ΟΠΤΙΚΗ ΑΝΑΓΝΩΡΙΣΗ ΧΑΡΑΚΤΗΡΩΝ

Η Διπλωματική Εργασία
παρουσιάστηκε ενώπιον
του Διδακτικού Προσωπικού του
Πανεπιστημίου Αιγαίου

Σε Μερική Εκπλήρωση
των Απαιτήσεων για το Μεταπτυχιακό Δίπλωμα του
Μηχανικού Πληροφοριακών και Επικοινωνιακών Συστημάτων

της

ΔΟΥΛΓΕΡΗ ΝΙΚΟΛΕΤΑΣ

ΚΑΡΛΟΒΑΣΙ-ΙΟΥΝΙΟΣ 2007

Η ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΔΙΔΑΣΚΟΝΤΩΝ ΕΠΙΚΥΡΩΝΕΙ
ΤΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ ΔΟΥΛΓΕΡΗ ΝΙΚΟΛΕΤΑΣ:

ΚΑΒΑΛΛΙΕΡΑΤΟΥ ΕΡΓΙΝΑ , Επιβλέπων
Ημερομηνία 22 ΙΟΥΝΙΟΥ 2007
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΣΤΑΜΑΤΑΤΟΣ ΕΥΣΤΑΘΙΟΣ, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΣΤΕΡΓΙΟΥ ΚΩΝΣΤΑΝΤΙΝΟΣ, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
2007

ΠΕΡΙΛΗΨΗ

Η οπτική αναγνώριση χαρακτήρων αποτελούσε από παλιά πρόκληση στους ερευνητές. Ειδικά στις μέρες μας με την πληθώρα των πληροφοριών που διακινούνται μέσω διαδικτύου, η ανάγκη για γρήγορες και αξιόπιστες μεθόδους που να ψηφιοποιούν τις πληροφορίες αυτές είναι μεγάλη. Ιστορικά έγγραφα που βρίσκονται σε βιβλιοθήκες θα μπορούσαν αν μετατραπούν σε ψηφιακή μορφή να γίνουν προσιτά από όλους, ενώ φόρμες συμπλήρωσης, προερχόμενες από διάφορες υπηρεσίες αν ψηφιοποιούνταν θα μπορούσαν να διευκολύνουν τις διάφορες εργασίες.

Έτσι στα πλαίσια της εργασίας αυτής, ασχοληθήκαμε με το πεδίο της οπτικής αναγνώρισης ολόκληρης λέξης, παραλείποντας το στάδιο της κατάτμησής της σε χαρακτήρες και την μετέπειτα αναγνώρισή τους. Συζητήθηκε η τεχνική αναγνώρισης που ακολουθήσαμε, που περιλαμβάνει ένα στάδιο εξαγωγής μορφολογικών χαρακτηριστικών των εικόνων-λέξεων, τέτοιων που να περιγράφουν το σχήμα της και στη συνέχεια την κατηγοριοποίησή τους με βάση κάποιον αλγόριθμο κατηγοριοποίησης, τον k-Means.

Σημαντική διαφοροποίηση με την υπάρχουσα βιβλιογραφία αποτέλεσε η μέθοδος που χρησιμοποιήσαμε για την κανονικοποίηση κατά μήκος, όπως και το γεγονός ότι ασχοληθήκαμε με την αναγνώριση κειμένου που προέρχεται από περισσότερους από έναν συγγραφείς. Πρέπει να τονιστεί ακόμα ότι για πρώτη φορά γίνεται απόπειρα να χρησιμοποιηθεί μέθοδος για αναγνώριση τυπωμένου και χειρόγραφου κειμένου μαζί, χωρίς να απαιτούνται ιδιαίτερα δεδομένα για εκπαίδευση.

ABSTRACT

Optical Character Recognition (OCR) has long been a challenge for researchers. Today the internet has revolutionized the way information is shared and accessed. However a huge volume of information contained in historical and other documents still lies in libraries, inaccessible to the public. Digitizing this information would not only make it widely available but also secure it for next generations. The same thing could happen with different kinds of information that can be distributed over the internet. This gives even greater value to fast, reliable methods for performing this task.

This study works on this field, dealing in particular with optical word recognition. We omit the quite common stage of segmenting the word into characters to be recognized one by one. Instead it proposes a technique for word recognition which consists of two stages. The first stage is the structural feature extraction from a word-image. These features describe the shape of the word. The second stage classifies all the words-images that need to be recognized, using the k-Means algorithm.

Significant differences from the existing methods found in the bibliography are the proposed method for normalizing the features, i.e. the Interpolation method, and the fact that we deal with documents from more than one writer. It should be emphasized that it is the first time an attempt is made to recognize combined printed and handwritten text, without requiring special training data.

In the next chapters we will see the proposed system analytically. Chapter one is an introduction to Optical Character Recognition Systems and chapter two covers some theory beyond our system. In chapter three we describe our system in detail and in the next chapter (chapter four) we present the experiments we performed with different kinds of documents. Finally in chapter five we discuss the conclusions we have reached about our system and suggest some future work in this direction in chapter six.

DOULGERI NIKOLETA
Department of Information and Communication Systems Engineering
UNIVERSITY OF THE AEGEAN

ΕΥΧΑΡΙΣΤΙΕΣ - ΑΦΙΕΡΩΣΕΙΣ

Στο σημείο αυτό θα ήθελα να ευχαριστήσω όλους όσους βοήθησαν άμεσα ή έμμεσα στην υλοποίηση της συγκεκριμένης εργασίας.

Πιο συγκεκριμένα την κα Καβαλλιεράτου Εργίνα που με καθοδήγησε σε όλα τα βήματα της εργασίας αυτής και η βοήθεια της ξεπέρασε κάθε προσδοκία. Οι προτάσεις και παρατηρήσεις της φάνηκαν ανεκτίμητες, ενώ παλαιότερη δουλειά της σε αυτόν τον τομέα φάνηκε πολύτιμη.

Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου και όλους τους στενούς μου φίλους σε Σάμο, Θεσσαλονίκη και Αθήνα, που μου συμπαραστάθηκαν όλο αυτό το διάστημα και με υποστήριξαν πλήρως ακόμα και τις δύσκολες ώρες.

Τέλος θα ήθελα να ευχαριστήσω τον διευθυντή και τους συναδέλφους μου στο σχολείο που με βοήθησαν σε θέματα ωραρίου και αδειών.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ....iii
ABSTRACT....iv
ΕΥΧΑΡΙΣΤΙΕΣ - ΑΦΙΕΡΩΣΕΙΣ....v
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ....vi
Κεφάλαιο 1
Εισαγωγή....1
1.1 Σύντομη Παρουσίαση της Τεχνικής μας....6
1.2 Τεχνικά Χαρακτηριστικά....7
1.3 Δομή Εργασίας....7
Κεφάλαιο 2
Θεωρητικό Υπόβαθρο....9
2.1 Εισαγωγή....9
2.2 Ιστογράμματα....9
2.3 Προφίλ....11
2.4 Μέθοδος Παρεμβολής (Interpolation Method)....12
2.5 k-Means αλγόριθμος....16
Κεφάλαιο 3
Περιγραφή Υποσυστήματος....19
3.1 Εισαγωγή....19
3.2 Αναγνώριση Λέξης....20
3.2.1 Προεπεξεργασία - Καθαρισμός λέξης....22
3.2.2 Εξαγωγή Χαρακτηριστικών....24
3.2.2.1 Κάθετο Ιστόγραμμα....25
3.2.2.2 Προφίλ....26
3.2.3 Κανονικοποίηση κατά ύψος....29
3.2.4 Κανονικοποίηση κατά μήκος με Παρεμβολή....31
3.2.5 Επιπλέον βελτίωση....33
3.2.5.1 Πάνω ουρές....36
3.2.5.2 Κάτω ουρές....38
3.2.5.3 Εξομάλυνση....40
3.2.6 Κατηγοριοποίηση με k-Means....43
3.3 Το υποσύστημά μας....45
Κεφάλαιο 4
Πειράματα....47
4.1 Εισαγωγή....47

4.2 Πειραματικά Δεδομένα 47
4.3 Παράμετροι 49
4.3.1 Πλήθος Δεδομένων Εκπαίδευσης 49
4.3.2 Μέγεθος Παρεμβολής 49
4.3.3 Εξομάλυνση 50
4.3.4 Αριθμός Συγγραφέων 50
4.4 Μέτρο αξιολόγησης 50
4.5 Περιγραφή Πειραμάτων 51
4.5.1 Εκπαίδευση και Έλεγχος σε Ιστορικό Ενός Συγγραφέα 51
4.5.2 Εκπαίδευση Τυπωμένο και Έλεγχος σε Ιστορικό 54
4.5.2.1 Ιστορικό Αγγλικό 54
4.5.2.2 Ιστορικό Ελληνικό 58
4.5.3 Εκπαίδευση Τυπωμένο και Έλεγχος σε Χειρόγραφο Διαφορετικών Συγγραφέων 63
Κεφάλαιο 5
Συμπεράσματα 69
Κεφάλαιο 6
Μελλοντική Εργασία 73
Κεφάλαιο 7
Βιβλιογραφία 75
Κατάλογος Πινάκων 79
Κατάλογος Σχημάτων και Διαγραμμάτων 81

Κεφάλαιο 1

Εισαγωγή

Πολλές χρήσιμες πληροφορίες, που αποτελούν κομμάτι της πολιτιστικής κληρονομιάς μιας χώρας και προάγουν την γνώση και την έρευνα, συναντάμε σε διάφορα ιστορικά έγγραφα, τόσο χειρόγραφα όσο και δακτυλογραφημένα. Ιστορικοί, μελετητές, εκπαιδευτικοί, αλλά και άλλες ομάδες ανθρώπων θέλουν να έχουν πρόσβαση σε αυτές τις πληροφορίες.

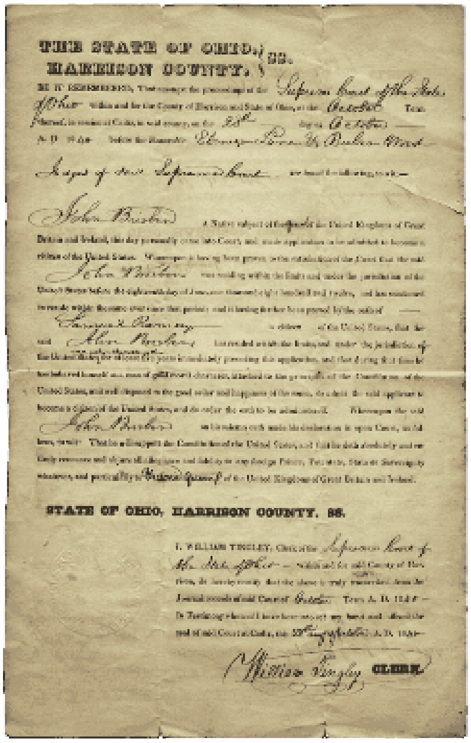
Σε μια προσπάθεια διατήρησης των εγγράφων αυτών, αλλά και συντήρησής τους, κάποιοι άνθρωποι άρχισαν το δύσκολο έργο της ψηφιοποίησης τους σε εικόνες εγγράφων, ώστε να είναι άμεσα διαθέσιμα, εύκολα ανταλλάξιμα και καταλαμβάνοντας πολύ λιγότερο χώρο σε σχέση με την πραγματική τους μορφή. Είναι όμως η μορφή αυτή κατάλληλη για την χρήση και αξιοποίηση των πληροφοριών των εγγράφων από όλους;

Συνήθως οι εικόνες εγγράφων είναι δύσκολο να διαβαστούν από ανθρώπους μη ειδικούς. Αυτό συμβαίνει εξαιτίας της ύπαρξης σημαδιών λόγω παλαιότητας, κακής ποιότητας χρωμάτων, σκισμένων άκρων ή και εξαιτίας του δυσανάγνωστου γραφικού χαρακτήρα εκείνων των εποχών.

Επίσης υπάρχουν σε διάφορα γραφεία και εταιρείες γενικότερα φόρμες προς συμπλήρωση που αφορούν διάφορα θέματα. Από τις φόρμες αυτές μπορεί να γίνονται παραγγελίες προϊόντων, στατιστικές έρευνες κα. Αν μπορούσαν οι απαντήσεις να περαστούν απευθείας σε ηλεκτρονικό υπολογιστή η δουλειά πολλών ανθρώπων θα διευκολύνονταν και θα γινόταν πιο γρήγορα και χωρίς κόπο.

Επίσης η δεικτοδότηση και η αναζήτηση σε ψηφιοποιημένα έγγραφα είναι μια δύσκολη αλλά και απαραίτητη εργασία σε βιβλιοθήκες, σχολεία, υπουργεία, συμβολαιογραφία και σε άλλους φορείς. Γρήγορες και αξιόπιστες μέθοδοι για το σκοπό αυτό θα έκαναν την εργασία πολλών ανθρώπων πιο εύκολη και γρήγορη.

Για παράδειγμα στο σχήμα 1.1α βλέπουμε ένα ιστορικό έγγραφο τυπωμένο, που αφορά την βράβευση ενός πολίτη του John Brisben και είναι κρατικό έγγραφο, στο σχήμα 1.1β ένα ιστορικό χειρόγραφο έγγραφο, που αφορά την Margaret Academy και στο σχήμα 1.1γ βλέπουμε μια συμπληρωμένη φόρμα για επιστροφή χρημάτων.

