting of 102 words per document and thus it made a total of 510 words.

The documents that are obtained from different authors are then scan with scanner. The total text images thus obtained was 225. Out of these225, 180 writing sample are chosen for system training and rest 45 was utilized for testing. For each author there we have taken 4 documents for training and one text image for testing. Diagram 3.1 shows some examples of handwritten Devanagari text images. The samples collected from different authors were containing variation in writing styles. The authors wrote word in different size and may vary number of words in a line for author to author. In some words there are distortions also introduced and amount of such distortion depends on quality of pen ink used to write in the documents and speed of the author.

3.3 PRE-PROCESSING

The individual documents are scanned and many preprocessing techniques are applied for feature extraction. As here we extract features from for whole text, so we use textural features and structural features for feature extraction of the documents. Here we apply preprocessing steps for whole image text.

3.3.1 RGB TO GRAY

Once a picture is given as input, the first thing is to change over the picture into gray scale picture by averaging the pixel esteems by methods for the accompanying condition. Following diagram 3.1 and diagram 3.2 are RGB and gray image.

$$P=0.299*r+0.587*g+0.114*b$$

Where P—is the gray value

r—estimation of red segment of the pixel

g—estimation of green segment of the pixel b—estimation of blue segment of the pixel

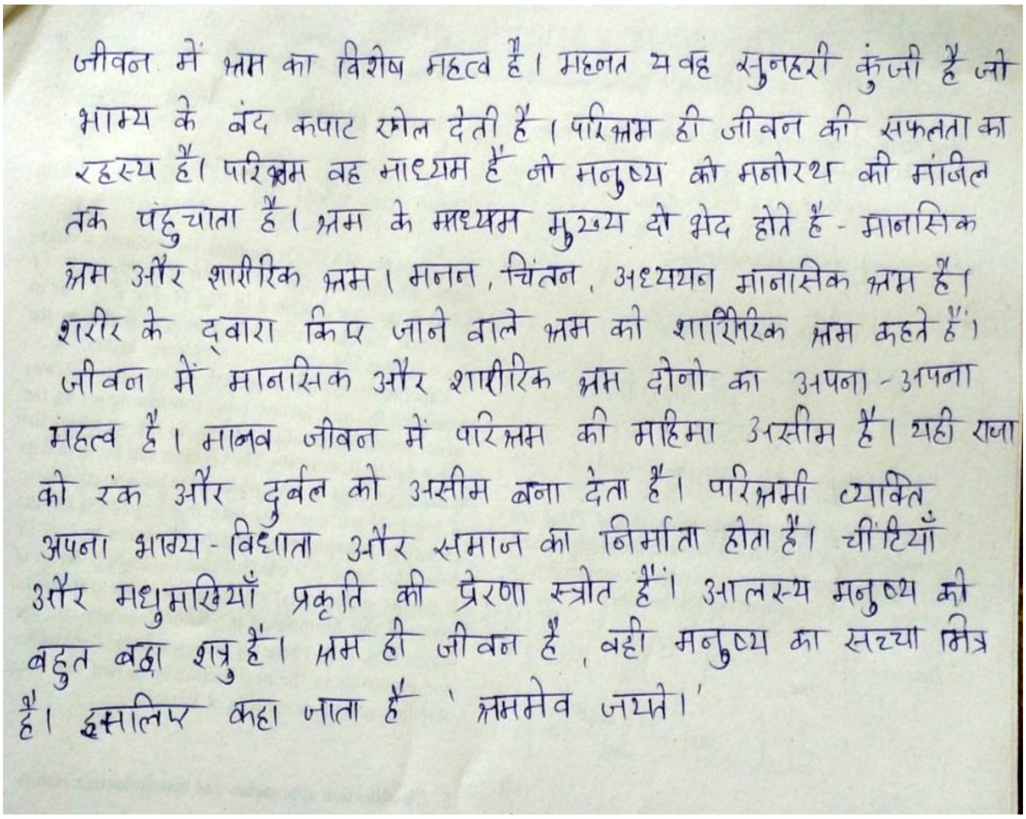

जीवन में श्रम का विशेष महत्व है। मेहनत य वह सुनहरी कुंजी है जो
भाग्य के बंद कपाट खोल देती है। परिश्रम ही जीवन की सफलता का
रहस्य है। परिश्रम वह माध्यम है जो मनुष्य को मनोरथ की मंजिल
तक पहुंचाता है। श्रम के माध्यम मुख्य दो भेद होते हैं - मानसिक
श्रम और शारीरिक श्रम। मनन, चिंतन, अध्ययन मानसिक श्रम है।
शरीर के द्वारा किए जाने वाले श्रम को शारीरिक श्रम कहते हैं।
जीवन में मानसिक और शारीरिक श्रम दोनो का अपना - अपना
महत्व है। मानव जीवन में परिश्रम की महिमा असीम है। यही राजा
को रंक और दुर्बल को असीम बना देता है। परिश्रमी व्यक्ति
अपना भाग्य - विधाता और समाज का निर्माता होता है। चींटियाँ
और मधुमक्खियाँ प्रकृति की प्रेरणा स्त्रोत हैं। आलस्य मनुष्य का
बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही मनुष्य का सच्चा मित्र
है। इसलिए कहा जाता है 'श्रममेव जयते।'

Figure 3.1 Original text image

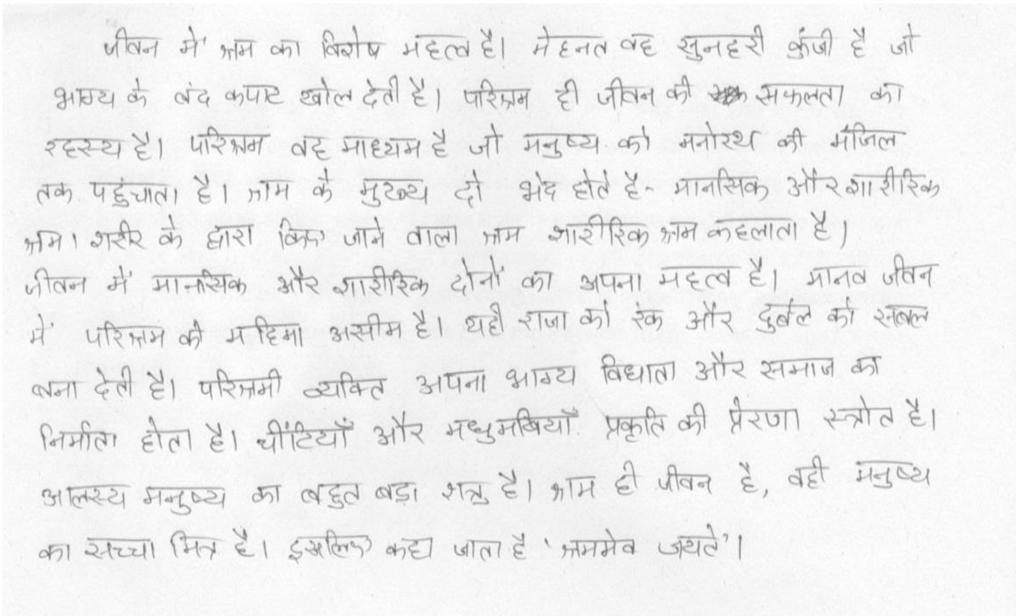

जीवन में श्रम का विशेष महत्व है। मेहनत वह सुनहरी कुंजी है जो
भाग्य के बंद कपाट खोल देती है। परिश्रम ही जीवन की ~~सफ~~ सफलता का
रहस्य है। परिश्रम वह माध्यम है जो मनुष्य को मनोरथ की मंजिल
तक पहुंचाता है। श्रम के मुख्य दो भेद होते हैं- मानसिक और शारीरिक
श्रम। शरीर के द्वारा किए जाने वाला श्रम शारीरिक श्रम कहलाता है।
जीवन में मानसिक और शारीरिक दोनों का अपना महत्व है। मानव जीवन
में परिश्रम की महिमा असीम है। यही राजा को रंक और दुर्बल को सबल
बना देती है। परिश्रमी व्यक्ति अपना भाग्य विधाता और समाज का
निर्माता होता है। चींटियाँ और मधुमक्खियाँ प्रकृति की प्रेरणा स्त्रोत है।
आलस्य मनुष्य का बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही मनुष्य
का सच्चा मित्र है। इसलिए कहा जाता है 'श्रममेव जयते'।

Figure 3.2 Gray image

3.3.2 BINARISATION

For converting the gray scale image to binary image we use Otsu's strategy. In PC vision and picture taking care of, Otsu's methodology is applied actually to perform gathering based picture thresholding or, the diminishing of a graylevel picture to a twofold picture [23]. The figuring acknowledge that the photo contains two classes of pixels following bi-measured histogram (front line pixels and establishment pixels), it by then learns as far as possible detaching the two classes so that their solidified spread (intra-class contrast) is inconsequential, or equivalently (in light of the way that the entire of pairwise squared partitions is steady), so that their between class change is maximal. Consequently, Otsu's system is around a one-dimensional basic of Fisher's Discriminant Analysis.Applying Otsu’s algorithm on gray scale image is displayed in diagram 3.3.

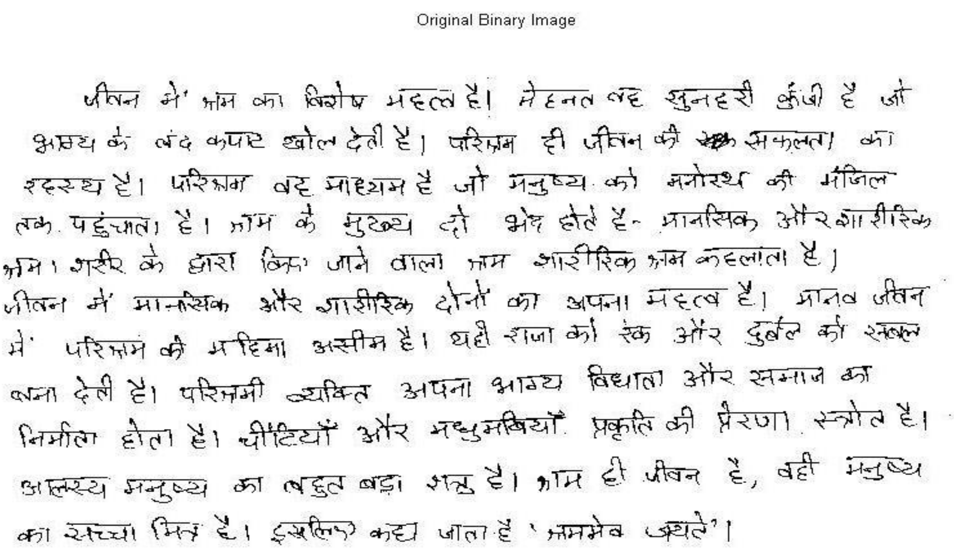


Figure 3.3 Original binary image.

3.3.3 GAUSSIAN SMOOTHING

There is clamor introduce amid picture acquisitions, to expel those commotion we utilize Gaussian channel. The Gaussian smoothing director is a 2-D convolution head that is applied to `blur' pictures and clear info and hullabaloo. In this sense it resembles the mean channel, yet it uses other part that addresses the condition of a Gaussian (`bell-shaped') knock. Before and after applying Gaussian filter on an image the horizontal projection profile is displayed in diagram 3.4 and diagram 3.5. For more about Gaussian Smoothing refer [38]

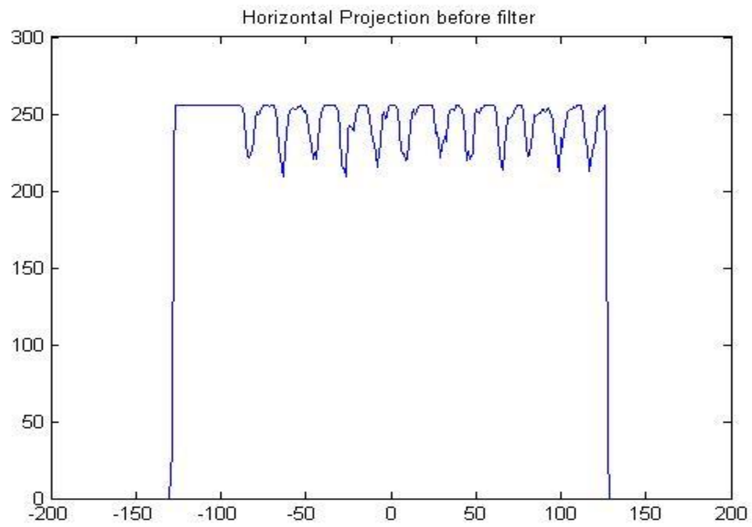


Figure 3.4 Horizontal Projection profile before applying Gaussian filter

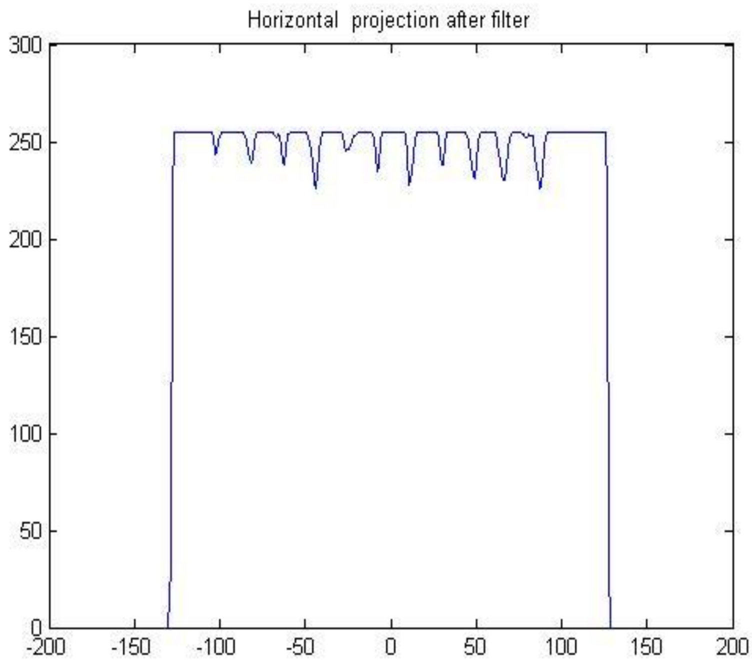


Figure 3.5 Horizontal Projection profile after applying Gaussian filter

3.3.4 PROJECTION PROFILE

Projection profile has been generally utilized as a part of line and word discovery [42]. We utilize a changed variant of a similar calculation reached out to dim level pictures [43]. To start with, the even projection profile is processed and after that smoothed with a low pass Gaussian channel diagram 3.5. In smoothed projection profile, the pinnacles relate to the space amongst lines and the valleys compare to the content lines. The pinnacles can be processed by setting the subordinate of the projection profile to zero. The smoothing and the subsidiary operation can be consolidated into one stage by convolving the projection profile with a Gaussian subordinate as takes after.

$$\frac{d}{dy} * G(y;\delta) * P(y) = \frac{dg(y;\delta)}{dy} * P(y) \tag{3.1}$$

3.3.5 PADDING OF AN IMAGE

If there is blank space at the bottom and side of the documents then we apply circular padding to those sections, so that features applied on documents are symmetric as displayed in diagram 3.6.

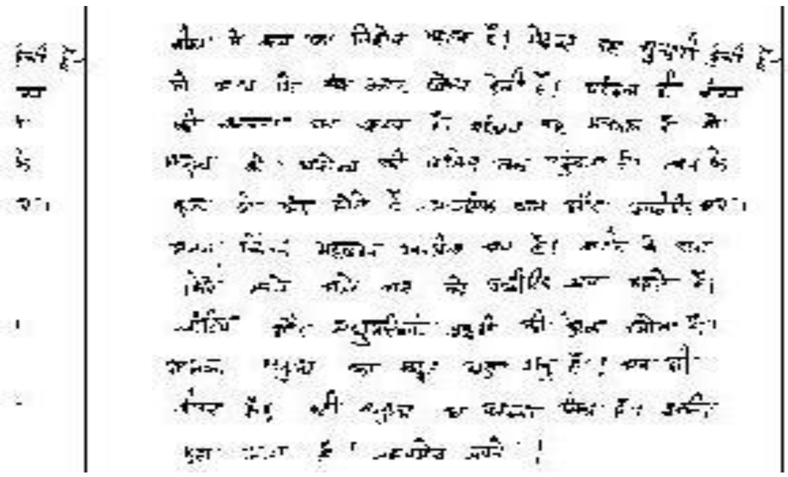


Figure 3.6 Padded image

3.3.6 NORMALIZATION

Finally we normalize the image documents to predefined size of 128×128 pixels. If the size of an image is large then applying texture features on that documents may become more complex and time consuming.

3.4 FEATURE EXTRACTION

3.4.1 GABOR FILTER

It is notable that the execution of an author distinguishing proof framework depends essentially on the elements utilized. Choice of a feature extraction method is one of the crucial decisions a developer has to take while designing such system. 2-D Gabor filters are selected to use for this purpose because of its excellent property of simulating the responsive domains of basic cells in the visual cortex [30]. Gabor filters have been utilized mostly in PC sight and surface investigation [28]. The theoretical details of Gabor filter have been discussed in section.

Gabor filter [35] is a popular feature extraction technique for pattern recognition problems. Gabor filter has the property of ideal joint spatial recurrence limitation and capacity to recreate the responsive domains of basic cells in the visual cortex [30]. Due to these attributes, the Gabor filter based elements appear to be like elements separated by human and subsequently, successful for design acknowledgment. Two-dimensional Gabor filter functions were mainly derived for use in picture processing, particularly for feature extraction and surface examination. Mathematically, a 2-D Gabor filter can be defined as a complex sinusoidally modulated Gaussian function represented in spatial domain by following equation.

$$h(x, y; \lambda, \theta, \sigma_x, \sigma_y) = exp\left[-\frac{1}{2}\left\{\frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2}\right\}\right] \times exp\left\{i\frac{2\pi R_1}{\lambda}\right\} \quad (3.2)$$

Where R_1=xcos$\boldsymbol{\theta}$+y sin$\boldsymbol{\theta}$,   and R_2= -xsin$\boldsymbol{\theta}$ +ycos$\boldsymbol{\theta}$

$\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ respectively represent the wavelength and introduction of the sinusoidal plane wave. σ_x and ,σ_y are the standard deviations of the Gaussian envelope along the x-hub and y-pivot separately. A revolution of the x-y plane by an edge $\boldsymbol{\theta}$ will bring about a Gabor channel at introduction $\boldsymbol{\theta}$ [28]. The estimation of $\boldsymbol{\theta}$ is given by π(k-1)/m, k=1,2,… ,m, where m means the quantity of introductions. For instance, when m is 4, the introductions will be 00, 450, 900 and 1350. An arrangement of Gabor channels with 5 spatial frequencies and 8 particular introductions making 40 distinctive Gabor channels is appeared in diagram 3.7 and diagram 3.8.

The Gabor highlight can be seen as the reaction of the Gabor channel situated at a testing point. The reaction is acquired by convolving the channel with a picture. Gabor channels separate the introduction subordinate recurrence substance, i.e. edge like components, from as little a region as could reasonably be expected.

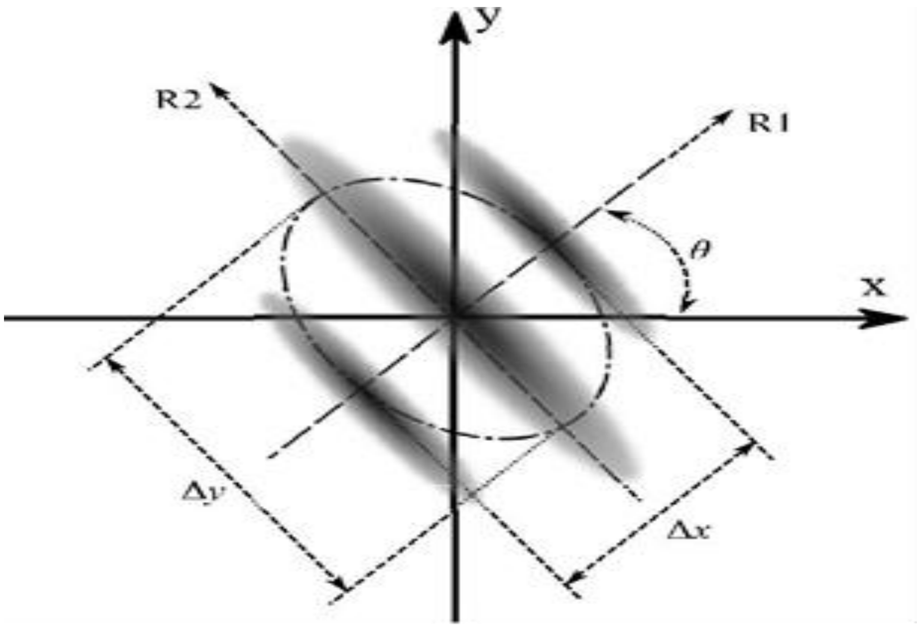
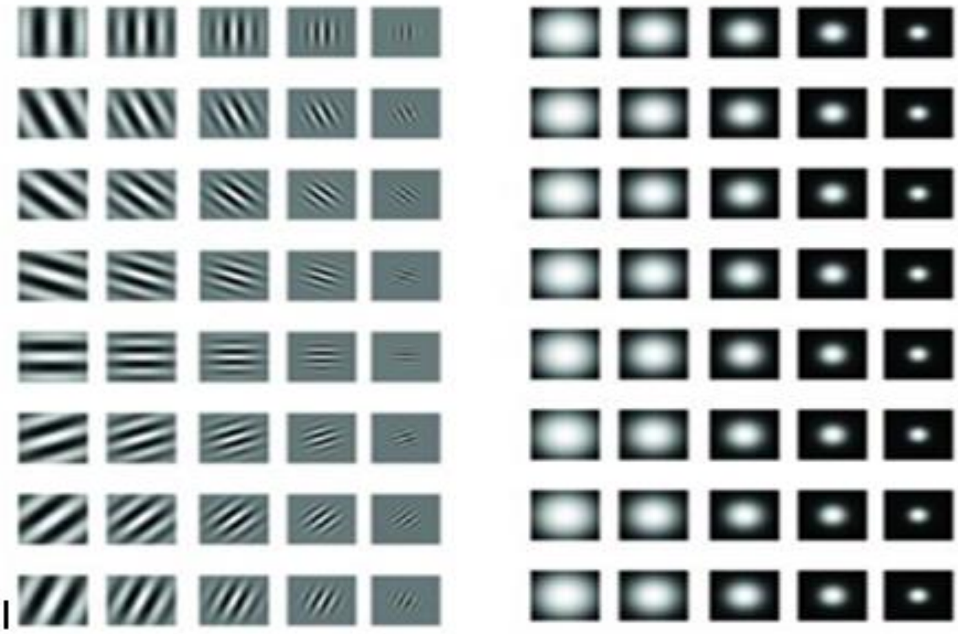


Figure 3.7 Top-perspectives of a Gabor channel in spatial space [31]

The Gabor feature corresponding to the filter with orientation $\boldsymbol{\theta}$ at a sampling point (X, Y) is defined as

$$g(X, Y, \theta, \alpha, \lambda, \sigma_x, \sigma_y) = \left| \sum_{x=-X}^{N-X} \sum_{y=-Y}^{N-Y-1} I(X+x, Y+y) f(.) \right| \qquad (3.3)$$

where I(x; y) means a N × N double picture, and |z| indicates the supreme estimation of an unpredictable number z. α is the variable by which the encircling casing is amplified [31]. For each inspecting point, m Gabor elements can be gotten for m introductions. The component vector is a 2-D lattice with highlights as lines and levels of the pictures in first segment. At the point when Gabor channels are connected to every pixels of the picture, the measurement of the sifted vector can be extensive (corresponding to the picture measurement). Along these lines, it will prompt costly calculation and capacity cost. To mitigate such issue and make the calculation strong, Gabor highlights are gathered at general interims of pixels as opposed to every last pixel of the picture.