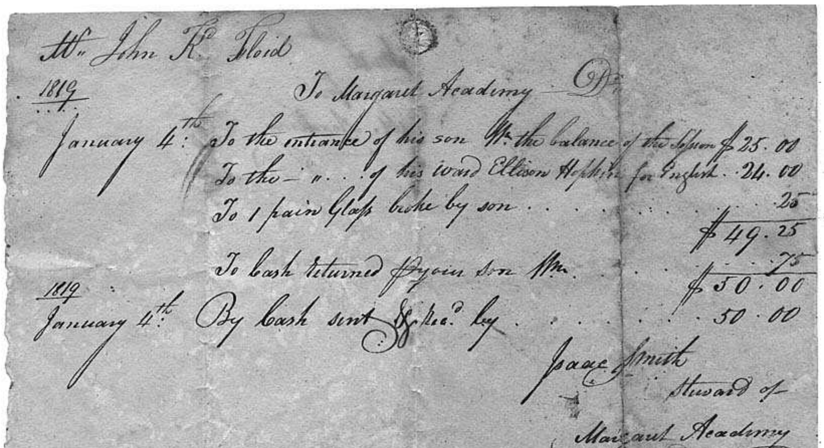

THE STATE OF OHIO, } SS.
HARRISON COUNTY.

John Brisben

STATE OF OHIO, HARRISON COUNTY, SS.

William Tingley CLERK.

(α)


Mr. John R. Floid

1819 To Margaret Academy Dr.

January 4th To the entrance of his son Wm. the balance of the session $25.00
To the — " . . of his ward Ellison Hopkins for English 24.00
To 1 pain Glass broke by son .25
$49.25
To cash returned by your son Wm. .75
$50.00
1819
January 4th By Cash sent [illegible] by 50.00

Isaac Smith
Steward of
Margaret Academy

(β)


WELCOME TO BEST BUY #147
CARY, NC 27511
(919)851-1868

0147 004 7008 02/02/05 18:51 0506024

REBATE FORM

OFFER: 50165

Name: Marshall Brain
Address: 106 Ashwyn Ct.
City: Cary
State: NC
Zip: 27511
Daytime Phone: 919-858-1002

BRAND: Sony
MODEL: DRU710A
SKU: 6782902

REBATE: $30.00

OFFER DATE: 08/13/2004 TO 02/25/2005
POSTMARK DATE: 30 Days From Purchase.

PROOF OF PURCHASE REQUIREMENTS:
This Rebate Form
Copy of Receipt
Copy of entire UPC from box
Serial Number
#

(γ)

Σχήμα 1.1: Διάφορα είδη εγγράφων

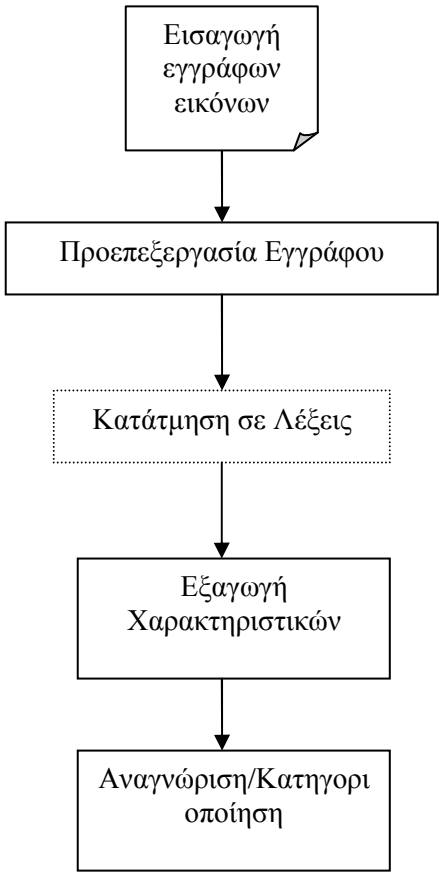
Για να μπορέσουμε να εξάγουμε αυτές τις πληροφορίες και να τις χρησιμοποιήσουμε είναι απαραίτητη η μετατροπή των εικόνων αυτών σε κείμενο, το οποίο είναι επεξεργάσιμο και ευανάγνωστο. Η αναγκαιότητα ύπαρξης τέτοιων συστημάτων συναντάται καθημερινά. Αυτή η διαδικασία είναι γνωστή ως οπτική αναγνώριση χαρακτήρων (optical character recognition – OCR).

Πολλοί έχουν ασχοληθεί με αυτό το ζήτημα και μπορούμε να πούμε ότι ο τομέας της αναγνώρισης λέξεων έχει κάνει σημαντική πρόοδο τις τελευταίες δεκαετίες. Μια πρώτη εμφάνιση της οπτικής αναγνώρισης χαρακτήρων συναντάμε το 1870, ως παροχή βοήθειας για οπτικά ανάπηρους και μια πρώτη επιτυχημένη προσπάθεια πραγματοποιείται από έναν Ρώσο επιστήμονα τον Tyurin το 1900. Οι πιο μοντέρνες μορφές OCR εμφανίζονται με την ανάπτυξη των ψηφιακών υπολογιστών στα μέσα της δεκαετίας του ’40, ενώ μια δεκαετία αργότερα (’50) οι πρώτες OCR μηχανές είναι διαθέσιμες στο εμπόριο. [28]

Αρχικά οι μεθοδολογίες περιλάμβαναν την αναγνώριση τυπωμένων αριθμών και περιορισμένου αριθμού αγγλικών γραμμάτων, ενώ τα τελευταία χρόνια περιλαμβάνει την αναγνώριση πολύπλοκων μορφών κειμένου, όπως χειρόγραφο, κείμενο με σύμβολα, γράμματα πιο πολύπλοκων γλωσσών, όπως κινέζικα [30], αραβικά [29], αρχαία ελληνικά [25] κα.

Τις τελευταίες δεκαετίες πολλές μέθοδοι, όπως και εφαρμογές αναπτύσσονται για τον σκοπό αυτό. Πληθώρα βιβλίων, αναφορών και εργασιών παρουσιάζονται στα συνέδρια του Pattern Recognition, του International System, Man and Cybernetics, της Ανάλυσης Εικόνας κα. και παρά το γεγονός ότι όλοι βλέπουν το πρόβλημα από διαφορετικές σκοπιές, ακολουθούν μια κοινή δομή.

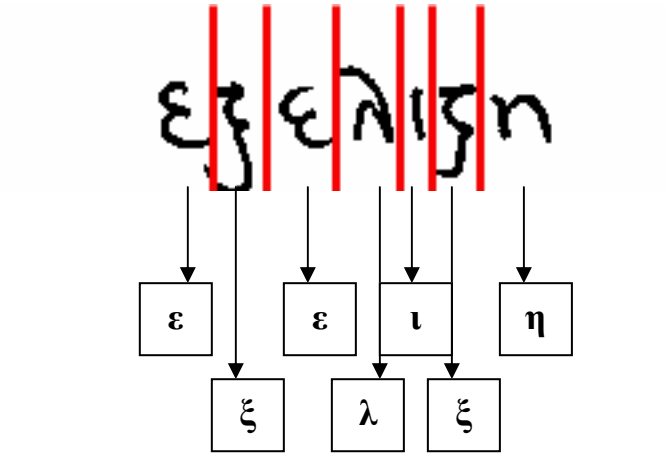
Συνήθως τέτοια συστήματα έχουν τέσσερα βασικά στάδια, όπως βλέπουμε και στο σχήμα 1.2: Προεργασία των εικόνων-λέξεων (preprocessing), μια φάση κατάτμησης των λέξεων σε χαρακτήρες (segmentation), η οποία μπορεί να παραλείπεται (segmentation-free), εξαγωγή χαρακτηριστικών των εικόνων-λέξεων (feature extraction) και η τελική αναγνώριση (recognition).



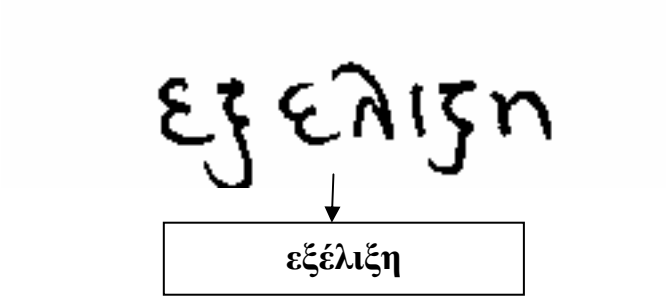
Σχήμα 1.2: Γενικό Σύστημα Οπτικής Αναγνώρισης Χαρακτήρων

Το πρώτο στάδιο, η προεργασία, περιλαμβάνει συνήθως εργασίες όπως η διόρθωση γωνίας που περιστρέφει την λέξη μέχρι να γίνει οριζόντια, η διόρθωση της κλίσης χαρακτήρων, η κανονικοποίηση, η μείωση θορύβου, η εύρεση γραμμής αναφοράς κτλ. [1]

Στο δεύτερο στάδιο συναντάμε δύο προσεγγίσεις: Από τη μια μεριά υπάρχουν οι μέθοδοι που προσπαθούν να κατατμήσουν την εικόνα-λέξη σε μικρότερα μέρη (σχήμα 1.3). Συνήθως τα κομμάτια αυτά γίνεται προσπάθεια να αντιστοιχούν σε χαρακτήρες, οι οποίοι πρόκειται να αναγνωριστούν ένας ένας. Αυτές οι μέθοδοι καλούνται segmentation-based [2][3][4]. Από την άλλη όταν τα χαρακτηριστικά που πρόκειται να εξαχθούν στο επόμενο βήμα είναι καθολικά, δηλαδή είναι χαρακτηριστικά που αφορούν ολόκληρη την εικόνα-λέξη, η κατάτμηση μπορεί να παραλείπεται και μιλάμε για segmentation-free μεθόδους (σχήμα 1.4). Αυτές οι τελευταίες κερδίζουν σημαντικό έδαφος τα τελευταία χρόνια [5][6][7][10].



Σχήμα 1.3: Μέθοδος Αναγνώρισης με Κατάτμηση



Σχήμα 1.4: Μέθοδος Αναγνώρισης χωρίς Κατάτμηση

Στην πρώτη περίπτωση, segmentation-based μέθοδοι, έχουν γίνει πολλές προσπάθειες ώστε η κατάτμηση σε χαρακτήρες να έχει καλά αποτελέσματα και εξάγονται χαρακτηριστικά για κάθε ένα χαρακτήρα χωριστά. Τα χαρακτηριστικά αυτά μπορούν να χωριστούν σε δυο κατηγορίες, σύμφωνα με τους [28]:

- σε μορφολογικά χαρακτηριστικά (structural features), όπως τομές τμημάτων γραμμών, strokes, τελικά σημεία κτλ και
- σε στατιστικά χαρακτηριστικά (statistical features), που προέρχονται από την στατιστική κατανομή.

Και οι δυο κατηγορίες πάντως συμπληρώνουν η μία την άλλη, αφού εστιάζουν σε διαφορετικές ιδιότητες των χαρακτήρων και προσεγγίσεις που συνδυάζουν και τις δυο προσφέρουν μια ικανοποιητική και πλήρη περιγραφή των χαρακτήρων.

Στην δεύτερη περίπτωση, segmentation-free μέθοδοι, δεν γίνεται καμιά προσπάθεια να χωριστεί η εικόνα-λέξη σε κομμάτια που σχετίζονται με τον χαρακτήρα και αντί για αναγνώριση ανά γράμμα, γίνεται προσπάθεια να αναγνωριστεί ολόκληρη η λέξη. Παίρνουμε λοιπόν κάποια γενικά ή ειδικά χαρακτηριστικά της εικόνας από τα οποία θα σχηματίσουμε μια περιγραφή της. Αυτό γίνεται αναζητώντας μια λέξη που να έχει την πιο όμοια περιγραφή με αυτήν που πήραμε από την εικόνα-λέξη.

Τέτοιες περιγραφές εικόνων μπορούν να διαχωριστούν σε τρεις κατηγορίες σύμφωνα με τους [1]:

- χαμηλού επιπέδου χαρακτηριστικά
- μεσαίου επιπέδου χαρακτηριστικά

- υψηλού επιπέδου χαρακτηριστικά

Οι δυο πρώτες κατηγορίες αναφέρονται σε περιγραφές που στηρίζονται σε χαμηλού επιπέδου χαρακτηριστικά, όπως εντοπισμός ιχνών του περιγράμματος της λέξης της εικόνας, εύρεση των ακμών ενός πιθανού πολυγώνου του σκελετού της εικόνας κα, με τη διαφορά ότι στην δεύτερη κατηγορία τα χαρακτηριστικά χρησιμοποιούνται για να δημιουργηθούν patterns που θεωρούνται πρωταρχικά. Στην τρίτη κατηγορία χρησιμοποιούνται υψηλού επιπέδου χαρακτηριστικά για να περιγράψουν την εικόνα, τα ολιστικά χαρακτηριστικά (holistic), όπως ουρές πάνω και κάτω, τελείες, κτλ που είναι ανεξάρτητα από το στυλ γραφής που έχει χρησιμοποιηθεί. Για αυτό και οι μέθοδοι αυτοί καλούνται και ολιστικές μέθοδοι (holistic methods).

Στο τελευταίο στάδιο ενός OCR συστήματος γίνεται η αναγνώριση της λέξης και η μετατροπή της σε κείμενο. Ουσιαστικά αυτό που γίνεται εδώ είναι η κατηγοριοποίηση είτε των μεμονωμένων χαρακτήρων είτε ολόκληρων των λέξεων σε κλάσεις, που τις εκπαιδεύουμε για να μπορούν να αναγνωρίσουν και άλλες όταν εισαχθούν στο σύστημα.

Σύμφωνα με τους [31] μπορούμε να διακρίνουμε δύο μορφές κατηγοριοποίησης:

- την στατιστική προσέγγιση, που περιλαμβάνει την αναπαράσταση ενός pattern σαν συγκεκριμένου μήκους, διατεταγμένη λίστα τιμών και
- την μορφολογική προσέγγιση που περιγράφει το pattern σαν μη-διατεταγμένη, μεταβλητού μήκους λίστα από απλά σχήματα.

Η δεύτερη προσέγγιση είναι πιο κοντά στον τρόπο που ο άνθρωπος αναγνωρίζει, αλλά είναι πιο δύσκολο να υλοποιηθεί. Μπορεί να χειριστεί καλά μορφολογικά χαρακτηριστικά, αλλά όχι στατιστικά, και από κάποιους θεωρείται καλύτερη. Η πρώτη προσέγγιση χειρίζεται εξίσου καλά μορφολογικά και στατιστικά χαρακτηριστικά.

Τα συστήματα αυτά χρησιμοποιούνται για να αναγνωρίσουν διάφορα είδη κειμένου, όπως απλό τυπωμένο, τυπωμένο ιστορικό, χειρόγραφο, που μπορεί να είναι μεμονωμένων χαρακτήρων, συνεχόμενων χαρακτήρων ή και γραφής χωρίς περιορισμούς.

Η πληθώρα των πληροφοριών, που βρίσκεται σε βιβλία και σε βιβλιοθήκες ανά τον κόσμο, κάνει απαραίτητη την ψηφιοποίηση τους, ώστε να είναι εύκολα προσβάσιμες από όλους. Το ίδιο συμβαίνει και με πολλά χειρόγραφα που διατηρούνται.

Από την άλλη και η ανάγκη για μαζική επεξεργασία εγγράφων από διάφορους φορείς, όπως βιβλιοθήκες, υπουργεία, συμβολαιογραφεία, σχολεία κτλ και η εύκολη αναζήτηση και δεικτοδότησή τους, απαιτεί την μετατροπή τους σε ψηφιακή μορφή, που να είναι εύκολα επεξεργάσιμη, προσπελάσιμη και μεταφέρσιμη.

Ένα πάντως είναι σίγουρο, για οποιαδήποτε ανάγκη και με οποιονδήποτε τρόπο κι αν γίνεται η οπτική αναγνώριση των χαρακτήρων, είναι απαραίτητη για την διατήρηση της πολιτιστικής μας κληρονομιάς και της μετάδοσής της στις επόμενες γενεές. Ζώντας στην εποχή της πληροφοριακής επανάστασης, με την εξέλιξη των υπολογιστικών συστημάτων και των τηλεπικοινωνιών, η διακίνηση των πληροφοριών διευκολύνεται και σε αυτό πρέπει να βοηθήσει και ο τομέας της ανάλυσης εικόνας και οπτικής αναγνώρισης χαρακτήρων.

1.1 Σύντομη Παρουσίαση της Τεχνικής μας

Στα πλαίσια αυτής της εργασίας παρουσιάζεται ένα υποσύστημα για την αναγνώριση τυπωμένου ιστορικού κειμένου και χειρόγραφου κειμένου χωρίς περιορισμούς, για οποιονδήποτε αριθμό συγγραφέων. Στη συνέχεια θα γίνει μια σύντομη περιγραφή του υποσυστήματος αυτού, σε σχέση με το γενικό σύστημα του σχήματος 1.2.

Πιο συγκεκριμένα στο στάδιο της προεπεξεργασίας των εγγράφων, οι εικόνες λέξεις που εισάγονται στο σύστημα προέρχονται από το σύστημα των [13], έχοντας εφαρμόσει διόρθωση γωνίας εκτροπής εγγράφου, κατάτμηση σε γραμμές και κατάτμηση σε χαρακτήρες., χωρίς κάποια άλλη επεξεργασία. Όσον αφορά το δεύτερο στάδιο επιλέχτηκε η χρήση μεθόδου οπτικής αναγνώρισης ολόκληρης λέξης, δηλαδή ολιστικής μεθόδου και όχι κατάτμησής της σε χαρακτήρες και η μετέπειτα αναγνώρισή τους.

Έτσι λοιπόν στο υποσύστημά μας εισάγονται ολόκληρες εικόνες-λέξεις, οι οποίες πρέπει να τονιστεί ότι δεν έχουν υποστεί καμία επεξεργασία, όπως διόρθωση γωνίας εκτροπής, διόρθωση κλίσης χαρακτήρων κτλ.

Στη συνέχεια εξάγονται τα χαρακτηριστικά των εικόνων-λέξεων, που επιλέχτηκαν να είναι μορφολογικά, σε αντίθεση με τα στατιστικά, ώστε να αποφύγουμε πολύπλοκους υπολογισμούς, που απαιτούν τα δεύτερα. Έτσι χρησιμοποιήθηκαν χαρακτηριστικά που περιγράφουν το σχήμα ολόκληρης της λέξης.

Τέλος όσον αφορά το στάδιο της κατηγοριοποίησης και τελικής αναγνώρισης επιλέχτηκε η χρήση ενός αλγορίθμου κατηγοριοποίησης που βασίζεται στην ελαχιστοποίηση της ευκλείδειας απόστασης των χαρακτηριστικών των λέξεων αυτών από κάποιες αρχικές κλάσεις. Η μέθοδος αυτή αποφεύγει τους χρονοβόρους και πολύπλοκους υπολογισμούς, στηρίζεται όμως στην ύπαρξη αρχικού λεξιλογίου.

Βασικός στόχος μας ήταν η απλότητα και ταχύτητα του συστήματος, ενώ στις μεθόδους που προτείνονται προσπαθήσαμε να πλησιάσουμε τον ανθρώπινο τρόπο σκέψης στην αναγνώριση των λέξεων

1.2 Τεχνικά Χαρακτηριστικά

Όλες οι συναρτήσεις που περιγράφονται παρακάτω για την υλοποίηση των διάφορων σταδίων του υποσυστήματος που προτείνεται έχουν υλοποιηθεί σε MATLAB 6.5, θεωρώντας ότι είναι ένα πολύ καλό και εργαλείο για την επεξεργασία των εικόνων.

Ο υπολογιστής που χρησιμοποιήθηκε για τα διάφορα πειράματα ήταν Pentium Mobile στα 1.6GHz και με μνήμη 512MB, που κρίθηκε επαρκής για το πλήθος των εικόνων που χρησιμοποιήθηκαν.

Όλες οι εικόνες που εισάγονται στο σύστημα είναι ασπρόμαυρες BMP εικόνες, στα 300dpi, εκτός από τις χειρόγραφες εικόνες που είναι στα 200dpi.

Κατά την έξοδό του το υποσύστημα, αποθηκεύει σε αρχείο την κατηγοριοποίηση που προέκυψε. Δημιουργεί επίσης φακέλους για κάθε αρχική κλάση λέξεων και σε καθέναν από αυτούς τοποθετεί τις εικόνες ανάλογα με την κατηγοριοποίηση που έκανε.

1.3 Δομή Εργασίας

Στη συνέχεια θα γίνει αναλυτική περιγραφή των σταδίων του προτεινόμενου υποσυστήματος.

Ποιο συγκεκριμένα, στο δεύτερο κεφάλαιο θα γίνει μια παρουσίαση του θεωρητικού υπόβαθρου πάνω στο οποίο στηρίζεται η δική μας τεχνική. Θα γίνει αναλυτική περιγραφή των διαφόρων χαρακτηριστικών των εικόνων-λέξεων που χρησιμοποιούνται σε αντίστοιχα συστήματα. Επίσης θα δούμε μια μέθοδο για την μετατροπή διανυσμάτων στο ίδιο μέγεθος, ώστε να μπορούν να συγκριθούν και τα γενικά στοιχεία ενός αλγορίθμου για την κατηγοριοποίηση δεδομένων.

Στο κεφάλαιο τρία θα γίνει αναλυτική περιγραφή του υποσυστήματος που αναπτύξαμε. Αυτό περιλαμβάνει παρουσίαση όλων των σταδίων της τεχνικής αναγνώρισης λέξεων που προτείνουμε. Έτσι θα παρουσιαστούν ακριβώς τα χαρακτηριστικά που επιλέξαμε να χρησιμοποιηθούν στο υποσύστημα, δηλαδή το κάθετο ιστόγραμμα, το πάνω και κάτω προφίλ, καθώς και τα χαρακτηριστικά που χρησιμοποιήθηκαν για επιπλέον βελτίωση, όπως είναι οι πάνω και κάτω ουρές. Θα δούμε πώς εξάγουμε τα χαρακτηριστικά αυτά για κάθε εικόνα-λέξη και στη συνέχεια πώς τα κανονικοποιούμε κατά ύψος και μήκος, ώστε να είναι άμεσα συγκρίσιμα. Συγκεκριμένα θα γίνει αναλυτική περιγραφή του τρόπου εφαρμογής της μεθόδου της παρεμβολής για την κανονικοποίηση κατά μήκος. Ακόμα θα δούμε την χρήση του αλγορίθμου k-Means για την κατηγοριοποίηση των εικόνων-λέξεων σε κλάσεις. Τέλος θα δούμε την μέθοδο της εξομάλυνσης των τιμών των χαρακτηριστικών και πώς αυτή επηρεάζει τα αποτελέσματα του υποσυστήματος μας.

Στο τέταρτο κεφάλαιο θα παρουσιαστούν αναλυτικά οι κατηγορίες των πειραμάτων που πραγματοποιήθηκαν. Αρχικά θα περιγραφούν οι παράμετροι που παίζουν ρόλο στα πειράματά μας και στη συνέχεια θα δούμε τα ποσοστά επιτυχίας για τις περιπτώσεις μόνο ιστορικού τυπωμένου κειμένου, λέξεις εκπαίδευσης τυπωμένο κείμενο και λέξεις ελέγχου ιστορικό τυπωμένο για διάφορες γλώσσες και τέλος λέξεις εκπαίδευσης τυπωμένο κείμενο και λέξεις ελέγχου χειρόγραφο, πολλών συγγραφέων. Σε κάθε κατηγορία πειραμάτων θα δούμε την μεταβολή των παραμέτρων και πώς αυτή επηρεάζει το ποσοστό επιτυχίας, με κατάλληλους πίνακας και διαγράμματα κάθε φορά.

Στο πέμπτο κεφάλαιο θα παρουσιαστεί μια σύνοψη της συγκεκριμένης εργασίας και τα συμπεράσματα στα οποία καταλήξαμε μέσα από τα πειράματα.

Στο κεφάλαιο έξι θα γίνει μια σύντομη παρουσίαση της εργασίας και των βελτιώσεων που πρόκειται να πραγματοποιηθούν στο μέλλον.

Τέλος στο έβδομο κεφάλαιο θα παρουσιαστεί αναλυτικά η βιβλιογραφία στην οποία στηριχτήκαμε και χρησιμοποιήσαμε για την εκπόνηση του προτεινόμενου υπόσυστήματος.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή

Στο κεφάλαιο που θα ακολουθήσει θα γίνει μια αναλυτική περιγραφή κάποιων θεωρητικών στοιχείων, που είναι απαραίτητα για την κατανόηση της συγκεκριμένης εργασίας.

Πιο συγκεκριμένα θα ασχοληθούμε με τα χαρακτηριστικά που εξάγουμε από τις εικόνες-λέξεις, όπως είναι τα ιστογράμματα και τα προφίλ, στη συνέχεια θα δούμε μια μέθοδο κανονικοποίησης των τιμών των χαρακτηριστικών αυτών, την παρεμβολή (Interpolation Method), ώστε να μπορούμε να τα συγκρίνουμε και τέλος θα δούμε τον αλγόριθμο που χρησιμοποιήθηκε για την κατηγοριοποίηση των εικόνων-λέξεων, βάση των εξαγόμενων χαρακτηριστικών και κατά επέκταση για την αναγνώρισή τους.

2.2 Ιστογράμματα

Μια εικόνα μπορεί να οριστεί ως μια συνάρτηση *f(x,y)*, όπου τα *x* και *y* είναι χωρικές συντεταγμένες και η τιμή της *f* για κάθε ζεύγος συντεταγμένων καλείται ένταση της εικόνας στο σημείο αυτό. Για τις μονόχρωμες εικόνες οι τιμές που μπορεί να πάρει η συνάρτηση αυτή είναι 0 αν τα pixels είναι άσπρα και 1 αν τα pixels είναι μαύρα.

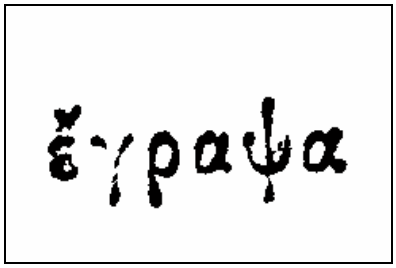
Ένα βασικό χαρακτηριστικό των εικόνων-εγγράφων, που συναντάται σε πολλές τεχνικές στα διάφορα στάδια της επεξεργασίας των εικόνων, είναι το οριζόντιο και το κάθετο ιστόγραμμα ή αλλιώς οριζόντια ή κάθετη προβολή *(horizontal/vertical histogram ή prozection profile).*

Αν θεωρήσουμε την εικόνα ως ένα πίνακα διαστάσεων MxN, ως κάθετο ιστόγραμμα ορίζεται το πλήθος των μαύρων pixels που συναντάμε σε κάθε στήλη της εικόνας, ενώ ως οριζόντιο ιστόγραμμα το πλήθος των μαύρων pixels που συναντάμε σε κάθε γραμμή της εικόνας.