(a) Real parts of the filters (b) Magnitude part of the filters

Figure 3.8: Example of a set of Gabor filters at 5 spatial frequencies and 8 distinct orientations

In this implementation, Gabor filters with 5 spatial frequencies (λ) and 8 different orientations (θ) are used. Value of m is set to be 8, hence the resulting orientation are: 0°, 22°, 45°, 67°, 90°, 112°, 135° and 157°.Downsampling factors along both the directions are set to 16, thus $(128 \times 128) \div (16 \times 16) = 64$ sampling points are generated. In each sampling point, Gabor filters at 8 different orientations and 5 different frequencies produce $8 \times 5 = 40$ features. Hence a total of $40 \times 64 = 2560$ Gabor features are obtained for an entire text image. Class label for each character is also appended with the feature vector.

3.4.2 GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM) FEATURES

A measurable technique for analyzing surface that considers the spatial relationship of pixels is the dim level co-event grid (GLCM) [41], otherwise called the dim level spatial reliance framework. The GLCM capacities portray the surface of a picture by computing how frequently combines of pixel with particular esteems and in a predetermined spatial relationship happen in a picture, making a GLCM, and after that separating factual measures from this grid. (The surface channel capacities, portrayed in Texture Analysis can't give data about shape, i.e., the spatial connections of pixels in a picture.)

In the wake of making the GLCMs, get a few measurements from them utilizing the graycoprops work. These insights give data about the surface of a picture. The accompanying table 3.1 records the insights.

Table 3.1 Statistical information about texture image

Statistic	Description
Contrast	Measures the local variations in the gray-level co-occurrence matrix.
Correlation	Measures the joint probability occurrence of the specified pixel pairs
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

To make a GLCM, utilize the graycomatrix work. The graycomatrix work makes a dark level co-event framework (GLCM) by figuring how regularly a pixel with the power (dim level) esteem i happens in a particular spatial relationship to a pixel with the esteem j. As a matter of course, the spatial relationship is characterized as the pixel of intrigue and the pixel to its prompt right (on a level plane contiguous), however indicate other spatial connections between the two pixels. Every component (i, j) in the resultant GLCM is just the entirety of the quantity of times that the pixel with esteem i happened in the predefined spatial relationship to a pixel with esteem j in the info picture.

The quantity of dim levels in the picture decides the span of the GLCM. Of course, graycomatrix utilizes scaling to lessen the quantity of force esteems in a picture to eight, however you can utilize the NumLevels and the GrayLimits parameters to control this scaling of dark levels. The dim level co-event lattice can uncover sure properties about the spatial appropriation of the dim levels in the surface picture. For instance, if the greater part of the passages in the GLCM are focused along the askew, the surface is coarse as for the predetermined counterbalance. We can likewise get a few factual measures from the GLCM.

To outline, the accompanying diagram demonstrates how graycomatrix computes the initial three esteems in a GLCM. In the yield GLCM, component (1, 1) contains the esteem 1 on the grounds that there is just a single occurrence in the information picture where two on a level plane adjoining pixels have the qualities 1 and 1, individually. GLCM (1, 2) contains the esteem 2 on the grounds that there are two examples where two on a level plane adjoining pixels have the qualities 1 and 2. Component (1, 3) in the GLCM has the esteem 0 in light of the fact that there are no cases of two on a level plane adjoining pixels with the qualities 1 and 3. Graycomatrix keeps preparing the information picture, checking the picture for other pixel sets (i, j) and recording the totals in the relating components of the GLCM.

Process used to create the GLCM

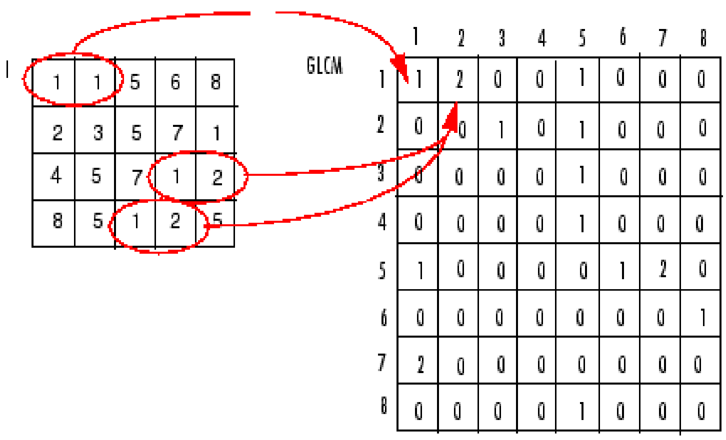


Figure 3.9 Detail process for creating GLCM

As a matter of course, the graycomatrix work makes a solitary GLCM, with the spatial relationship, or balance, characterized as two evenly contiguous pixels. In any case, a solitary GLCM won't not be sufficient to depict the textural elements of the info picture. For instance, a solitary level balance won't not be delicate to surface with a vertical introduction. Consequently, graycomatrix can make various GLCMs for a solitary information picture. The diagram 3.9 demonstrates the detail procedure of making GLCM.

To make numerous GLCMs, indicate a variety of balances to the graycomatrix work. These counterbalances characterize pixel connections of shifting heading and separation. For instance, characterize a variety of balances that indicate four bearings (level, vertical, and two diagonals) and four separations. For this situation, the information picture is spoken to by 16 GLCMs. When we assure insights from these GLCMs, we can take the normal. We indicate these counterbalances as a p-by-2 exhibit of whole numbers. Each line in the cluster is a two-component vector, [row_offset, col_offset], that indicates one balance. Row_offset is the quantity of lines between the pixel of intrigue and its neighbor. Col_offset is the quantity of sections between the pixel of intrigue and its neighbor. This illustration makes a counterbalance that indicates four bearings and 4 separations for every heading. For more data about indicating counterbalances, see the graycomatrix reference page.

```
offset = [ 0 1; 0 2; 0 3; 0 4;...
      -1 1; -2 2; -3 3; -4 4;...
      -1 0; -2 0; -3 0; -4 0;...
      -1 -1; -2 -2; -3 -3; -4 -4];
```

The diagram 3.10 delineates the spatial connections of pixels that are characterized by this variety of balances, where D speaks to the separation from the pixel of intrigue. In factual surface investigation, surface components are diagramd from the measurable dispersion of watched mixes of forces at determined positions in respect to each other in the picture. As per the quantity of power focuses (pixels) in every mix, measurements are grouped into first-arrange, second-arrange and higher-arrange insights [40].

The Gray Level Co-event Matrix (GLCM) strategy is a method for extricating second request measurable surface elements. The approach has been utilized as a part of various applications, e.g. [22], [27], [29], [33], [26], [25], [20], [19] [32]. A GLCM is a framework where the quantity of lines and sections is equivalent to the quantity of dark levels, G, in the picture. The grid component P (i, j | Δx, Δy) is the relative recurrence with which two pixels, isolated by a pixel separate (Δx, Δy), happen inside a given neighborhood, one with force i and the other with power j.

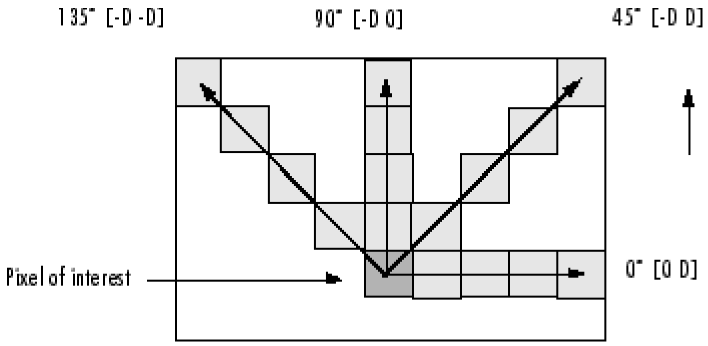


Figure 3.10 Spatial relationships of pixels

One may likewise say that the grid component P (i, j | d, θ) contains the second request likelihood esteems for changes between dark levels i and j at a specific relocation remove d and at a specific point (θ).

Given a M × N neighborhood of an information picture containing G dim levels from 0 to G − 1, let f(m, n) be the force at test m, line n of the domain.

Then

$$P(i, j \mid \Delta x, \Delta y) = W\, Q(i, j \mid \Delta x, \Delta y) \tag{3.4}$$

Where

$$W = \frac{1}{(M - \Delta x)(N - \Delta y)} \tag{3.5}$$

$$Q(i, j \mid \Delta x, \Delta y) = \sum_{n=1}^{N-\Delta y} \sum_{m=1}^{M-\Delta x} A \tag{3.6}$$

$$A = \begin{cases} 1 & \text{if } f(m, n) = i \text{ and } f(m + \Delta x, n + \Delta y) = j \\ 0 & \text{elsewhere} \end{cases} \tag{3.7}$$

A small (5 × 5) sub-image with 4 gray levels and its corresponding gray level co-occurrence matrix P (i, j | Δx = 1, Δy = 0) is illustrated below.

IMAGE		P (i, j; 1, 0)			
0 1 1 2 3		j=0	1	2	3
0 0 2 3 3	i= 0	1/20	2/20	1/20	0
0 1 2 2 3	1	0	1/20	3/20	0
1 2 3 2 2	2	0	0	3/20	5/20
2 2 3 3 2	3	0	0	2/20	2/20

Utilizing countless levels G suggests putting away a ton of transitory information, i.e. a G × G grid for every mix of (Δx, Δy) or (d, θ). One some of the time has the dumbfounding circumstance that the networks from which the surface components are separated are more voluminous than the first pictures from which they are inferred. It is likewise evident that as a result of their substantial dimensionality, the GLCM's are extremely touchy to the measure of the surface examples on which they are assessed. In this way, the quantity of dim levels is regularly decreased. Indeed, even outwardly, quantization into 16 dark levels is regularly adequate for separation or degmentation of surfaces. Utilizing few levels is comparable to review the picture on a coarse scale, though more levels give a picture with more detail. Be that as it may, the execution of a given GLCM-based component, and in addition the positioning of the elements, may rely on upon the quantity of dim levels utilized.

Since a G × G framework (or histogram cluster) must be aggregated for each sub-picture/window and for every detachment parameter set (d, θ), it is normally computationally important to confine the (d, θ)- qualities to be tried to a predetermined number of qualities. Diagram 1 beneath delineates the geometrical connections of GLCM estimations made for four separations (d = max {| Δx |, | Δy |}) and points of θ = 0, π/4, π/2 and 3π/4 radians under the presumption of rakish symmetry.

So as to acquire a measurably dependable gauge of the joint likelihood conveyance, the lattice must contain a sensibly expansive normal inhabitance level. This can be accomplished either by limiting the quantity of dim esteem quantization levels or by utilizing a moderately huge window. The previous approach brings about lost surface portrayal exactness in the examination of low abundancy surfaces, while the last causes instability and blunder if the surface changes over the expansive window. A regular trade off is to utilize 16 dim levels and a window of around 30 to 50 pixels on each side.

Straightforward connections exist among specific sets of the assessed likelihood conveyances P (d, θ) .Let P^t(d, θ) signifies the transpose of the grid P (d, θ). At that point

$$P\ (d, 0^0) = P^t(d,\ \ 180^0)$$

$$P\ (d, 45^0) = P^t(d,\ \ 225^0)$$

$$P\ (d, 90^0) = P^t(d,\ \ 270^0)$$

$$P\ (d, 135^0) = P^t(d,\ \ 315^0)$$

Along these lines, the information of $P\ d$,1800, $P\ d$,2250, P (d, 2700), and $P(d$,3150) adds nothing to the detail of the surface.

The Gray-Level Co-event Matrix (GLCM) depends on the extraction of a dim scale picture. It considers the connection between two neighboring pixels, the principal pixel is known as a kind of perspective and the second is known as a neighbor pixel [34]. The GLCM is a square grid with Ng measurement, where Ng parallels the quantity of dim levels in the picture.