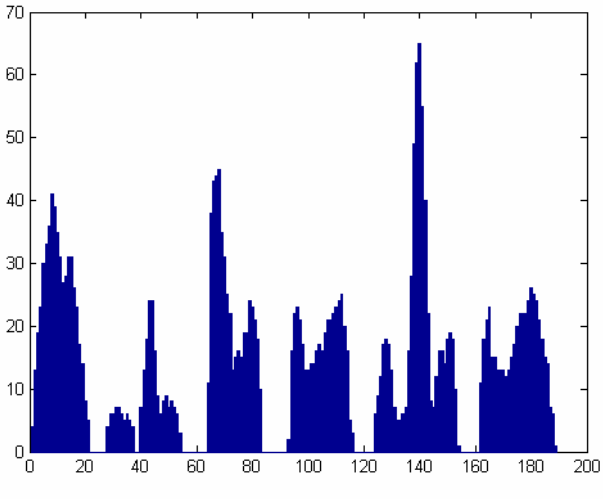
Στην βιβλιογραφία τα ιστογράμματα αυτά χρησιμοποιούνται κυρίως στο στάδιο της διόρθωσης της γωνίας εκτροπής του εγγράφου, όπως στις περιπτώσεις των [2], [11], [12], [14]. Σε κάποιες άλλες που προχωρούν σε κατάτμηση χαρακτήρων, χρησιμοποιούνται ως χαρακτηριστικά των χαρακτήρων αυτών στην προσπάθεια να αναγνωριστούν, όπως στους [2].

Σε πολλές, όμως, άλλες περιπτώσεις όπως στους [10], αλλά και στο δικό μας σύστημα θεωρήθηκε απαραίτητη η εξαγωγή του κάθετου ιστογράμματος για κάθε εικόνα-λέξη, ολόκληρη, ως ένα από τα χαρακτηριστικά που πρόκειται να συγκριθούν. Έτσι μπορούμε να έχουμε πληροφορία της κατανομής του μελανιού σε κάθε στήλη της εικόνας-λέξης, τιμή αρκετά χαρακτηριστική για κάθε εικόνα.

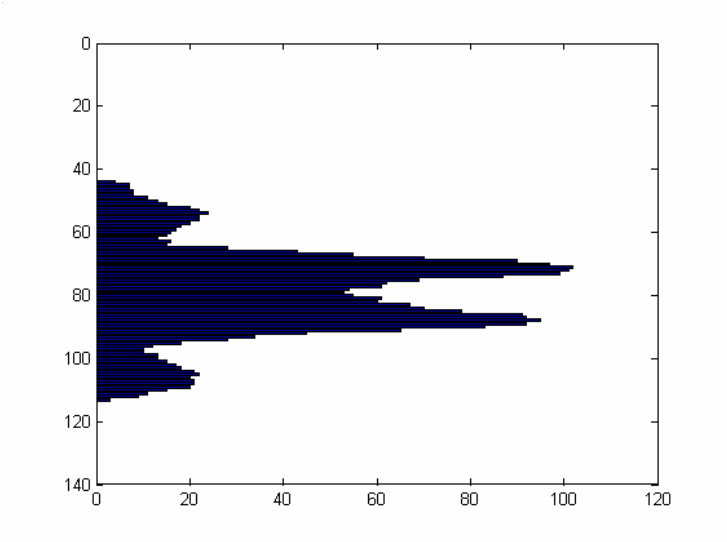
Παρακάτω στα σχήμα 2.1α βλέπουμε μια εικόνα-λέξη και τα αντίστοιχα κάθετο ιστόγραμμα της στο σχήμα 2.1β και οριζόντιο ιστόγραμμά της στο σχήμα 2.1γ.



(α) Λέξη



(β) Κάθετο Ιστόγραμμα



(γ) Οριζόντιο Ιστόγραμμα

Σχήμα 2.1: Κάθετο και Οριζόντιο Ιστόγραμμα Λέξης

2.3 Προφίλ

Τα προφίλ (*profiles*) είναι ένας βασικός τρόπος περιγραφής διαφόρων πραγμάτων στον κόσμο. Για παράδειγμα το προφίλ ενός ανθρώπου μπορεί να περιλαμβάνει το όνομά του, την ηλικία του, μια περιγραφή της εμφάνισής του, όπως επίσης και κάποιων πνευματικών χαρακτηριστικών του. Έχοντας στην κατοχή μας τα προφίλ διαφόρων ανθρώπων μπορούμε να δούμε ποιοι μοιάζουν μεταξύ τους, ποιοι διαφέρουν, ακόμα κι αν μιλάμε για το ίδιο άτομο.

Μια μέθοδος οπτικής αναγνώρισης χαρακτήρων ή λέξεων εκμεταλλεύεται τις γενικές ιδιότητες των προφίλ και τα χρησιμοποιεί για την αναγνώρισή τους. Αποκτούμε έτσι μια γενική αναπαράσταση του σχήματος των χαρακτήρων ή των λέξεων, έτοιμη να συγκριθεί με άλλες. Βέβαια στην βιβλιογραφία συναντάμε διάφορα είδη προφίλ, τα οποία χρησιμοποιούνται ανάλογα με τις εκάστοτε ανάγκες.

Θεωρώντας όπως προηγουμένως ότι η εικόνα είναι ένας πίνακας διαστάσεων MxN μπορούμε να εξάγουμε τους παρακάτω ορισμούς:

- *Πάνω προφίλ:* ορίζεται ως η θέση του πρώτου μαύρου pixel που συναντάμε σε κάθε στήλη της εικόνας, ξεκινώντας από την πρώτη γραμμή και διατρέχοντας την εικόνα προς τα κάτω
- *Κάτω προφίλ:* ορίζεται ως η θέση του πρώτου μαύρου pixel που συναντάμε σε κάθε στήλη της εικόνας, ξεκινώντας από την τελευταία γραμμή και διατρέχοντας την εικόνα προς τα πάνω
- *Αριστερό προφίλ:* ορίζεται ως η θέση του πρώτου μαύρου pixel που συναντάμε σε κάθε γραμμή της εικόνας, ξεκινώντας από την πρώτη στήλη και διατρέχοντας την εικόνα προς τα δεξιά
- *Δεξιό προφίλ:* ορίζεται ως η θέση του πρώτου μαύρου pixel που συναντάμε σε κάθε γραμμή της εικόνας, ξεκινώντας από την τελευταία στήλη και διατρέχοντας την εικόνα προς τα αριστερά

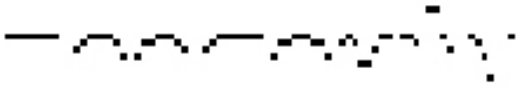
- *Ακτινικά προφίλ:* ορίζονται ως η θέση του πρώτου μαύρου pixel που συναντάμε αν κοιτάξουμε την εικόνα προς το κέντρο ή από το κέντρο για μια θέση ακτίνας

Έτσι για παράδειγμα οι [27] εισήγαγαν την έννοια του δεξιού και αριστερού προφίλ για την κατάτμηση και αναγνώριση χειρόγραφων συνδεδεμένων αριθμών, ενώ οι [14] χρησιμοποίησαν τα πάνω και κάτω προφίλ για την αναγνώριση χαρακτήρων, θεωρώντας ότι τα δυο προηγούμενα δεν είναι αρκετά για να περιγράψουν τις διαφοροποιήσεις των χαρακτήρων. Επίσης οι [13] χρησιμοποίησαν τα δεξί, αριστερό και ακτινικά προφίλ για την περιγραφή των χαρακτήρων και οι [10] προχώρησαν στην χρήση των πάνω και κάτω προφίλ για ολόκληρη την εικόνα-λέξη

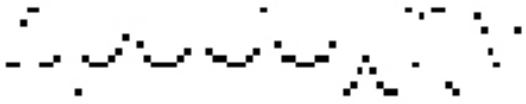
Στο δικό μας υποσύστημα θεωρήθηκε απαραίτητη η περιγραφή του σχήματος της εικόνας-λέξης, ως μέτρο σύγκρισης. Έτσι κατάλληλα προφίλ για τον σκοπό αυτό θεωρήθηκαν τα πάνω και κάτω προφίλ ολόκληρης της εικόνας-λέξης.

Στο σχήμα 2.2α βλέπουμε μια λέξη, ενώ στο σχήμα 2.2β και σχήμα 2.2γ το πάνω και κάτω προφίλ αντίστοιχα της λέξης.

(α) Λέξη



(β) Πάνω Προφίλ



(γ) Κάτω Προφίλ

Σχήμα 2.2: Διάφορα Είδη Προφίλ Λέξεων

2.4 Μέθοδος Παρεμβολής (Interpolation Method)

Η παρεμβολή (*interpolation*) είναι ουσιαστικά μια μέθοδος υπολογισμού ενδιάμεσων τιμών σε ένα σύνολο τιμών, κι όχι μόνο. Ο E. Whittaker, καθηγητής του Πανεπιστημίου του Εδιμβούργου από το 1913 παρατήρησε ότι «η πιο κοινή μορφή παρεμβολής προκύπτει όταν αναζητούμε δεδομένα από έναν πίνακα, ο οποίος δεν έχει τις ακριβής τιμές που θέλουμε».

Στην ιστορία η παρεμβολή έχει χρησιμοποιηθεί με ποικίλους τρόπους για πάρα πολλούς σκοπούς. Οι πρώτες ενδείξεις χρήσης της παρεμβολής προέρχονται από τους αρχαίους Βαβυλώνιους και Έλληνες, οι οποίοι το 300π.Χ δεν χρησιμοποιούσαν μόνο την απλή γραμμική παρεμβολή αλλά και πιο πολύπλοκες μορφές της, για να προβλέψουν τις θέσεις του ήλιου, του φεγγαριού και των ως τότε γνωστών πλανητών. Το 150π.Χ μάλιστα,

στην Ελλάδα ο Ίππαρχος της Ρόδου χρησιμοποίησε γραμμική παρεμβολή για να κατασκευάσει μια «αρμονική» συνάρτηση που υπολογίζει την θέση ουράνιων σωμάτων.

Στην Κίνα χρησιμοποιήθηκε μια μορφή της Gregory-Newton παρεμβολής για την κατασκευή ενός «Αυτοκρατορικού Ημερολογίου», ενώ στην Ινδία προτάθηκε μια μέθοδος παρεμβολής για άνισα-διαστήματα δεδομένων (unequal-interval data).

Μέσα στους αιώνες συναντάμε την παρεμβολή σε πολλές εφαρμογές, με αρκετά σημαντική την θαλάσσια πλοήγηση. Πίνακες από ειδικές τιμές συναρτήσεων υπήρχαν για τον υπολογισμό του γεωγραφικού πλάτους και μήκους, που ήταν δύσκολο να κατασκευαστούν. Το ίδιο συμβαίνει και με άλλους πίνακες ειδικών συναρτήσεων, όπως πίνακες με εκτίμηση ορίου ζωής κτλ. Στο πέρασμα των χρόνων πολλοί ήταν αυτοί που προσπάθησαν να λύσουν το πρόβλημα με διάφορους τρόπους. Τελικά σημαντική βελτίωση έγινε όπως ήταν φυσικό με την χρήση υπολογιστών, αλλά και μοντέρνων μεθόδων παρεμβολής. [17]

Τα τελευταία χρόνια, λοιπόν, με την αλματώδη ανάπτυξη της τεχνολογίας και την αυξημένη χρήση των υπολογιστών, η παρεμβολή όχι μόνο δεν ξεχάστηκε, αλλά κρίνεται απαραίτητο εργαλείο σε πολλές εφαρμογές επεξεργασίας σημάτων και ψηφιακών εικόνων [16], [18]. Αποτελεί μια από τις πιο γνωστές λύσεις σε πολλά προβλήματα για ένα πλήθος εφαρμογών, όπως είναι η μηχανική όραση, η ψηφιακή φωτογραφία, η επεξεργασία γραφικών, στο image calibration και registration, στην αναδειγματοληψία, στην αλλαγή κλίμακας κτλ, [15].

Όπως είδαμε υπάρχουν πολλές παραλλαγές της παρεμβολής και ανάλογα με την εφαρμογή στην οποία πρόκειται να χρησιμοποιηθεί επιλέγεται και η κατάλληλη. Έτσι συναντάμε την γραμμική παρεμβολή (linear interpolation), την δι-γραμμική παρεμβολή (bilinear interpolation), την bicubic παρεμβολή, την πολυωνυμική παρεμβολή, την spline παρεμβολή, την παρεμβολή πλησιέστερου γείτονα (nearest neighbor interpolation) κτλ.

Στο δικό μας υποσύστημα, αφού εξαχθούν τα διάφορα χαρακτηριστικά των εικόνων-λέξεων, δηλαδή το κάθετο ιστόγραμμα και το πάνω και κάτω προφίλ, ακολουθεί η σύγκριση των τιμών τους για τις διάφορες λέξεις. Επειδή όμως οι εικόνες-λέξεις μπορεί να προέρχονται από διαφορετικά έγγραφα και να έχουν διαφορετικά μεγέθη απαραίτητη είναι η αναγωγή των παραπάνω χαρακτηριστικών σε ένα κοινό μήκος, ώστε να μπορεί να γίνει η σύγκριση χωρίς λάθη

Συγκεκριμένα στο υποσύστημά μας επιλέχτηκε η γραμμική παρεμβολή αφού τα χαρακτηριστικά των λέξεων είναι μονοδιάστατα διανύσματα και θεωρείται μια από τις πιο γρήγορες μεθόδους με καλής ποιότητας αποτελέσματα. Μια υλοποίηση του αλγορίθμου της παρεμβολής σε MATLAB, που χρησιμοποιήσαμε φαίνεται σχήμα 2.3.

```
function g=Interpolation(Vector,NewSize);
[Rows,Columns]=size(Vector);
step=Columns/NewSize;
NewVector(1)=Vector(1);
position=1+step;
for i=2:NewSize-1
    Integer_part=floor(position);
    Decimal_part=position-Integer_part;
    if Integer_part<Columns
     if Decimal_part==0
```

```
        NewVector(i)=Vector(Integer_part);
     elseif Decimal_part<=0.5
        NewVector(i)=Decimal_part*Vector(Integer_part)+(1-
        Decimal_part)*Vector(Integer_part+1);
     else
        NewVector(i)=(1-Decimal_part)*
        Vector(Integer_part)+Decimal_part*
        Vector(Integer_part+1);
     end
     position = position+step;
    else
      NewVector(i)=Vector(Integer_part);
    end
end
NewVector(NewSize)=Vector(Columns);
g=NewVector;
```

Σχήμα 2.3: Αλγόριθμος Μεθόδου Παρεμβολής

Η παραπάνω συνάρτηση δέχεται σαν ορίσματα ένα διάνυσμα (Vector) και το νέο μέγεθος (NewSize) στο οποίο θέλουμε να μετατρέψουμε το διάνυσμα. Αυτό που πρόκειται να κάνουμε είναι να παρεμβάλουμε σημεία ανάμεσα στα υπάρχοντα του διανύσματος, ώστε να αλλάξουμε το μέγεθός του, είτε να το μεγαλώσουμε είτε να το μικρύνουμε.

Το βήμα με βάση το οποίο θα καθορίζεται η θέση των νέων σημείων είναι στην ουσία το αποτέλεσμα της διαίρεσης του παλιού μήκους (Columns) με το νέο μήκος. Έτσι για παράδειγμα αν έχουμε ένα διάνυσμα 15 σημείων *Vector=[3 5 2 6 8 9 4 6 3 7 8 2 4 6 7]* (σχήμα 2.6α) και θέλουμε να το μετατρέψουμε σε ένα διάνυσμα μήκους 25 το βήμα θα είναι 0.6.

Αρχίζουμε από το πρώτο σημείο (το οποίο θα είναι ίδιο και στο νέο διάνυσμα) και προσθέτουμε το βήμα, στο παράδειγμά μας το αποτέλεσμα θα είναι 1.6, άρα αυτή θα είναι και η θέση του νέου σημείου.

Η τιμή του τώρα θα υπολογιστεί από τα δυο γειτονικά του σημεία, στην προκειμένη περίπτωση το Σημείο1 και Σημείο2. Δανειζόμαστε και κάποια στοιχεία από την nearest neighbor παρεμβολή και τα δύο αυτά σημεία θα συμμετέχουν στην τιμή του νέου σημείου όχι ισάξια αλλά με κάποιο βάρος. Στην περίπτωση μας αφού έχουμε θέση 1.6, πιο κοντά στο Σημείο2, θα είναι πιο σωστό η τιμή του νέου σημείου να καθορίζεται περισσότερο από το Σημείο2 και λιγότερο από το Σημείο1. Έτσι έχουμε:

*Νέα Τιμή = 0.4 *Τιμή_ Σημείου1 +0.6 * Τιμή_Σημείου2*

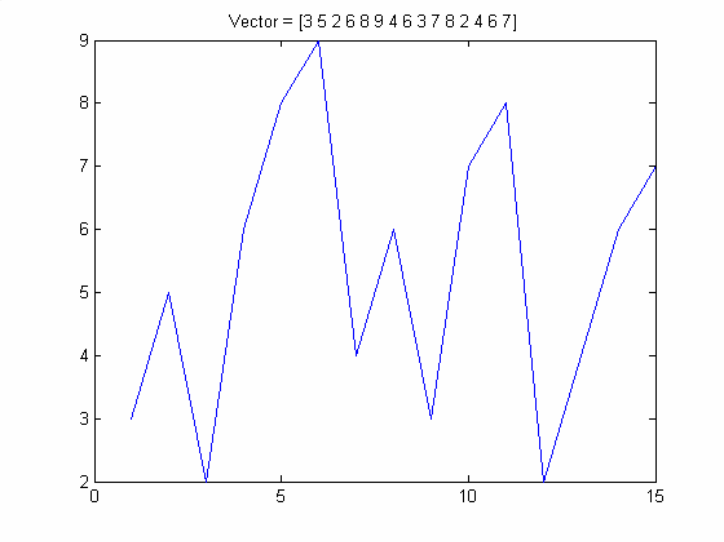
Αν η θέση ήταν 1.2, δηλαδή πιο κοντά στο Σημείο1, αυτό θα είχε μεγαλύτερη βαρύτητα σε σχέση με το Σημείο 2 και θα ίσχυε:

*Νέα Τιμή = 0.8 * Τιμή_Σημείου1 +0.2 * Τιμή_Σημείου2*

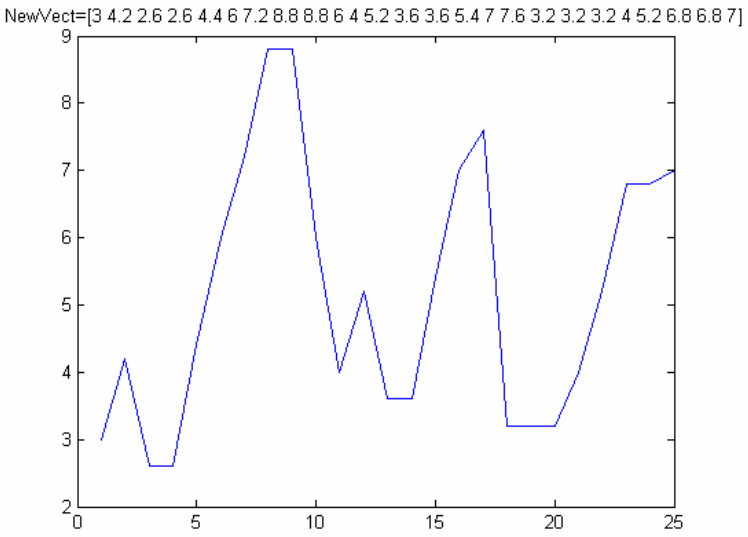
Έτσι συνεχίζουμε διατρέχοντας όλο το διάνυσμα και στο τέλος θα έχουμε σαν αποτέλεσμα ένα νέο διάνυσμα με το επιθυμητό μέγεθος. Στο σχήμα 2.4α βλέπουμε τη γραφική παράσταση του παραπάνω διανύσματος *Vector*, στο σχήμα 2.4β βλέπουμε το νέο

διάνυσμα που προκύπτει με την εφαρμογή του παραπάνω αλγόριθμου παρεμβολής και έχει μήκος 25 και στο 2.6γ βλέπουμε ένα νέο διάνυσμα μήκους 10.

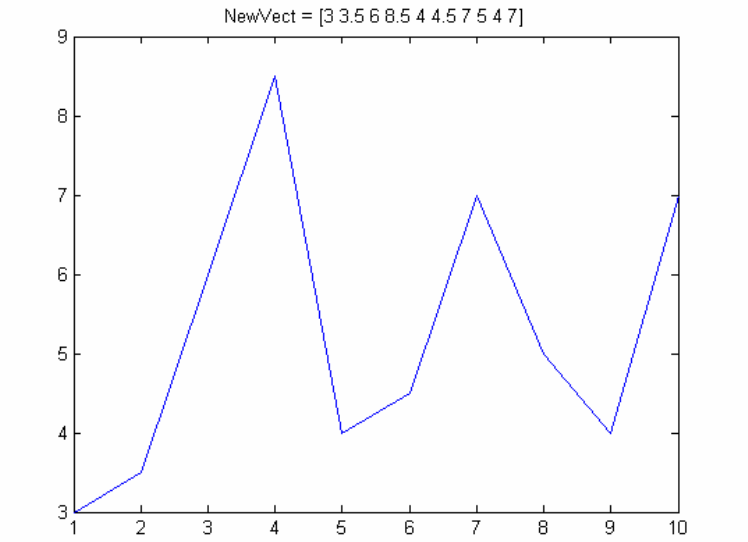
(α)



(β)



(γ)



Σχήμα 2.4: Εφαρμογή Αλγορίθμου Παρεμβολής σε Διάνυσμα

2.5 k-Means αλγόριθμος

Προβλήματα κατηγοριοποίησης (classification) συναντάμε σε πολλές εφαρμογές, όπως στη συμπίεση δεδομένων, στην αναζήτηση γνώσης, στην αναγνώριση προτύπων (pattern recognition) και στην κατηγοριοποίηση προτύπων (pattern classification).

Γενικά η κατηγοριοποίηση περιλαμβάνει το διαχωρισμό ενός συνόλου σημείων δεδομένων (data points) σε μη επικαλυπτόμενες κατηγορίες ή κλάσεις (classes), όπου τα σημεία σε μια κλάση είναι «πιο κοντά» το ένα στο άλλο από τα σημεία σε άλλες κλάσεις. Με τον όρο «πιο κοντά» όταν αναφερόμαστε σε σημεία κλάσεων εννοούμε πιο κοντά βάση ενός μέτρου προσέγγισης. Όταν ένα σύνολο δεδομένων κατηγοριοποιείται, κάθε σημείο ανατίθεται σε μια κλάση και κάθε κλάση μπορεί να χαρακτηριστεί με ένα μοναδικό σημείο αναφοράς (συνήθως ο μέσος όρος των σημείων που περιέχονται στην κλάση), που λέγεται κεντρικό σημείο (centroid)

Μια από τις πιο γνωστές εφαρμογές κατηγοριοποίησης είναι ο διαχωρισμός των φυτών ή των ζώων σε ξεχωριστές ομάδες ή είδη. Επίσης η κατηγοριοποίηση μαθητών ή φοιτητών με βάση την επίδοσή τους είναι ένα άλλο κοινό παράδειγμα, αυτής της λειτουργίας.[8]

Το αν μια μέθοδος θεωρείται καλή εξαρτάται από τα χαρακτηριστικά της ίδιας της εφαρμογής στην οποία πρόκειται να χρησιμοποιηθεί. Στην βιβλιογραφία συναντάμε πολλές μεθόδους για την εύρεση κλάσεων (classes) βασισμένες σε διάφορα κριτήρια, είτε τυχαίες είτε συστηματικές. Έτσι συναντάμε προσεγγίσεις που στηρίζονται στις διαδικασίες του διαχωρισμού και της συνένωσης (ISODATA), κάποιες τυχαίες προσεγγίσεις (CLARA, CLARANS), μεθόδους που σχετίζονται με νευρωνικά δίκτυα καθώς και πολλές άλλες τεχνικές.[9]

Τι όμως είναι αυτό που κάνει μια κατηγοριοποίηση να θεωρείται «καλή»; Ας θεωρήσουμε μια μοναδική κλάση σημείων μαζί με το κεντρικό σημείο της ή μέσο. Αν τα υπόλοιπα σημεία βρίσκονται πολύ κοντά στο κεντρικό σημείο, τότε αυτό θα είναι και αντιπροσωπευτικό όλων των σημείων της κλάσης. Το βασικό μέτρο μέτρησης της διασποράς των σημείων σε μια ομάδα γύρω από το μέσο είναι συνήθως το άθροισμα των τετραγώνων της απόστασης ανάμεσα σε κάθε σημείο και το μέσο. Αν τα σημεία είναι κοντά στο μέσο, η διασπορά θα είναι μικρή. Μια γενίκευση αυτής της διασποράς, όπου το κεντρικό σημείο αντικαθίσταται από ένα σημείο αναφοράς, που να είναι ή να μην είναι το κεντρικό, χρησιμοποιείται στην ανάλυση κλάσεων (classes analysis) και το άθροισμα όλων των διασπορών αποτελεί ένα μέτρο λάθους (error measure) το *E*, που είναι ουσιαστικά μια αντικειμενική μέθοδος σύγκρισης των διαφόρων μεθόδων κατηγοριοποίησης. [8]

Μια από τις πιο διαδεδομένες μεθόδους κατηγοριοποίησης, που βασίζεται στην ελαχιστοποίηση μιας αντικειμενικής συνάρτησης είναι η k-Means κατηγοριοποίηση (k-Means classification). Προτάθηκε το 1967 από τον J.MacQueen ως μια μη-επιβλεπόμενη τεχνική κατηγοριοποίησης. Δοσμένου ενός συνόλου n σημείων δεδομένων σε έναν χώρο d-διαστάσεων, R^d , και ενός ακεραίου k, το πρόβλημα είναι να καθοριστεί ένα σύνολο k σημείων στο R^d , που καλούνται κεντρικά σημεία (centroids), τέτοια ώστε να ελαχιστοποιηθεί η μέση τετράγωνη απόσταση από κάθε δεδομένο στο κοντινότερό του κέντρο.