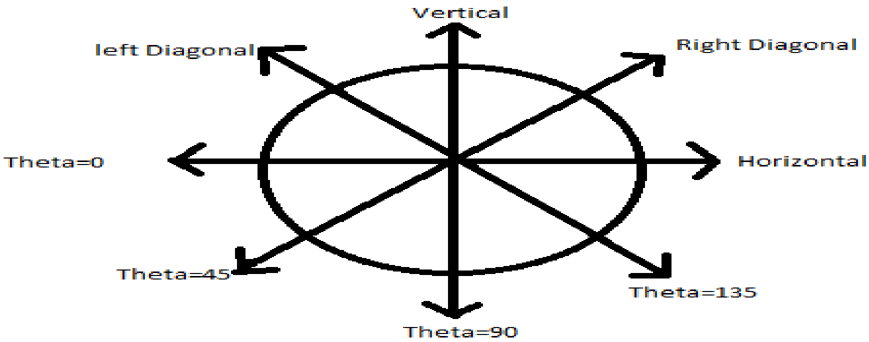


Figure 3.11 Diagram of angles, the Haralick texture features are calculated in each of these directions.

Haralick et al [27] characterized 14 surface components, these elements contain the data about the picture such homogeneity, differentiate, the many-sided quality of the picture, and so on. They are utilized as a part of numerous applications and picture recovery. This nearness can happen in four headings in view of the point, level, vertical, right corner to corner, and left slanting. Diagram 3.11 demonstrates these bearings.

The accompanying conditions are required for figuring Haralick surface element

.

$$p_x(i) = \sum_{j=1}^{N_g} p_{d,\theta}(i,j) \tag{3.8}$$

$$p_y(j) = \sum_{i=1}^{N_g} p_{d,\theta}(i,j) \tag{3.9}$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=0}^{N_g} p_{d,\theta}(i,j), \qquad k = \{2,3,\dots 2N_g\}, \qquad k = i+j \tag{3.10}$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=0}^{N_g} p_{d,\theta}(i,j), \qquad k = \{0,1,\dots N_g\}, \qquad k = |i-j| \tag{3.11}$$

Haralick Texture Features: With this strategy, 14 surface components are taken for each picture. The components are as per the following**:**

Contrast (CON): Contrast a measure of force or dark level varieties between the reference pixel and its neighbor. The visual discernment is the distinction in appearance of at least two sections of a domain appear to be at the same time or progressively.

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{d,\theta}(i,j) \right\} \ |i-j| = n \tag{3.13}$$

Angular Second Moment (ASM): ASM otherwise called consistency or vitality, measures the picture homogeneity. ASM is high when pixels are fundamentally the same as.

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{d,\theta}(i,j)^2 \tag{3.12}$$

Correlation (COR): Connection diagrams the direct reliance of the dim level esteems in the co-event network [29]. It demonstrates how the reference pixel is identified with its neighbor.

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{d,\theta}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3.14}$$

Where σ_y ,σ_x,μ_y and μ_x are the standard deviations and means of p_y ,and p_x.

Sum of Squares: Variance: Following is evaluation of gray tone variance.

$$f_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-\mu)^2 p_{d,\theta}(i,j) \tag{3.15}$$

Inverse Difference Moment (IDM): IDM likewise in some cases named homogeneity, measures the nearby homogeneity of an advanced picture. IDM restores the measures of the closeness of the appropriation of the GLCM components to the GLCM slanting.

$$f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1+(i-\mu)^2} p_{d,\theta}(i,j) \tag{3.16}$$

Sum Average (mean)

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{3.17}$$

Sum Variance

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i) \tag{3.18}$$

Sum Entropy

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log p_{x+y}(i) \tag{3.19}$$

On the off chance that the likelihood measures up to zero then the log(0) is not characterized. To keep this issue, it is prescribed to utilize log(p+ε) that ε is a subjectively little positive steady, rather than log(p).

Entropy (ENT): Entropy demonstrates the measure of data of the picture that is required for picture pressure.

$$f_9 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{d,\theta}(i,j) \log (p_{d,\theta}(i,j)) \tag{3.20}$$

The high entropy picture has an extraordinary difference from one pixel to the its neighbor and can't be compacted as a low entropy picture which has a low differentiation (a considerable measure of measure of pixels have the same or comparable esteem) [18].

Difference Variance

$$f_{10} = variance\ of\ p_{x-y} \tag{3.21}$$

Difference Entropy

$$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) \, log \, p_{x-y}(i) \tag{3.22}$$

Information Measures of Correlation 2

$$f_{13} = (1 - \exp[-2.0(HXY@ - HXY)])^{1/2} \tag{3.24}$$

Information Measures of Correlation 1

$$f_{12} = \frac{HXY - HXY1}{\max(HX, HY)} \tag{3.23}$$

Where

$$HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{d,\theta}(i,j) \log(p_{d,\theta}(i,j)) \tag{3.25}$$

HY and HX are entropies of p_y and p_x

$$HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{d,\theta}(i,j) \log(p_x(i)\, p_y(j)) \tag{3.26}$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i) p_y(j) \log(p_x(i)\, p_y(j)) \tag{3.27}$$

Maximal Correlation Coefficient

$$f_{14} = (second\ largest\ eigenvalue\ of\ Q)^{1/2} \tag{3.28}$$

Where

$$Q(i,j) = \sum_{k} \frac{p_{d,\theta}(i,k)p_{d,\theta}(j,k)}{p_x(i)p_y(k)} \tag{3.29}$$

The variance is a evaluation of the distribution of the values around the mean, it is similar to the entropy.

3.4.3 FIRST ORDER STATISTICAL METHOD

First-order statistics approximate characteristics (e.g. average and variance) [45] of single pixel values by waiving the spatial interaction between picture pixels. This technique gives the 1D histogram of a picture in light of its dim level. The histogram is basically a rundown of the measurable data about the picture. The likely thickness (p (i)) of event of the power levels is computed by separating the qualities h(i) in the aggregate number of pixels in the Nx × Ny picture.

$$p(i) = \frac{h(i)}{N_x} \times N_y \ \ , i = \{0,1, \ldots N_g - 1\} \tag{3.30}$$

The histogram characterizes the qualities of the picture, for instance, a barely disseminated histogram showed the low-differentiate picture. A bimodal histogram regularly recommends that the picture contained a question with a tight force run against a foundation of contrasting power [34].

The features achieved are:

Mean: The mean characterizes the average level of luminosity of the picture or surface

$$\mu = \sum_{i=0}^{N_g - 1} ip(i) \tag{3.31}$$

Variance: This characterizes the variety of intensity around the mean

$$\sigma^2 = \sum_{i=0}^{N_g-1} (i-\mu)^2 p(i) \tag{3.32}$$

Skewness: It characterizes the symmetry.

$$\mu^3 = \sigma^{-3} \sum_{i=0}^{N_g-1} (i-\mu)^3 p(i) \tag{3.33}$$

Kurtosis: This is an evaluation of the flatness of the histogram

$$\mu^4 = \sigma^{-4} \sum_{i=0}^{N_g-1} ((i-\mu)^4 p(i)) - 3. \tag{3.34}$$

Energy: That returns the sum of squared elements

$$E = \sum_{i=0}^{N_g-1} [p(i)]^2. \tag{3.35}$$

Entropy:

$$H = -\sum_{i=0}^{N_g-1} p(i) log_2 [p(i)]. \tag{3.36}$$

3.5 CLASSIFICATION

3.5.1 K-NEAREST NEIGHBOUR (k-NN) CLASSIFIER

In application, as a principle we have a depiction of a surface specimen and we need to discover which component of a database best matches that example. Along these lines is arrangement: to relate the fitting class mark (sort of surface) with the test by utilizing the estimations that depict it. One approach to make the affiliation is by finding the individual from the class (the example of a known surface) with estimations which vary by minimal sum from the test's estimations [37]. Regarding Euclidean separation, the distinction d between the M depictions of a specimen, s, and the portrayal of a known surface, k, is which is additionally called the L2 standard.

$$d = \sqrt{\sum_{i=1}^{M} (s_i - k_i)^2} \tag{3.37}$$

Alternative distance metrics include: the L1 norm which is the sum of the modulus of the differences between the measurements

$$L_1 = \sum_{i=1}^{M} |s_i - k_i| \tag{3.38}$$

and the Bhattacharyya distance B

$$B = -ln \sum_{i=1}^{M} \sqrt{s_i \times k_i} \tag{3.39}$$

in any case, this gives off an impression of being utilized less, as different measurements, for example, the Matusita distinction. On the off chance that we have M estimations of N known specimens of surfaces and we have O tests of every, at that point we have a M-dimensional component space that contains the N × O focuses. In the event that we select the point, in the component space, which is nearest to the present example, at that point we have chosen the specimens closest neighbor. This is outlined in diagram 3.3 where we have a two-dimensional component space created by the two measures made on each example, measure1 and measure 2. Each example gives diverse esteems for these measures however the specimens of various classes offer ascent to bunches in the component space where each group is related with a solitary class. In Diagram 3.12 we have seven specimens of two known surfaces: Class An and Class B delineated by × and O separately. We need to arrange a test, delineated by +, as having a place either with Class An or to Class B (i.e. we expect that the preparation information contains agents of every single conceivable class). Its closest neighbor, the specimen with minimum separation, is one of the examples of Class A so we could then say that our test gives off an impression of being another example of Class An (i.e. the class mark related with it is Class A). Unmistakably, the bunches will be far separated for measures that have great oppressive capacity while the groups will cover for measures that have poor unfair capacity.

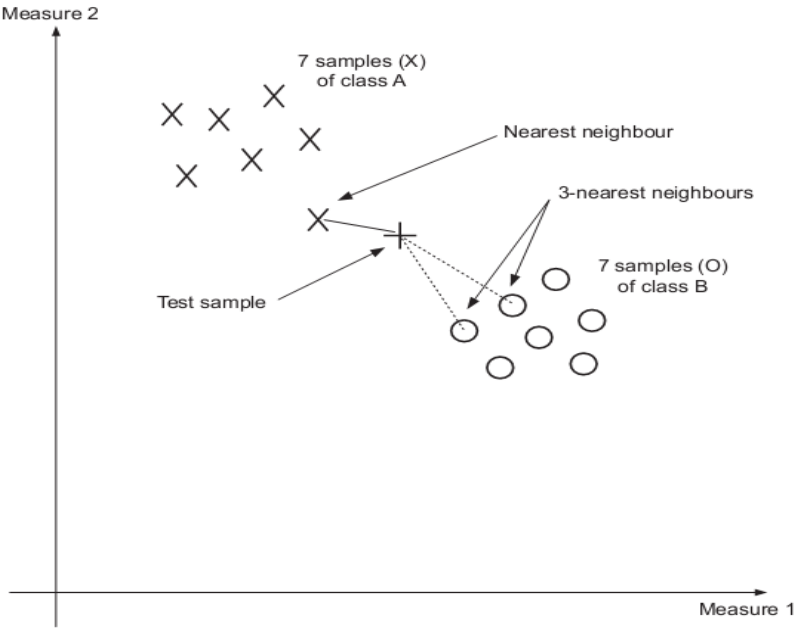


Figure 3.12 Feature space and classification [37]

That is the manner by which we can pick measures for specific assignments. Prior to that, let us take a gander under the most favorable conditions to relate a class mark with our test. Arranging a test as the preparation test it is nearest to in include space is really a particular instance of a general characterization run known as the k-closest neighbor run the show. In this control, the class chose is the method of the examples closest k neighbors. By the k-closest neighbor run, for k = 3, we select the closest three neighbors (those three with the minimum separation) and their mode, the maximally spoke to class, is credited to the example. In diagram 3.12, the 3-closest neighbor is really Class B since the three closest specimens contain one from Class An (its closest neighbor) and two from Class B. Since there are two components of Class B, at that point the specimen is ascribed to this class by the 3-closest neighbor run the show. In that capacity, choice from more than one point presents a type of highlight space smoothing and permits the order choice not to be influenced by uproarious anomaly focuses. Plainly, this smoothing has more prominent impact for bigger estimations of k.

k-NN classifier utilizes the example based learning by relating obscure example to the known by some separation or some other closeness work. It arranges the protest by larger part vote of its neighbor. Since it considers just neighbor question a specific level, it utilizes nearby guess of separation work. It implies languid or occurrence learning is utilized as a part of k-NN while in different classifiers as SVM enthusiastic learning is utilized. K indicates the quantity of closest neighbors to be considered and the class of greater part of these neighbors is resolved as the class of obscure example. In the diagram 3.13 the characterization of articles with k=3 (strong line circle) and k=5 (dabbed circle) is appeared.

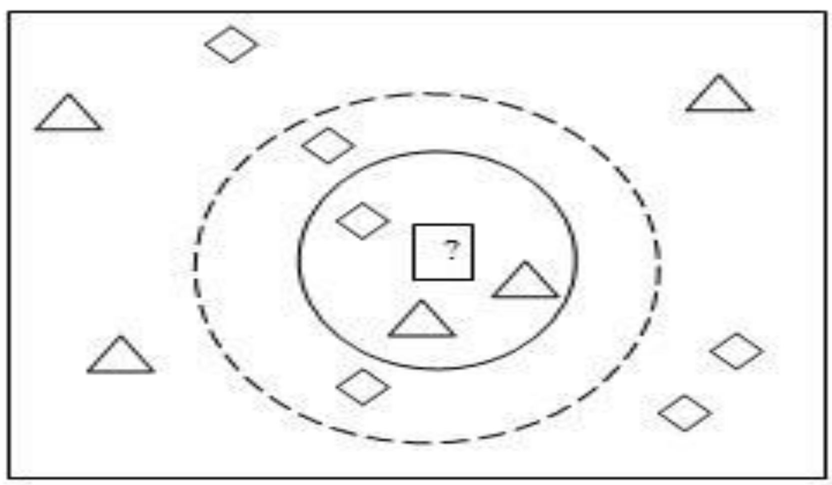


Figure 3.13 Classification of objects using K-NN with k=3 and k=5

It is outstanding that class of protest is modified in both cases. At k=3 the obscure example will be named triangular shape protest while at k=5 it will be named precious stone shape question in light of the fact that at k=3 triangular items are in larger part (inside inward strong hover) while at k=5 jewel formed articles are in greater part (inside external dashed circle).
The separation work used to discover the closest neighbors are: Euclidian separation, aggregate of supreme contrasts, cosine, relationship and rate of bits that vary.

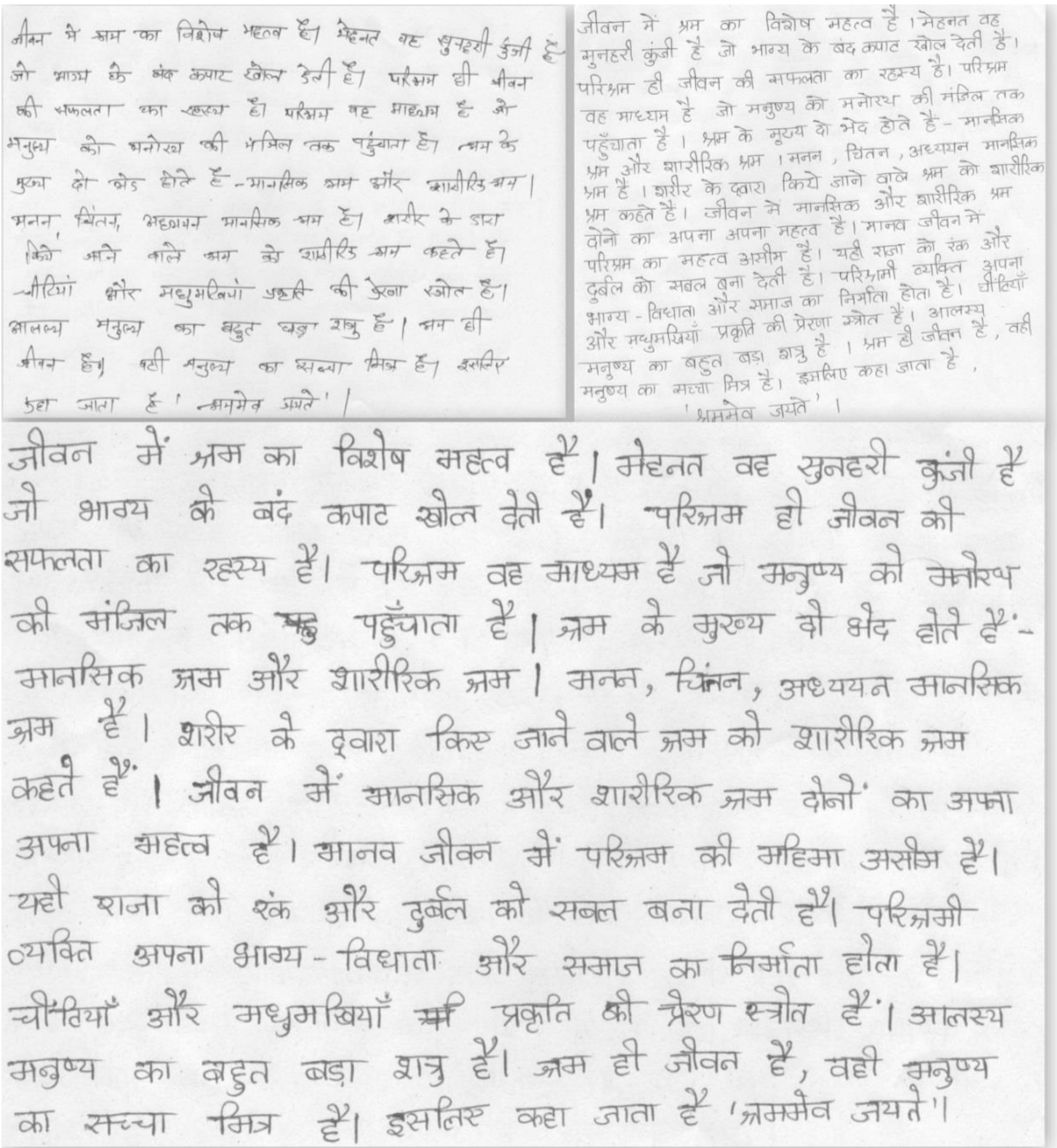
Parameter Selection: More generous models can be refined by discovering k, where k > 1, neighbors and allowing the bigger part to vote pick the consequence of the class naming. A higher estimation of k achieves a smoother, less locally sensitive, work. Generally, greater estimations of k diminish the effect of upheaval on the gathering, however make confines between classes less unmistakable. The best choice of k depends on the data. A not too bad k can be picked by various heuristic systems, for example, cross-endorsement. The phenomenal circumstance where the class is expected to be the class of the closest get ready test (i.e. exactly when k = 1) is known as the nearest neighbor estimation. More portrayal about K-NN classifier can be found at [37] and [44].

In our usage we have utilized Euclidean separation as the separation parameter, while utilizing other separation capacities we acquired debased outcomes. Utilizing k-closest neighbor classifier we discovered productive and great outcomes.

CHAPTER 4

EXPERIMENTAL SETUP AND RESULTS

This chapter presents the analysis of the experimental results conducted on datasets that we have made for Devanagari script. All tests are directed utilizing MATLAB tool and tested on core i3 dual core CPU with 2 GB RAM on Windows -7. The samples of document that we have used for our proposed work is displayed in figure 4.1



जीवन मे श्रम का विशेष महत्व है। मेहनत वह सुनहरी कुंजी है जो भाग्य के बंद कपाट खोल देती है। ~~पसीन~~ परिश्रम ही जीवन की सफलता का रहस्य है। परिश्रम वह माध्यम है जो मनुष्य को मनोरथ की मंज़िल तक पहुँचाता है। श्रम के मुख्य दो भेद होते हैं - मानसिक श्रम और शारीरिक श्रम। मनन, चिंतन, अध्ययन मानसिक श्रम हैं। शरीर के द्वारा किये जाने वाले ~~श्रम को~~ श्रम को शारीरिक श्रम कहते हैं। जीवन में मानसिक और शारीरिक श्रम दोनों का अपना अपना महत्व है। मानव जीवन मे परिश्रम की महिमा असीम है। यही राजा को रंक और दुर्बल को सबल बना देती है। परिश्रमी व्यक्ति अपना भाग्य-विधाता और समाज का निर्माता होता है। चींटियाँ ~~श्रम~~ और मधुमक्खियाँ प्रकृति की प्रेरणा स्रोत हैं। आलस्य मनुष्य का बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही मनुष्य का सच्चा मित्र है। इसलिए कहा जाता है 'श्रमेव जयते'।

जीवन मे श्रम का विशेष महत्व है। मेहनत वह सुनहरी कुंजी है जो भाग्य के बंद कपाट खोल देती है। परिश्रम ही जीवन की सफलता का रहस्य है। परिश्रम वह माध्यम है जो मनुष्य को मनोरथ की मंजिल तक पहुँचाता है। श्रम के मुख्य दो भेद होते हैं - मानसिक श्रम और शारीरिक श्रम। मनन, चिंतन, अध्ययन मानसिक श्रम है। शरीर के द्वारा किये जाने वाले श्रम को शारीरिक श्रम कहते हैं। जीवन में मानसिक और शारीरिक श्रम दोनों का अपना अपना महत्व है। मानव जीवन में परिश्रम की महिमा असीम है। यही राजा को रंक और दुर्बल को सबल बना देती है। परिश्रमी व्यक्ति अपना भाग्य - विधाता और समाज का निर्माता होता है। चीटियाँ और मधुमखियाँ प्रकृति के प्रेरणा स्रोत हैं। आलस्य मनुष्य का बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही मनुष्य का सच्चा मित्र है। इसलिए कहा जाता है श्रमेव जयते।

जीवन में श्रम का विशेष महत्व है। मेहनत वह सुनहरी कुंजी है जो भाग्य के बंद कपाट खोल देती है। परिश्रम ही जीवन की सफलता का रहस्य है। परिश्रम वह माध्यम है जो मनुष्य को मनोरथ की मंजिल तक पहुंचाता है। श्रम के मुख्य दो भेद होते हैं- मानसिक श्रम और शारीरिक श्रम। मनन, चिंतन, अध्ययन मानसिक श्रम है। शरीर के द्वारा किये जाने वाले शर्म को शारीरिक श्रम कहते है। जीवन में मानसिक और शारीरिक श्रम दोनों का अपना अपना महत्व है। मानव जीवन में परिश्रम की महिमा असीम है। यही राजा को रंक और दुर्बल को सबल बना देती है। चीटियाँ और मधुमखियाँ प्रकृति की प्रेरणा स्त्रोत है। आलस्य मनुष्य का बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही मनुष्य का सच्चा मित्र है। इसलिए कहा जाता है 'श्रमेव जयते'।