Για την επίλυση του k-mean προβλήματος υπάρχει ένας ευριστικός αλγόριθμος που βασίζεται σε ένα απλό επαναληπτικό σχήμα για την εύρεση μιας τοπικά ελάχιστης λύσης, που καλείται k-mean αλγόριθμος, με πλήθος παραλλαγών[9]. Μια γενική περιγραφή του αλγορίθμου φαίνεται στο σχήμα 2.5 και ένα παράδειγμα κατηγοριοποίησης με τη χρήση του k-Means στο σχήμα 2.6. Πάντως ο k-Mean αλγόριθμος ανεξάρτητα παραλλαγής επειδή συνεχώς ανανεώνει τις κλάσεις, δεν απαιτεί πολλές επαναλήψεις και γενικά θεωρείται αρκετά γρήγορος αλγόριθμος.

k-Means Αλγόριθμος

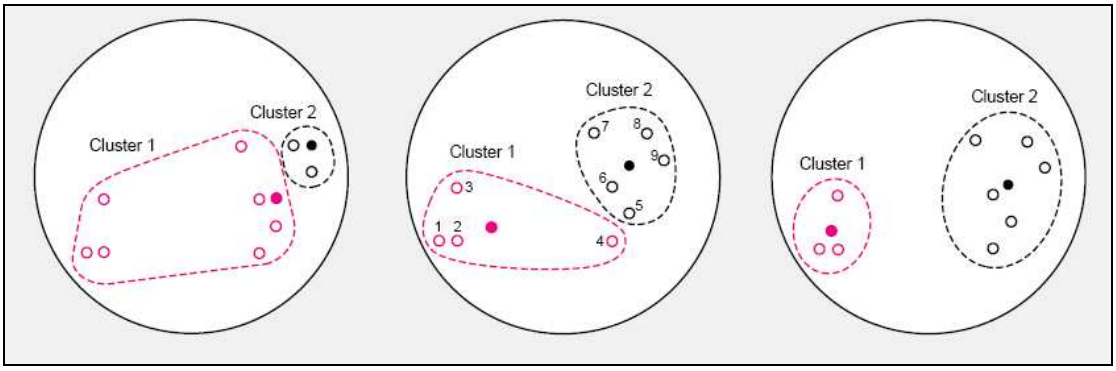
1. Όρισε το πλήθος των κλάσεων
2. Αρχικοποίησε τις κλάσεις με:
    a. Τυχαία κατανομή παραδειγμάτων στις κλάσεις

    Ή

    b. Τυχαία επιλογή κέντρων κλάσεων
3. Υπολόγισε το μέσο κάθε κλάσης
4. Απέδωσε κάθε δείγμα στο πλησιέστερο μέσο
5. Αν η κατανομή των δειγμάτων δεν άλλαξε τερμάτισε αλλιώς πήγαινε στο βήμα 3

Σχήμα 2.5: Αλγόριθμος k-Means

α) Αρχικοποίηση	**β) Πρώτη Επανάληψη**	**γ) Δεύτερη Επανάληψη**
Σημείο Αναφοράς 1 (κόκκινος κύκλος με γέμισμα) και Σημείο Αναφοράς 2 (μαύρος κύκλος με γέμισμα) Κάθε δεδομένο συμμετέχει σε μια κλάση, με βάση την απόσταση από τα σημεία αναφοράς	Σημεία Αναφοράς γίνεται τα κεντρικά σημεία κάθε κλάσης και τα υπόλοιπα δεδομένα ταξινομούνται στις δυο κλάσης με βάση την απόστασή τους από τα νέα σημεία αναφοράς	Το προηγούμενο βήμα εκτελείται ξανά και προκύπτει σταθερή κατηγοριοποίηση



Σχήμα 2.6: Παράδειγμα Κατηγοριοποίησης με τη χρήση του K-Means

Στο δικό μας υποσύστημα αφού εξαχθούν τα χαρακτηριστικά των εικόνων λέξεων το επόμενο βήμα είναι η κατηγοριοποίησή τους

Με βάση τα προηγούμενα δεδομένα η χρήση του παραπάνω αλγορίθμου θεωρήθηκε ικανοποιητική για το δικό μας υποσύστημα. Έτσι χρησιμοποιήθηκε μια υλοποίηση του k-

Means σε MATLAB. Ως μέτρο σύγκρισης χρησιμοποιήθηκε η Ευκλείδεια απόσταση ανάμεσα στις τιμές των χαρακτηριστικών των διάφορων δεδομένων.

Έτσι αφού εξαχθούν τα χαρακτηριστικά των εικόνων-λέξεων εισάγονται στο k-Means αλγόριθμο, μαζί με τα χαρακτηριστικά κάποιων αρχικών κλάσεων, δηλαδή κάποιων αρχικών λέξεων-εικόνων (το λεξιλόγιο που έχουμε) και υπολογίζονται οι αποστάσεις τους από αυτές. Με βάση το σχήμα 2.5 γίνεται η τελική κατηγοριοποίηση τους σε αυτές τις κλάσεις.

Κεφάλαιο 3

Περιγραφή Υποσυστήματος

3.1 Εισαγωγή

Όπως αναφέρθηκε και προηγουμένως το υποσύστημά μας ασχολείται με την αναγνώριση ολόκληρης λέξης κι όχι την κατάτμησή της σε χαρακτήρες και την μετέπειτα αναγνώρισή τους. Αντίθετα με τις μεθόδους αυτές εμείς επιλέγουμε να αποφύγουμε το στάδιο της κατάτμησης σε χαρακτήρες, κερδίζοντας σε χρόνο και προσπαθώντας να φέρουμε το σύστημα πιο κοντά στον τρόπο με τον οποίο σκέφτεται ο άνθρωπος.

Ένα γενικό σύστημα αναγνώρισης λέξης φαίνεται στο σχήμα 3.1 και αποτελείται από εισαγωγή της εικόνας εγγράφου, κάποια στάδια προεπεξεργασίας, όπως είναι η διόρθωση της γωνίας εκτροπής του εγγράφου, η κατάτμηση σε γραμμές, στη συνέχεια η διόρθωση της κλίσης των χαρακτήρων και τέλος η κατάτμηση του εγγράφου σε λέξεις.

Έτσι με την προεπεξεργασία παίρνουμε τις εικόνες-λέξεις, οι οποίες είναι έτοιμες να εισαχθούν στο σύστημα της αναγνώρισης λέξεων. Τελικό στάδιο είναι η κατηγοριοποίησή τους σε κλάσεις, με βάση κάποιες αρχικές κλάσεις[13].

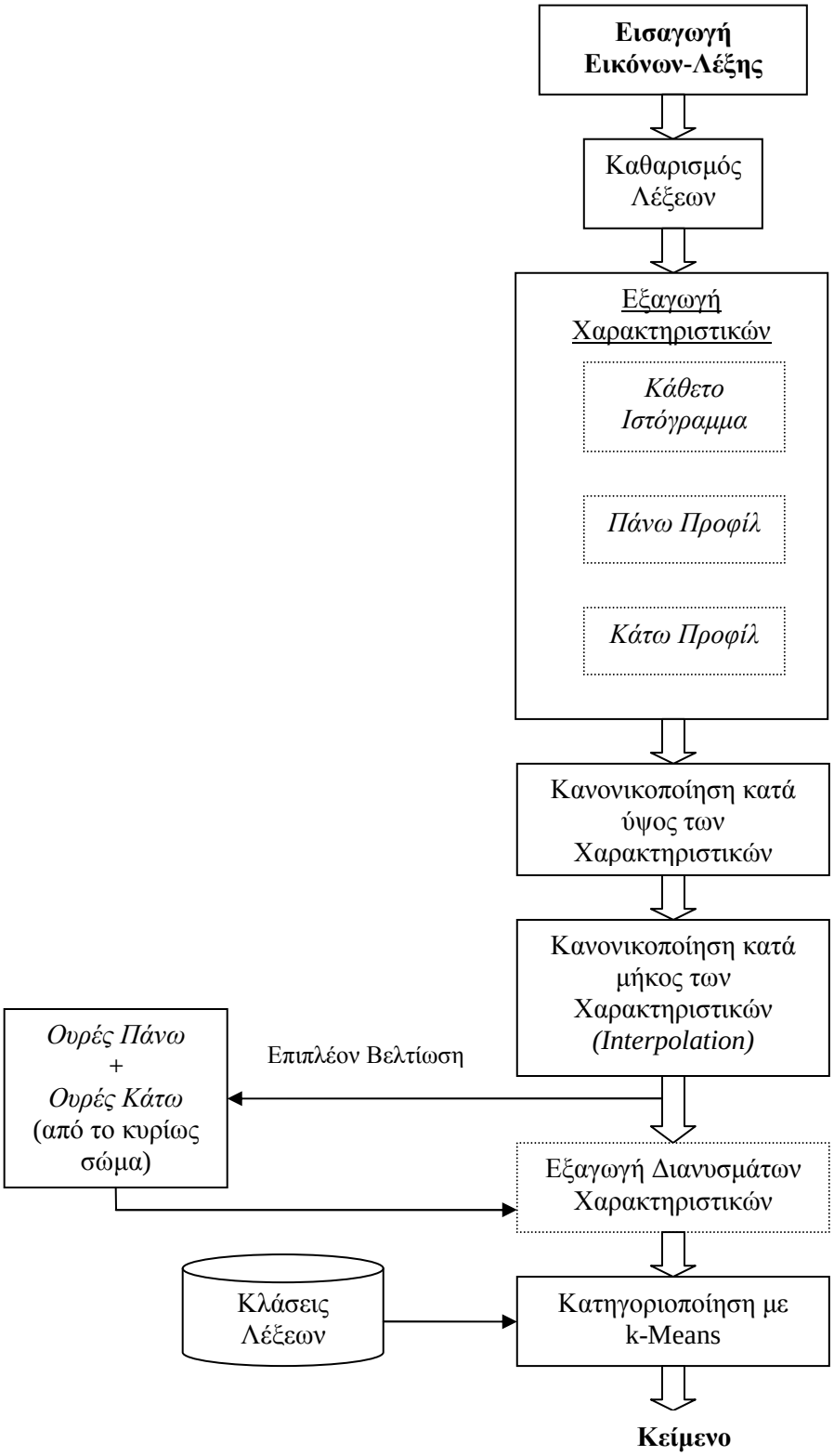
Σχήμα 3.1: Γενικό Σύστημα Αναγνώρισης Ολόκληρης Λέξης

Εμείς αναπτύξαμε ένα υποσύστημα του παραπάνω συστήματος στο οποιο ασχολούμαστε μόνο με το στάδιο της αναγνώρισης λέξης (που φαίνεται με κόκκινο στο παραπάνω σχήμα), θεωρώντας ότι έχουμε έτοιμες τις εικόνες-λέξεις. Βέβαια πρέπει να τονιστεί ότι στα πειράματα που θα ακολουθήσουν στο Κεφάλαιο 4 οι εικόνες-λέξεις δεν έχουν υποστεί καμία επεξεργασία, όπως διόρθωση κλίσης χαρακτήρων.

3.2 Αναγνώριση Λέξης

Τα βασικά στάδια του υποσυστήματός μας για την αναγνώριση μιας εικόνας-λέξης φαίνονται στο σχήμα 3.2.

Γενικά μπορούμε να πούμε ότι εισάγονται στο υποσύστημα οι εικόνες-λέξεις, χωρίς, όπως είπαμε, να έχουν υποστεί κάποια προεπεξεργασία. Αφού καθαριστούν από κενά, εξάγονται τα μορφολογικά χαρακτηριστικά τους, κάθετο ιστόγραμμα, πάνω προφίλ και κάτω προφίλ και αφού με παρεμβολή τα φέρουμε στο ίδιο μήκος για όλες τις λέξεις, εξάγεται ένα διάνυσμα για κάθε εικόνα-λέξη. Επιπλέον βελτίωση προσφέρεται με την καταγραφή της σχετικής θέσης των πάνω και κάτω ουρών κάθε εικόνας-λέξης. Τέλος πραγματοποιείται η κατηγοριοποίησή τους με βάση αρχικές κλάσεις, με τη χρήση του k-Means αλγορίθμου.

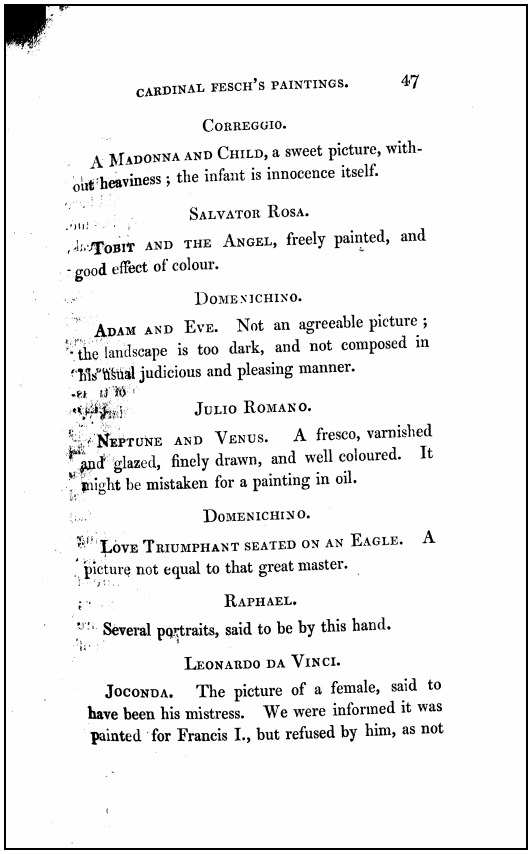


Σχήμα 3.2: Στάδια Προτεινόμενου Υποσυστήματος

Παρακάτω θα δούμε τα διάφορα στάδια του υποσυστήματος μας αναλυτικά.