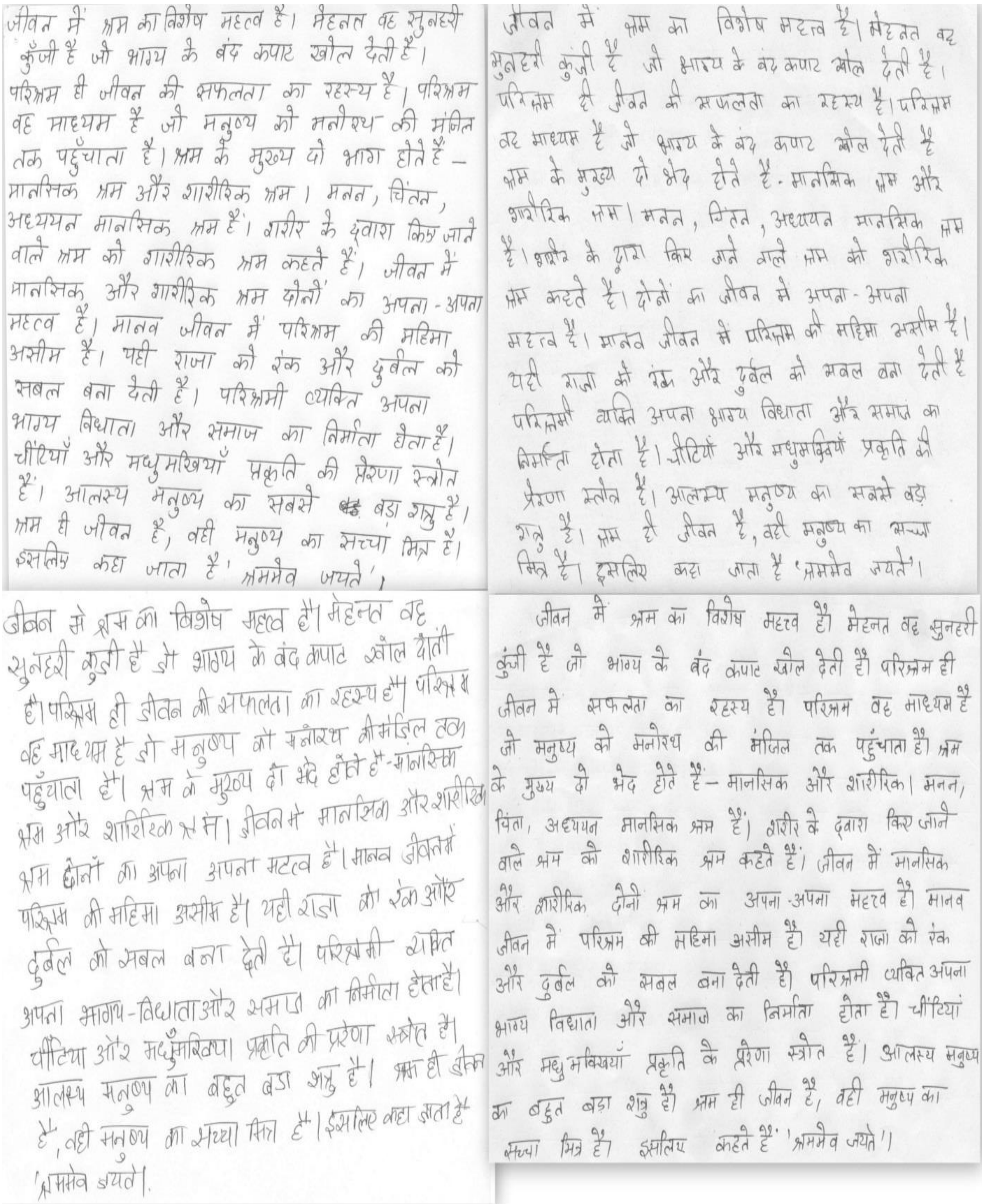


Figure 4.1 Samples of handwritten Devanagari documents

4.1 PERFORMANCE ANALYSIS OF HANDWRITTEN DEVANAGARI TEXT

The handwritten Devanagari text is identified using different features with k-NN classifier. The evaluated features are first order statistical method, Gabor filter and gray level co-occurrence matrix (GLCM). By using k-NN classifier we have identified the author of the given text.

The numbers of features that are evaluated from 2-D Gabor filter are more (2560), so it is not feasible to take all the features for recognition of the author. Here we have taken features at an interval of 40 so total features from this are 64. For more see in section 3.4

Practically speaking, k-ought to be vast so blunder rate is limited, k too little prompt loud choice limits. k ought to be sufficiently little so that exclusive close-by tests are incorporated, k too extensive will prompt over smoothed limits. Above expressed explanation is not minor, this is an intermittent issue, need to smooth information, yet not all that much. Table 4.1 shows recognizable proof rate for various estimations of k for various elements. In principle, when substantial number of tests is accessible, the bigger the k, better is characterization. In practice, k- should be large so that error rate is minimized, k too small lead to noisy decision boundaries. k should be small enough so that only nearby samples are included, k too large will lead to over smoothed boundaries. Above stated statement is not trivial, this is a recurrent issue, need to smooth data, but not too much. Table 4.1 shows recognition rate for different values of k for different features. In theory, when large number of samples is available, the larger the k, better is classification. As here we have small datasets available we get better results for small values of k. In diagram 4.2, diagram 4.3 and diagram 4.4 show the result using first order statistical method, Gabor filter and GLCM features.

Table 4.1 Results from different features using k-NN for different values of k

K-NN	Recognition rate		
	First order statistical features	Gabor filter	GLCM features

K=1	90.90	91.67	92.30
K=2	90.90	91.67	92.30
K=3	85	66.70	92.30
K=4	75	83.33	69.20
K=5	75	75	84.62
K=6	85	85	76.92
K=7	85	75	76.92

The main concern is to selection value of k in k-nearest neighbor. We have checked different values of k for recognition of author for different features evaluated. First we checked for first order statistical feature and we found that for k=1 and k=2 the recognition rate is maximum.

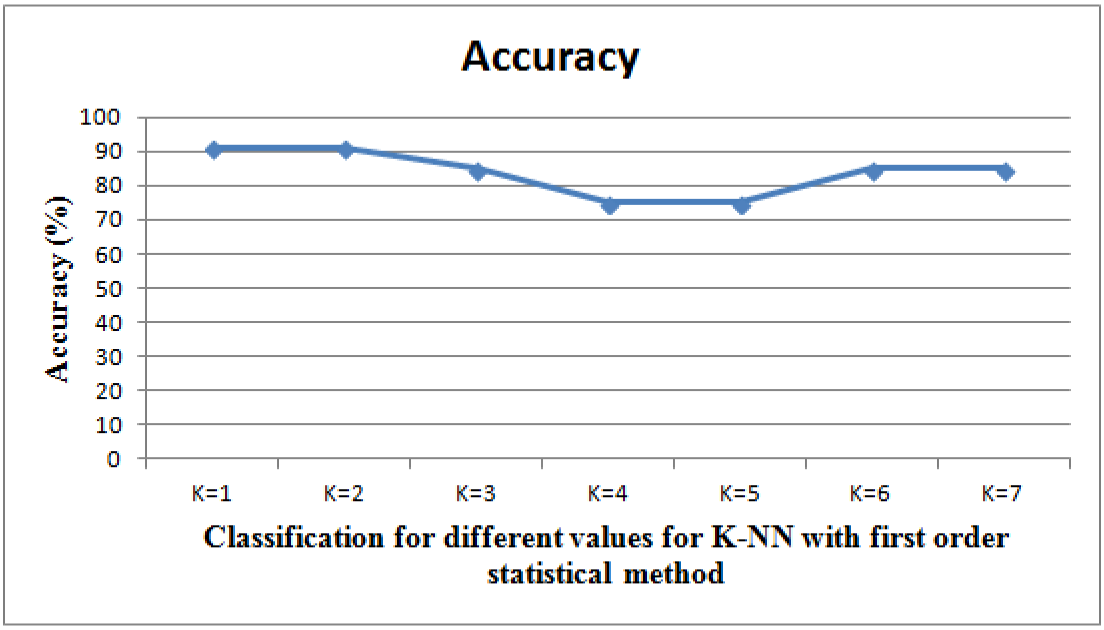


Figure 4.2 Results for author recognition with first order statistical features with k-NN for different values of k

On the contrary recognition rate for Gabor filter is maximum for k=1 and k=2 and we get the accuracy of 91.67%.

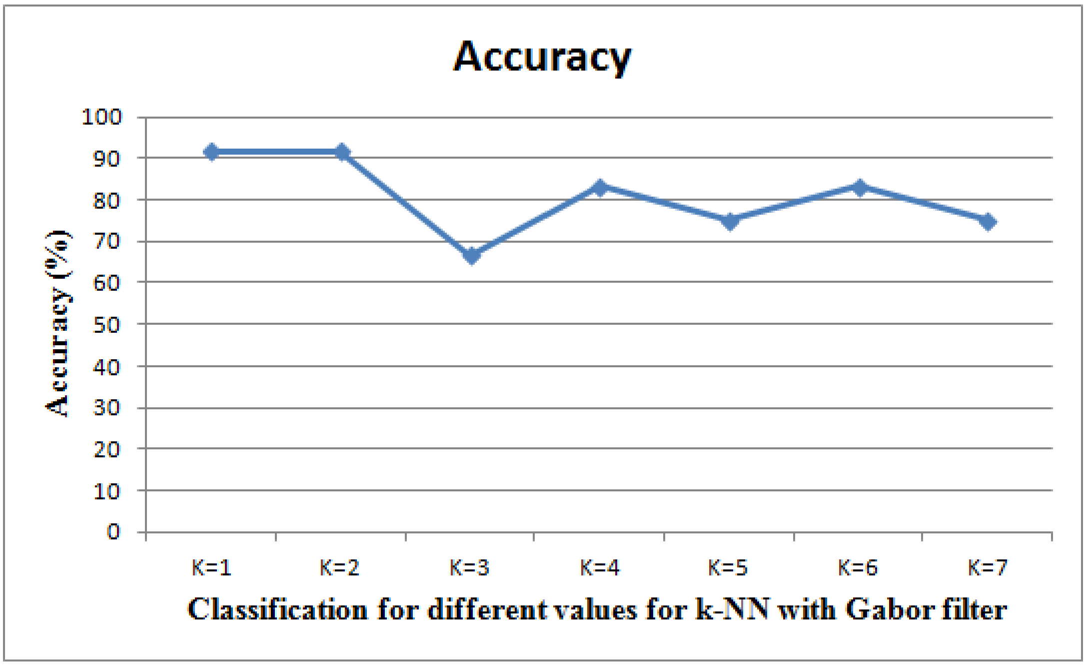


Figure 4.3 Results for author recognition with Gabor filter with k-NN for different values of k

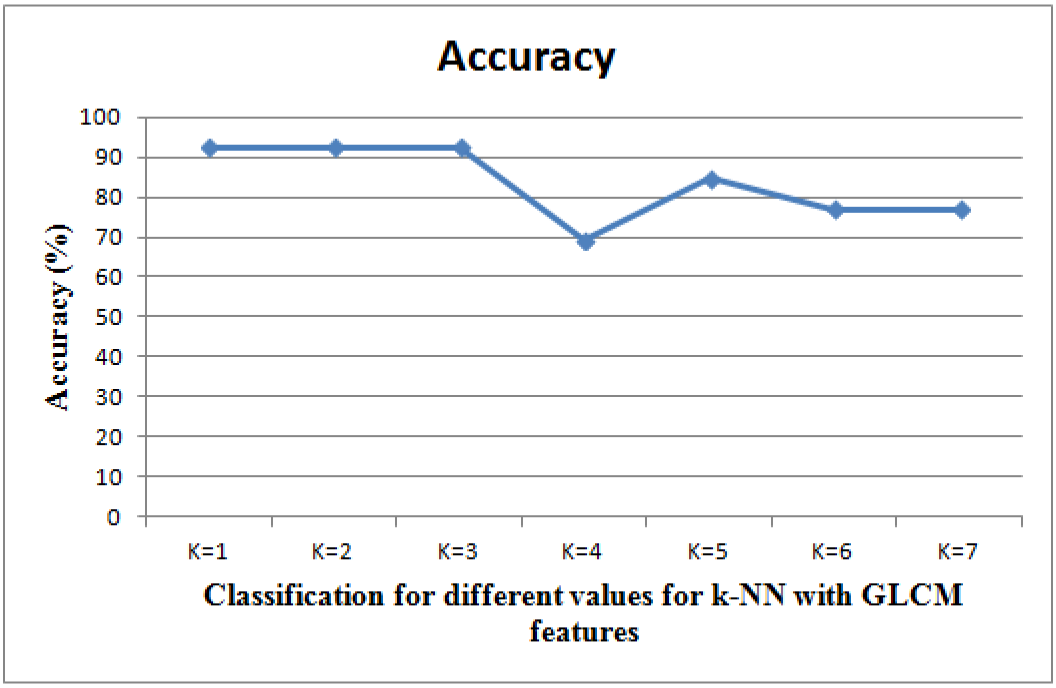


Figure 4.4 Results for writer identification with GLCM features with k-NN for different values of k

The following diagram 4.5 shows the recognition rate with k-NN for different values of k for different features used. Maximum accuracy for all features is get on small value of k and among them best result is get through GLCM features.

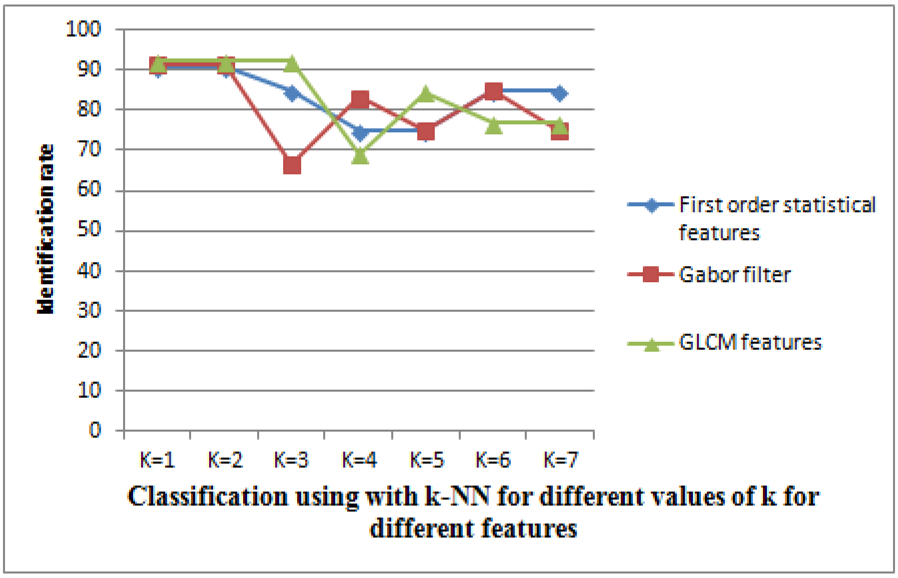


Figure 4.5 Results for author recognition for different features with k-NN for different values of k

The recognition rates for different features are given in the following table.

Table 4.2 Results from different features using K-NN classifier

Serial No.	Types of Features	Accuracy using k-NN classifier
1.	GLCM	92.30%
2.	2-D Gabor filter	91.67%
3.	First order statistical features	90.90%

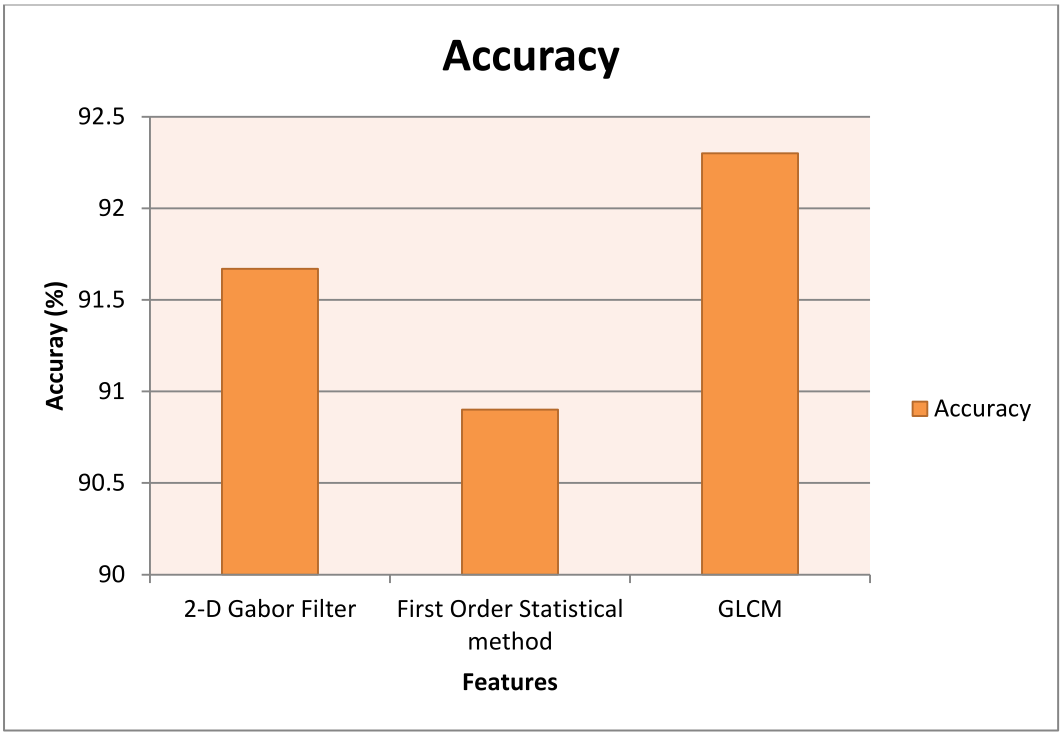


Figure 4.6 Graphical representation of accuracy of the system

The table 4.2 depicts the comparison of our proposed with all previous approaches.

Table 4.3 Comparison of proposed approach with previous approaches

Proposed by	Year	Feature used	Classifier	Language	Accuracy
Biswas and Das [13]	2012	Radon transform projection profile	Euclidean distance	Bangla	92.27%
Srihari et al. [1]	2002	Micro and Macro features are evaluated	k-NN	English	N.R
Chayan Halder et al.[16]	2015	Margins, interline spacing, inter-word spacing and intra-word spacing.	k-NN	Devanagari	89%

Yan et al. [9]	2009	Fast fourier transformation is used for feature extraction	Weighted Euclidean distance	Chinese	N.R
Our proposed work		GLCM	K-NN	Devanagari	92.30%

From the above result and table we have seen that maximum accuracy is obtained through GLCM features. The text document that has not been identified is due to noise and classification error. For the classification we have used k-NN classifier and k-NN classifier has some advantages and disadvantages. The fundamental points of interest of k-NN for arrangement are:

Extremely basic execution.

- Robust with respect to the hunt space; for example, classes don't need to be straightly detachable.
- Classifier can be refreshed online at next to no cost as new cases with known classes are introduced.

- Few parameters to tune: separate metric and k.

The fundamental detriments of the calculation are:

- Expensive testing of each case, as we have to register its separation to every single known occurrence. Specific estimations and heuristics exist for specific issues and partition limits, which can diminish this issue. This is dangerous for datasets with a significant number of attributes. Right when the amount of events is significantly greater than the amount of attributes, a R-tree or a kd-tree can be used to store illustrations, considering speedy right neighbor unmistakable confirmation.

- Sensitiveness to boisterous or unimportant characteristics, which can bring about less important separation numbers. Scaling as well as highlight determination are normally utilized as a part of blend with k-NN to alleviate this issue.

- Sensitiveness to extremely uneven datasets, where most elements have a place with one or a couple of classes, and rare classes are in this way regularly overwhelmed in many neighborhoods. This can be eased through adjusted inspecting of the more well known classes in the preparation arrange, conceivably combined with outfits.

The test data which is miss classify is due to above reason that we have discussed about k-NN and noise. The diagram 4.7 shows train data and diagram 4.8 shows the miss classify result.