3.2.1 Προεπεξεργασία - Καθαρισμός λέξης

Οι εικόνες-λέξεις που εισάγονται στο υποσύστημα που αναπτύξαμε προέρχονται από παλιά ιστορικά έγγραφα ή χειρόγραφα έγγραφα, με αποτέλεσμα να έχουν πολλές ατέλειες. Είναι πιθανόν να έχουν κλίση, τη λεγόμενη γωνία εκτροπής (σχήμα 3.3α), λόγω κακού σαρώματος, αλλά και οι ίδιοι οι χαρακτήρες να έχουν κλίση (σχήμα 3.3β), ενώ ολόκληρη η λέξη όχι, ειδικά στο χειρόγραφο κείμενο. Και αφού μιλάμε κυρίως για ιστορικά έγγραφα είναι φυσικό να παρατηρείται θόρυβος στις εικόνες αυτές (σχήμα 3.3γ).



CARDINAL FESCH'S PAINTINGS. 47

CORREGGIO.

A MADONNA AND CHILD, a sweet picture, without heaviness; the infant is innocence itself.

SALVATOR ROSA.

TOBIT AND THE ANGEL, freely painted, and good effect of colour.

DOMENICHINO.

ADAM AND EVE. Not an agreeable picture; the landscape is too dark, and not composed in his usual judicious and pleasing manner.

JULIO ROMANO.

NEPTUNE AND VENUS. A fresco, varnished and glazed, finely drawn, and well coloured. It might be mistaken for a painting in oil.

DOMENICHINO.

LOVE TRIUMPHANT SEATED ON AN EAGLE. A picture not equal to that great master.

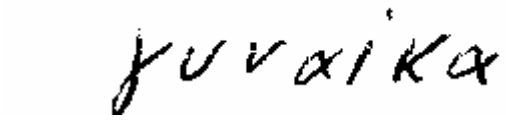
RAPHAEL.

Several portraits, said to be by this hand.

LEONARDO DA VINCI.

JOCONDA. The picture of a female, said to have been his mistress. We were informed it was painted for Francis I., but refused by him, as not

(α) Έγγραφο με Γωνία Εκτροπής



(β) Λέξη με Κλίση Χαρακτήρων

(γ) Θόρυβος

Σχήμα 3.3: Ατέλειες Εγγράφων

Διάφορες τεχνικές διόρθωσης γωνίας εκτροπής κατηγοριοποιεί ο [19], ενώ νέες μέθοδοι προτείνονται από τους [20], [21], καθώς και από ένα πλήθος άλλων ερευνητών, αφού

το θέμα αυτό έχει εξαιρετικό ενδιαφέρον. Αντίστοιχα και ο τομέας της διόρθωσης κλίσης χαρακτήρων είναι πολύ σημαντικός, αφού διευκολύνει την αναγνώριση και μειώνει την πιθανότητα εμφάνισης λάθους. Πλήθος τεχνικών προτείνονται, όπως ενδεικτικά φαίνεται στην βιβλιογραφία οι [20] και [22].

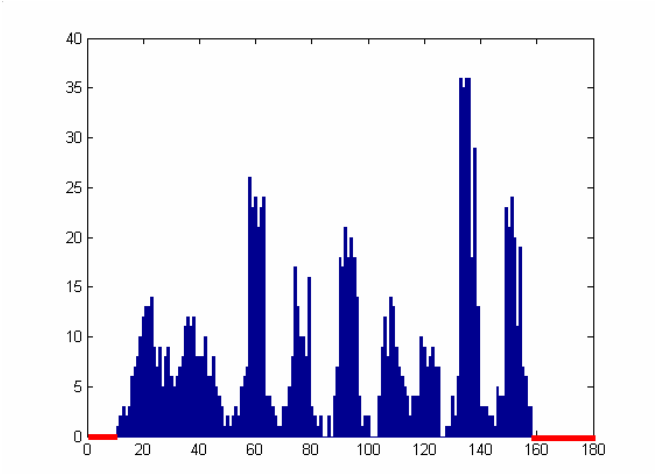
Στο υποσύστημά μας οι εικόνες-λέξεις που εισήχθησαν προέρχονται από το σύστημα των [13] με διόρθωση μόνο της γωνίας εκτροπής του εγγράφου και δεν ασχοληθήκαμε με άλλη προεπεξεργασία των εικόνων, γεγονός που δυσκολεύει την αναγνώριση λέξεων ακόμη περισσότερο. Η μόνη μορφή προεπεξεργασίας που πραγματοποιήσαμε είναι ο καθαρισμός των εικόνων-λέξεων.

Οι εικόνες που εισάγονται στο σύστημά μας περιέχουν επιπλέον λευκά κομμάτια πάνω, κάτω, δεξιά και αριστερά της λέξης. Το γεγονός αυτό επηρεάζει την αναγνώριση των λέξεων, αφού κατά την εξαγωγή των χαρακτηριστικών στα επόμενα στάδια θέλουμε να γίνει η κανονικοποίησή τους κατά το «πραγματικό» ύψος και μήκος της λέξης, και όχι αυτό που προκύπτει με την ύπαρξη των λευκών κομματιών.

Επίσης ο καθαρισμός της εικόνας-λέξης απαλλάσσει τα εξαγόμενα χαρακτηριστικά από επιπλέον τιμές, που στην ουσία θα είναι μηδενικές. Για παράδειγμα το κάθετο ιστόγραμμα πριν το καθαρισμό θα είχε στην αρχή και στο τέλος κάποιες μηδενικές τιμές από την ύπαρξη των λευκών κομματιών αριστερά και δεξιά της λέξης (σχήμα 3.4β-κόκκινες γραμμές).

(α)

(β)



Σχήμα 3.4: Κάθετο Ιστόγραμμα Λέξης πριν τον Καθαρισμό

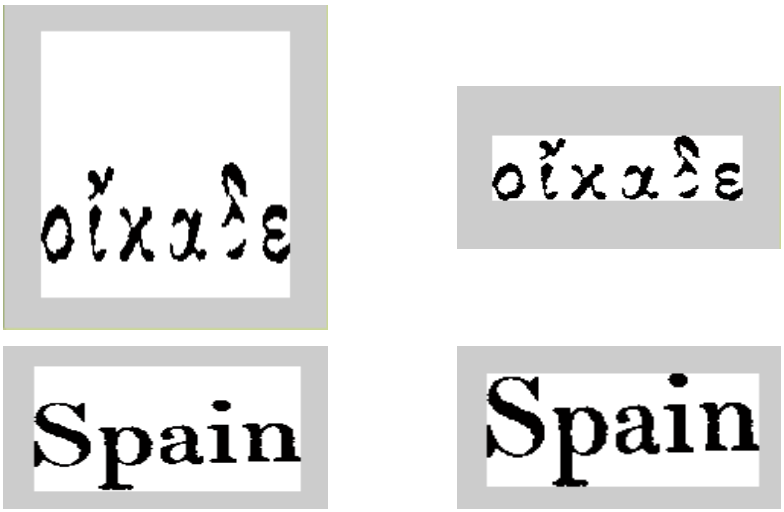
Όπως αναφέραμε και σε προηγούμενο κεφάλαιο, η εικόνα-λέξη είναι στην ουσία ένας πίνακας MxN, με τιμές 0 αν τα αντίστοιχα pixels είναι άσπρα και 1 αν είναι μαύρα. Η διαδικασία του καθαρισμού της εικόνας, ουσιαστικά περιλαμβάνει τέσσερα στάδια εντοπισμού λευκών γραμμών

- *πάνω από τη λέξη*, σαρώνοντας την εικόνα από την πρώτη γραμμή και προς τα κάτω μέχρι να εντοπιστούν σε κάποια στήλη τα πρώτα μαύρα pixels

- *κάτω από τη λέξη*, σαρώνοντας την εικόνα από την τελευταία γραμμή και προς τα πάνω μέχρι να εντοπιστούν σε κάποια στήλη τα πρώτα μαύρα pixels
- *αριστερά από τη λέξη*, σαρώνοντας την εικόνα από την πρώτη στήλη και προς τα δεξιά μέχρι να εντοπιστούν σε κάποια γραμμή τα πρώτα μαύρα pixels
- *δεξιά από τη λέξη*, σαρώνοντας την εικόνα από την τελευταία στήλη και προς τα αριστερά μέχρι να εντοπιστούν σε κάποια γραμμή τα πρώτα μαύρα pixels

Έτσι τελικά, κρατάμε από την εικόνα μόνο το μέρος εκείνο που είναι η λέξη, διευκολύνοντας με αυτόν τον τρόπο την κανονικοποίηση και κατ' επέκταση την αναγνώριση της λέξης.

Παράδειγμα λέξεων πριν και μετά τον καθαρισμό φαίνονται στο σχήμα 3.5α και 3.5β αντίστοιχα.



(α) Πριν **(β) Μετά**

Σχήμα 3.5: Παραδείγματα Λέξεων Πριν και Μετά τον Καθαρισμό

3.2.2 Εξαγωγή Χαρακτηριστικών

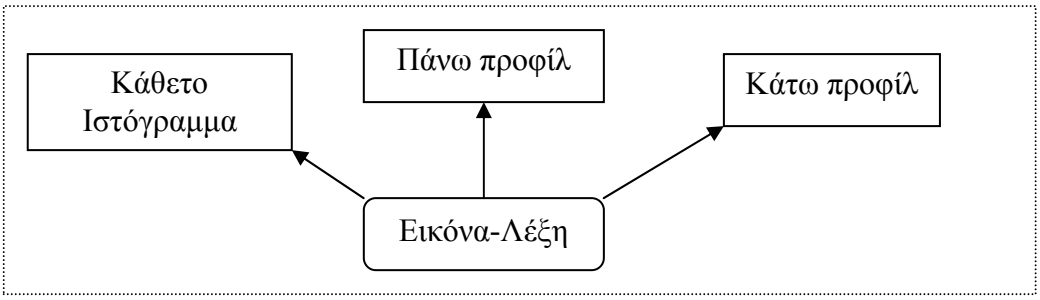
Το επόμενο στάδιο μετά τον καθαρισμό των εικόνων-λέξεων είναι, όπως αναφέρθηκε η εξαγωγή των μορφολογικών χαρακτηριστικών τους και η δημιουργία ενός διανύσματος για κάθε εικόνα-λέξη ξεχωριστά. Στη συνέχεια τα διανύσματα αυτά θα συγκριθούν μεταξύ τους για να γίνει η κατηγοριοποίηση των λέξεων σε κλάσεις.

Γενικά, όπως αναφέραμε και στο κεφάλαιο 1, τα χαρακτηριστικά που εξάγουμε για ολόκληρες λέξεις, μπορούν να κατηγοριοποιηθεί σύμφωνα με τους [1] σε τρεις κατηγορίες: χαμηλού, μεσαίου και υψηλού επιπέδου. Τα υψηλού επιπέδου χαρακτηριστικά αποτελούν ουσιαστικά μια περιγραφή της μορφής της λέξης και έρχονται πιο κοντά στον τρόπο με τον οποίο οι άνθρωποι αναγνωρίζουν μια λέξη.

Έτσι αποφασίστηκε στο υποσύστημά μας να χρησιμοποιηθούν μόνο μορφολογικά χαρακτηριστικά, που μπορούν να λειτουργήσουν με απλούς αλγορίθμους κατηγοριοποίησης.

Για το σκοπό αυτό επιλέχτηκαν τρία μορφολογικά χαρακτηριστικά, το κάθετο ιστόγραμμα, το πάνω προφίλ και το κάτω προφίλ (σχήμα 3.6). Στη συνέχεια θα τα δούμε πιο αναλυτικά.

Εξαγωγή χαρακτηριστικών



Σχήμα 3.6: Εξαγωγή Χαρακτηριστικών Λέξεων

3.2.2.1 Κάθετο Ιστόγραμμα

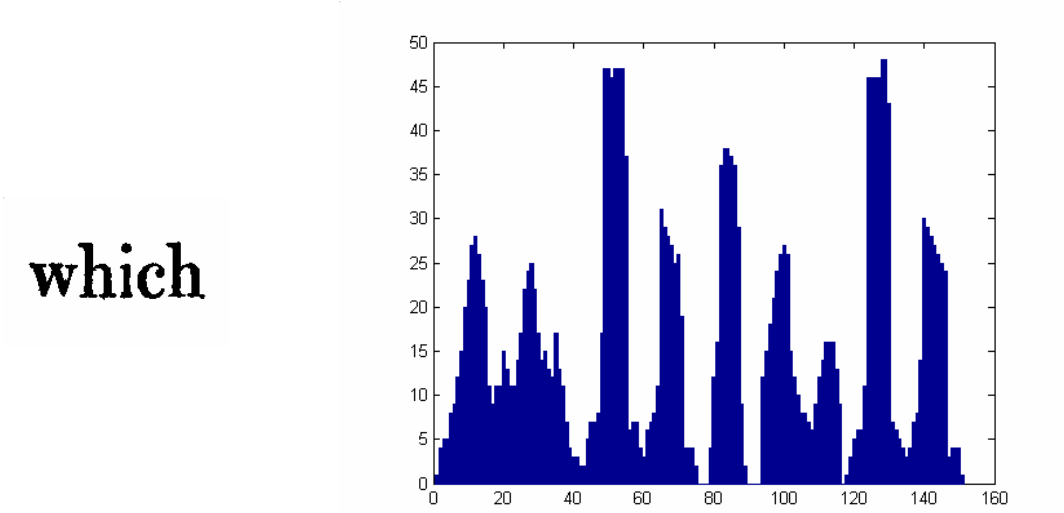
Αρχικά θεωρήθηκε αναγκαίο να υπάρχει μια ένδειξη του πλήθους των μαύρων pixels που υπάρχουν σε κάθε εικόνα-λέξη, που να τις διαφοροποιεί αρκετά μεταξύ τους.