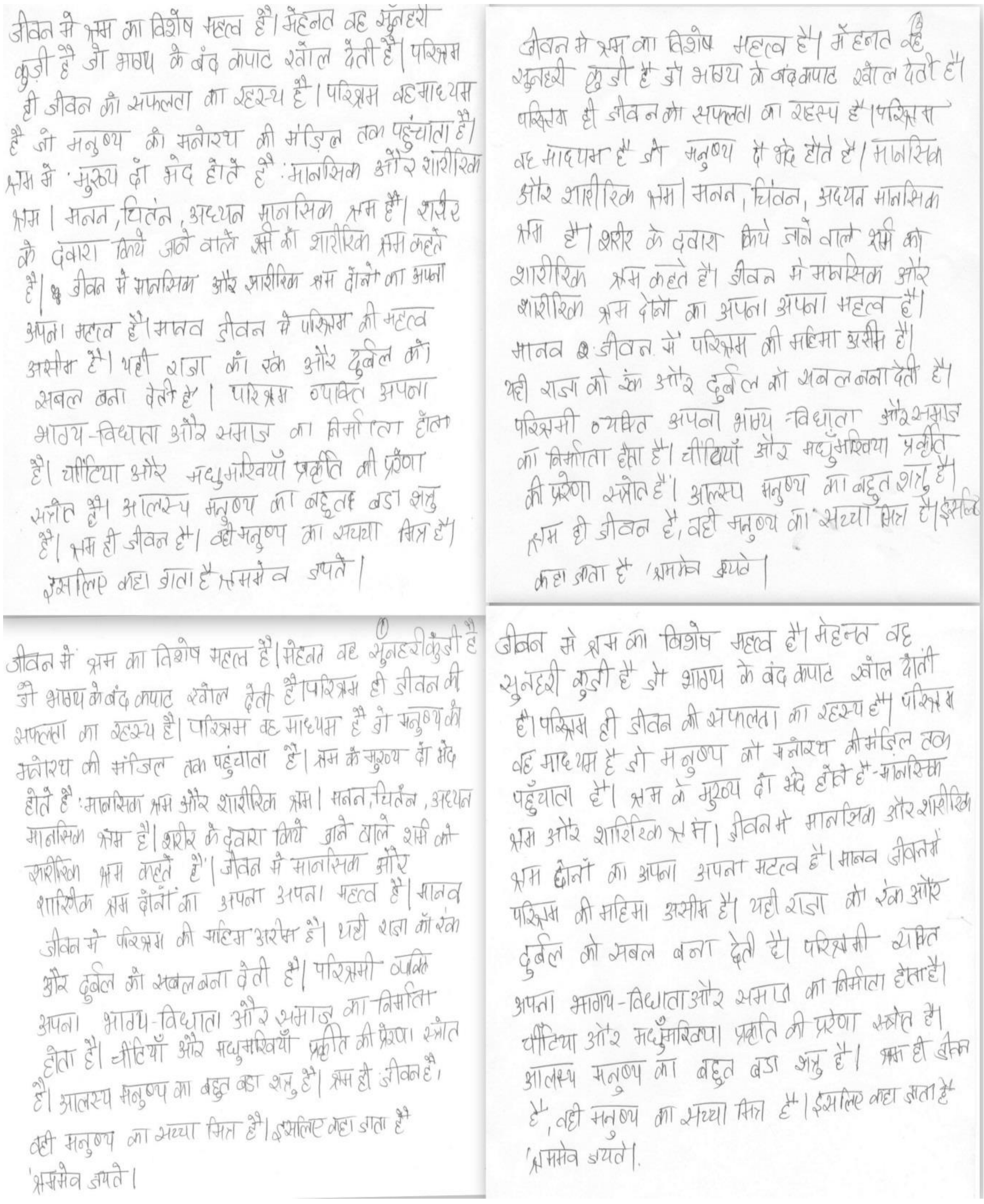


Figure 4.7 Training data for author

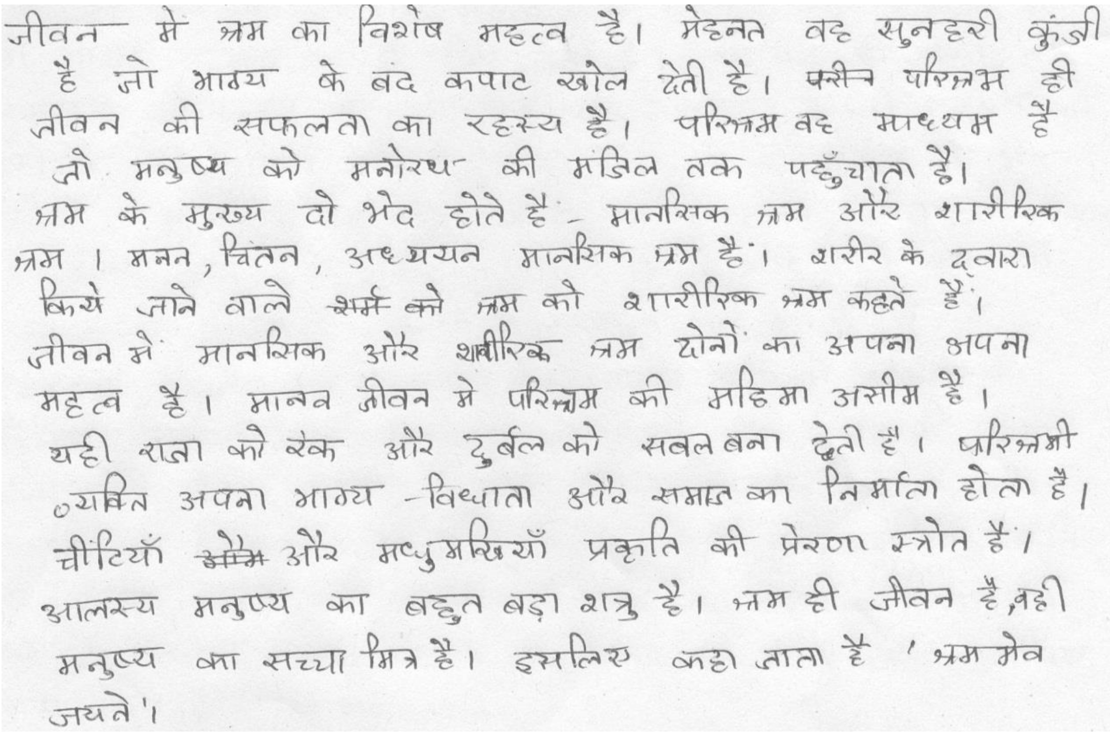

जीवन मे श्रम का विशेष महत्व है। मेहनत वह सुनहरी कुंजी
है जो भाग्य के बंद कपाट खोल देती है। ~~पसीन~~ परिश्रम ही
जीवन की सफलता का रहस्य है। परिश्रम वह माध्यम है
जो मनुष्य को मनोरथ की मंजिल तक पहुँचाता है।
श्रम के मुख्य दो भेद होते हैं - मानसिक श्रम और शारीरिक
श्रम। मनन, चिंतन, अध्ययन मानसिक श्रम हैं। शरीर के द्वारा
किये जाने वाले ~~श्रम को~~ श्रम को शारीरिक श्रम कहते हैं।
जीवन मे मानसिक और शारीरिक श्रम दोनो का अपना अपना
महत्व है। मानव जीवन मे परिश्रम की महिमा असीम है।
यही राजा को रंक और दुर्बल को सबल बना देती है। परिश्रमी
व्यक्ति अपना भाग्य-विधाता और समाज का निर्माता होता है।
चीटियाँ ~~और~~ और मधुमक्खियाँ प्रकृति की प्रेरणा स्रोत हैं।
आलस्य मनुष्य का बहुत बड़ा शत्रु है। श्रम ही जीवन है, वही
मनुष्य का सच्चा मित्र है। इसलिए कहा जाता है 'श्रम मेव
जयते'।

Figure 4.8 Miss classify result of author

As the features applied in our proposed work is on whole document, the accuracy obtained is also good. We have not segmented the documents into line or character level so, it is very simple.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

The system proposed in previous chapter was extraction of features from handwritten Devanagari text. The extracted features such as Gabor filters, first order statistical and Gray level co-occurrence matrix features are classified through K-NN classifier. The accuracy of different features is different. Overall we got efficient results from classification. There are 14 features are evaluated from gray level co-occurrence matrix, 6 features from first order statistical features and 2560 features are evaluated from 2-D Gabor filter which are efficient to identify the author of the text. K-NN is a powerful classifier which outperforms many other existing classifiers. The purpose of K-NN is to correctly classify the test data according to train data. The system performance is measured how accurately the test data is recognized. All the work has been done on our own datasets. The proposed approach gives an accuracy of 90.90% for First Order Statistical method, 91.67% for 2-D Gabor filter and 92.30% for GLCM using K-NN classifier.

5.2 FUTURE WORK

As there is no standard datasets available for handwritten Devanagari scripts. The datasets we used here are small, further we can work for large datasets. The features we evaluated are at document level, further we can use line, word and character level for enhanced results. We can also use more powerful classifier to enhance the classification results. Also multiple classifier combined together to get the better results.

REFERENCES

[1] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, “Individuality of handwriting.,” *Journal of Forensic Science*, vol. 47, no. 4, pp. 856–872, 2002.

[2] M. Bulacu, L. Schomaker, and L. Vuurpijl, “Author recognition using edge-based directional features,” *Seventh International Conference on Document Analysis and Recognition,* vol. II, no. 2, pp. 937-941, 2003.

[3] L. Schomaker and M. Bulacu, “Automatic author recognition using connected-component contours and edge-based features of uppercase western script,” *IEEE Transaction Pattern Analysis Machine Intelligence,* vol. 26, no. 6, pp. 787–798, 2004.

[4] A. Schlapbach and H. Bunke, “A author recognition and verification system using HMM based recognizers,” *Pattern Analysis Application*, vol. 10, no. 1, pp. 33–43, 2006.

[5] M. Bulacu and L. Schomaker, “Combining Multiple Features for Text-Independent Author Recognition and Verification,” *Proc. 10th International Workshop on Frontiers in Hanwriting Recognition, IWFHR*, pp. 281–286, 2006.

[6] L. Schomaker, K. Franke, and M. Bulacu, “Using codebooks of fragmented connected-component contours in forensic and historic author recognition,” *Pattern Recognition*, vol. 28, no. 6, pp. 719–727, 2007.

[7] I. A. Siddiqi and N. Vincent, “Author recognition in handwritten documents,” *Proc. International Conference on Document Analysis and Recognition, ICDAR*, vol. 1, pp. 108–112, 2007.

[8] M. Bulacu and L. Schomaker, "Text-independent author recognition and verification using textural and allographic features," *IEEE Transacton on Pattern Analysis Machine Intelligence*, vol. 29, no. 4, pp. 701–717, 2007.

[9] Y. Yan, Q. Chen, W. Deng, and F. Yuan, "Chinese handwriting recognition based on stable spectral feature of texture images," *International Journal of Intellgence Engineering System*, vol. 2 no. 1, pp. 17-22, 2009.

[10] F. Shahabi and M. Rahmati, "A new method for author recognition of handwritten Farsi documents," *Proc. Internation Conference on Document Analysis and Recognition, ICDAR*, pp. 426–430, 2009.

[11] U. Garain and T. Paquet, "Off-line multi-script author recognition using AR coefficients," *Proc. International Conference on Document Analysis and Recognition, ICDAR*, pp. 991–995, 2009.

[12] S. Chanda, K. Franke, U. Pal, and T. Wakabayashi, "Text independent author recognition for Bengali script," *Proc. International Conference on Pattern Recognition.*, pp. 2005–2008, 2010.

[13] S. Biswas and A. K. Das, "Author recognition of Bangla handwritings by radon transform projection profile," *Proc. - 10th IAPR Internation Workshop on Document Analysis System*, pp. 215–219, 2012.

[14] P. Krishnammal, "Text Dependent Author Recognition using Support Vector Machine,"*Pattern Recognition*, vol. 65, no. 2, pp. 6–11, 2013.

[15] H. Ding, H. Wu, X. Zhang, and J. Chen, "Author Recognition Based on Local Contour Distribution Feature.," *Internation Journal of Signal Processing, Pattern Recognition,* vol. 7, no. 1, pp. 169–180, 2014.

[16] C. Halder, K. Thakur, and S. Phadikar, "Author Recognition from Handwritten Devanagari Script," *Springer,* pp. 497–505, 2015.

[17] R. Kumar, K. K. Ravulakollu, and R. Bhat, "Fuzzy-Membership Based Author Recognition from Handwritten Devnagari Script," *Journal of Information Processing Systems,*2015.

[18] A. Materka and M. Strzelecki, "Texture Analysis Methods – A Review," *Technical University of Lodz,* vol. 11, pp. 1–33, 1998.

[19] M. M. Trivedi, C. a. Harlow, R. W. Conners, and S. Goh, "Object detection based on gray level cooccurrence," *Computer Vision, Graphics, and Image Processing,* vol. 28, no. 2, pp. 199–219, 1984.

[20] L. H. Siew, R. M. Hodgson, and E. J. Wood, "Texture Measures for Carpet Wear Assessment.," *IEEE Transaction on Pattern Analysis Machine Intelligence,* vol. 10, no. 1, pp. 92–105, 1988.

[21] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121–167, 1998.

[22] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "Comparative Study of Texture Measures for Terrain Classification.," *IEEE Transaction System Man Cybernetics,* vol. SMC-6, no. 4, pp. 269–285, 1976.

[23] T. O. Method, “Automatic Thresholding,” *International Conference on Circuits and Systems,* pp. 325–327, 1991.

[24] J. Zupan, “Introduction to artificial neural network (ANN) methods: what they are and how to use them,” *Acta Chimica Slovenica*, no. 2, September, pp. 327–352, 1994.

[25] M. Iizuka, “Quantitative evaluation of similar images with quasi-gray levels,” *Computer Vision, Graphics and Image Processing.*, vol. 38, no. 3, pp. 342–360, 1987.

[26] D. He, L. Wang, and J. Guibert, “Texture feature extraction,” *Pattern Recognition,* vol. 6, no. September, pp. 269–273, 1987.

[27] R. W. Conners and C. a Harlow, “A theoretical comparison of texture algorithms.,” *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 204–222, 1980.

[28] M. Porat and Y. Y. Zeevi, “Generalized Gabor Scheme of Image Representation in Biological and Machine Vision.,” *IEEE Transaction Pattern Analysis Machine Intelligence*, vol. 10, no. 4, pp. 452–468, 1988.

[29] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transaction System Man. Cybernetics*, vol. 3, no. 6, 1973.

[30] J. G. Daugman, “Unsurety relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters.,” *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1160–1169, 1985.

[31] Y. Hamamoto, S. Uchimura, M. Watanabe, T. Yasuda, Y. Mitani, and S. Tomita, "A gabor filter-based method for recognizing handwritten numerals," *Pattern Recognition Society*, vol. 31, no. 4, pp. 395–400, 1998.

[32] Shapiro and Stockman, "Chapter 7: Texture," *Computer Vision*, 2000.

[33] R. M. Haralick and K. Shanmugam, "Computer Classification of Reservoir Sandstones," *IEEE Transaction on Geoscience and Electronics system*, vol. 11, no. 4, 1973.

[34] F. Albregtsen, "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," *Image Processing Laboratory Department of Informatics, University of Oslo*, pp. 1–14, 2008.

[35] D. Gabor, "Theory of Communication," *Journal Iee*, vol. 93, no. 26. pp. 429–457, 1946.

[36] V. Cheung and K. Cannons, "An Introduction to Probabilistic Neural Networks," *Signal and Data Compression Laboratory University of Manitoba*, pp. 1–19, 2002.

[37] J. C. Bezdek, S. K. Chuah, and D. Leep, "Generalized k-nearest neighbor principles," *Pattern Regognition*, vol. 18, no. 3, pp. 237–256, 1986.

[38] Gaussian smoothing, educational website [Online]. Available at: http://homepages.inf.ed.ac.uk/rbf/HIPR2/gsmooth.htm

[39] Devanagari script, Wikipedia website [Online]. Available at: https://en.wikipedia.org/wiki/Devanagari.

[40] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Transaction System and Man, and Cybernetics*, vol. 3, no. 6, 1973.

[41] Gray-level co-occurrence matrix (GLCM), MathWorks website [Online]. Available at: http://in.mathworks.com/help/images/gray-level-co-occurrence-matrix-glcm.html.

[42] R. Manmatha and L. Rothfeder, “A scale space approach for automatically segmentation words from historical handwritten documents”, *IEEE Transaction On PAMI*, vol. 27,no.8,pp. 1212-1225, 2005.

[43] S. Peake and T.N. Tan, “Script and language recognition from document images”,*In Proc. of the BMVC* '97, vol. 2, pp. 610-619, 1997.

[44] K-nearest neighbor (K-NN), Wikipedia website [Online]. Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[45] M. Roumi, “MSc THESIS Implementing Texture Feature Extraction Algorithms on FPGA,” *Electrical Engineering*, 2009.

[46] A. Nosary, L. Heutte, T. Paquet, and Y. Lecourtier, “Defining author’s invariants to adapt the recognition task,” *Proc. Fifth International Conference on Document Analysis Recognition. ICDAR ’99 (Cat. No.PR00318)*, 1999.

[47] A. Bensefia, A. Nosary, T. Paquet, L. Heutte, U. F. R. Sciences, U. De Rouen, and F.-M. Cedex, “Author Recognition By Author ’ s Invariants,”*Pattern Recognition* pp. 274–279, 2002.

[48] T. D. E. Doctorat and A. Seropian, “Analyse de Document et Recognition de Scripteurs Résumé,” 2003.

[49] Andreas Schlapbach, "Author Recognition and Verification". *Berlin: Akademische Verlagsgesellschaft Aka GmbH*, 2008.

[50] R. Plamondon and G. Lorette, “Automatic signature verification and author recognition — the state of the art,” *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989.