Ως κάθετο ιστόγραμμα ορίστηκε, σε προηγούμενο κεφάλαιο, το πλήθος των μαύρων pixels που συναντάμε σε κάθε στήλη της εικόνας, ενώ ως οριζόντιο ιστόγραμμα το πλήθος των μαύρων pixels που συναντάμε σε κάθε γραμμή της εικόνας.

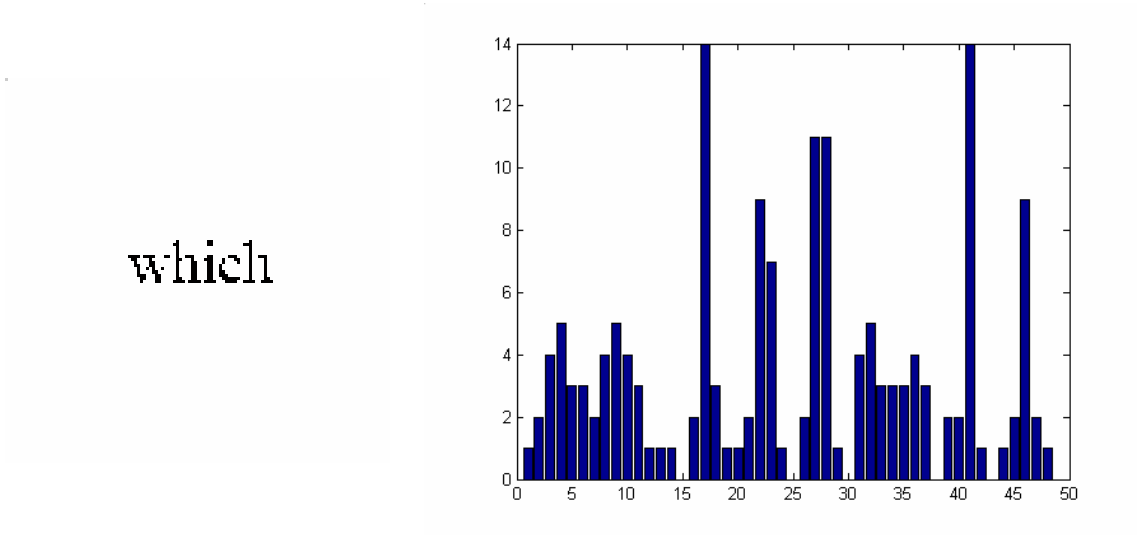
Το οριζόντιο ιστόγραμμα ενώ αποτελεί βασικό χαρακτηριστικό για την αναγνώριση ανεξάρτητων χαρακτήρων [13], ή στην κατάτμηση σε χαρακτήρες, στην αναγνώριση ολόκληρης λέξης δεν προσφέρει ουσιαστικά βελτιστοποίηση, αφού οι λέξεις με ίδιο ή διαφορετικό μήκος μπορεί να έχουν παρόμοια οριζόντια ιστογράμματα.

Από την άλλη μεριά, το κάθετο ιστόγραμμα προσφέρει ένα μέτρο ομοιότητας, αφού μας παρουσιάζει μια κατανομή του μελανιού σε στήλες. Με αυτό τον τρόπο έχουμε ένα δείγμα των ίδιων των γραμμάτων στην λέξη, των ουρών πάνω και κάτω από το κύριο σώμα της λέξης, τον τόνων που ίσως υπάρχουν και άλλων κοινών σημείων των λέξεων.

Για παράδειγμα στο σχήμα 3.7α βλέπουμε μια εικόνα από ιστορικό έγγραφο και το αντίστοιχο κάθετο ιστόγραμμά της, ενώ στο σχήμα 3.7β βλέπουμε την ίδια λέξη από τυπωμένο κείμενο και το αντίστοιχο κάθετο ιστόγραμμά της.



(α)



(β)

Σχήμα 3.7: Διαφορετικές Εμφανίσεις Ίδιων Λέξεων και τα αντίστοιχα Κάθετα Ιστογράμματα τους

Παρατηρούμε ότι παρά την διαφορά μήκους, στα κάθετα ιστογράμματα σχηματίζονται αντίστοιχα υψώματα και κοιλώματα. Για παράδειγμα στις στήλες που σχηματίζουν το γράμμα *h* η κατανομή του μελανιού είναι αντίστοιχη, όπως και στα υπόλοιπα γράμματα και φαίνονται επίσης και αντίστοιχα τα κενά ανάμεσα στα γράμματα.

Από την εφαρμογή του κάθετου ιστογράμματος παίρνουμε ένα μονοδιάστατο διάνυσμα με μήκος (στήλες) ίσο με αυτό ολόκληρης της λέξης και τιμή για κάθε στήλη το πλήθος των μαύρων pixels της αντίστοιχης στήλης της εικόνας. Ένα παράδειγμα του παραγόμενου διανύσματος για τη λέξη 3.7α φαίνεται στο σχήμα 3.8.

Vertical Histogram (which)

[1 4 5 5 8 9 12 15 20 23 27 28 26 23 20 11 9 11 11 15
13 11 11 14 17 22 24 25 22 17 14 15 13 12 17 13 11 7 4
3 3 2 2 5 7 7 8 17 47 47 46 47 47 47 37 6 7 7 4 3
6 7 8 11 31 29 28 27 25 26 19 4 4 4 2 0 0 0 4 12
16 36 38 38 37 36 29 9 2 0 0 0 0 12 15 18 21 24 26 27
26 15 12 10 8 8 7 6 9 12 14 16 16 16 13 9 0 1 3 5
6 6 11 46 46 46 46 48 48 43 7 6 5 4 3 4 7 8 14 30
29 28 27 26 25 24 3 4 4 4 1]

Σχήμα 3.8: Παράδειγμα Διανύσματος Κάθετου Ιστογράμματος

3.2.2.2 Προφίλ

Τα προφίλ χρησιμοποιήθηκαν αρκετά στην αναγνώριση μεμονωμένων χαρακτήρων, όπως στους [13] και [14]. Επίσης οι [10] και [23] τα χρησιμοποίησαν σε ολόκληρες λέξεις.

Το σίγουρο πάντως είναι πως αποτελούν μια βασική περιγραφή του σχήματος είτε μιας ολόκληρης λέξης, είτε ενός χαρακτήρα. Για αυτό και επιλέχτηκαν ως βασικά μορφολογικά χαρακτηριστικά για το δικό μας υποσύστημα.

Βέβαια δεν είναι απαραίτητα όλα τα είδη των προφίλ για την περιγραφή της λέξης. Έτσι το αριστερό και δεξιό προφίλ, ενώ δίνουν σημαντικές πληροφορίες του σχήματος ενός χαρακτήρα ([2]) δεν προσφέρουν ουσιαστικά τίποτα στην αναγνώριση της ολόκληρης λέξης. Το ίδιο συμβαίνει και με τα ακτινικά προφίλ. Σε αντίθεση όμως με αυτά, τα πάνω και κάτω προφίλ μας δίνουν σημαντικές πληροφορίες.

Χρησιμοποιώντας έτσι το πάνω προφίλ μιας εικόνας-λέξης παίρνουμε στην ουσία το εξωτερικό-πάνω σχήμα της λέξης με όλες τις ουρές, τόνους κτλ που βρίσκονται πάνω από το κυρίως σώμα της λέξης. Όπως αναφέραμε, το πάνω προφίλ είναι στην ουσία τα πρώτα μαύρα pixels που συναντάμε σε κάθε στήλη της εικόνας ξεκινώντας από την πρώτη γραμμή και προχωρώντας προς τα κάτω. Δημιουργούμε με αυτόν τον τρόπο ένα μονοδιάστατο διάνυσμα με μήκος (στήλες) ίσο με αυτό της λέξης, που οι τιμές του θα είναι για κάθε στήλη ο αριθμός της γραμμής στην οποία εμφανίστηκε το πρώτο μαύρο pixel. Η συνάρτηση που κατασκευάστηκε σε MATLAB και χρησιμοποιήθηκε στο σύστημά μας φαίνεται στο σχήμα 3.9.

```
function g=upprofil(image)
%Δημιουργεί το πάνω προφίλ της λέξης
[Rows, Columns]=size(image);
count=ones(Rows, Columns);
for i=1:Columns
    for j=1:Rows
        if image(j,i)==0
            count(j,i)=0;
            upRow(i)=j;
            break
        end
    end
end
g=upRow;
```

Σχήμα 3.9: Συνάρτηση Υπολογισμού Πάνω Προφίλ Λέξης

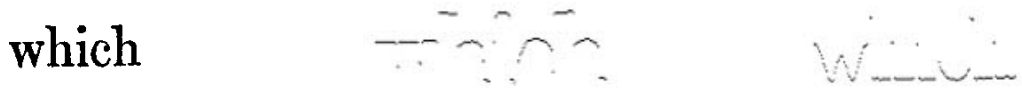
Κατά αναλογία το κάτω προφίλ, προσφέρει μια περιγραφή του εξωτερικού-κάτω σχήματος της λέξης και ορίζεται ως τα πρώτα μαύρα pixels που συναντάμε σε κάθε στήλη της εικόνας ξεκινώντας από την τελευταία γραμμή και συνεχίζοντας προς τα πάνω. Παίρνουμε πάλι ένα μονοδιάστατο διάνυσμα μήκους (στήλες) ίσου με αυτό της λέξης και τιμή για κάθε στήλη τον αριθμό της γραμμής στην οποία εμφανίστηκε το πρώτο μαύρο pixel. Ο κώδικας σε MATLAB που χρησιμοποιήθηκε φαίνεται στο σχήμα 3.10.

```
function g=downprofil(image)
%Δημιουργεί το κάτω προφίλ της λέξης
[Rows,Columns]=size(image);
count=ones(Rows,Columns);
```

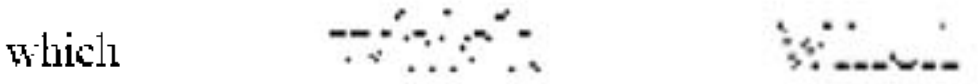
```
for i=Columns:-1:1
    for j=Rows:-1:1
        if image(j,i)==0
            count(j,i)=0;
            downRow(i)=j;
            break
        end
    end
end
g=downRow;
```

Σχήμα 3.10: Συνάρτηση Υπολογισμού Κάτω Προφίλ Λέξης

Παρακάτω στο σχήμα 3.11α φαίνεται μια λέξη ιστορικού κειμένου και το αντίστοιχο πάνω και κάτω προφίλ, ενώ στο 3.11β φαίνεται η ίδια λέξη αλλά τυπωμένου κειμένου και τα αντίστοιχα προφίλ της.



(α)



(β)

Σχήμα 3.11: Διαφορετικές Εμφανίσεις Ίδιων Λέξεων και τα αντίστοιχα Πάνω και Κάτω Προφίλ τους

Στις πιο πάνω εικόνες παρατηρούμε τις ομοιότητες. Αν και το μήκος των δυο λέξεων είναι διαφορετικό τα σχήματα που προκύπτουν μοιάζουν πάρα πολύ. Έτσι θεωρούμε ότι το πάνω και το κάτω προφίλ είναι ένα καλό μέτρο ομοιότητας.

Στο σχήμα 3.12 βλέπουμε τα δυο διανύσματα, πάνω και κάτω προφίλ, που προκύπτουν από τη λέξη του σχήματος 3.11α.

UpperProfile(which)

21	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*22*	*37*	*21*	*21*	*21*	*21*
21	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*21*	*34*	*35*	*21*	*20*	*20*	*20*	*20*	*20*	*20*	*20*

20 21 0 2 2 2 2 2 2 3 3 3 3 3 3 24 23 23 22 21 21
21 21 21 20 21 21 21 22 23 24 33 45 45 46 46 0 0 21 5
4 2 2 2 2 2 2 3 4 45 45 47 0 0 33 29 27 25 23 21
20 20 20 19 19 19 19 19 20 20 20 21 21 23 24 26 40 3 2
2 2 2 2 2 2 2 1 1 1 2 23 22 21 21 20 20 20 19 19
19 20 20 21 21 22 24 44 44 44 44 45

(α) Διάνυσμα Πάνω Προφίλ

DownProfil(which)

23 23 24 26 28 31 33 37 39 42 47 49 49 50 49 43 40 36 34
32 31 32 36 38 40 44 46 48 49 49 47 43 39 36 33 29 25 24
24 23 0 4 4 47 48 48 47 47 49 49 48 48 48 48 48 48 48
22 22 47 48 48 48 48 48 48 48 48 48 48 48 48 48 47 0 0
23 47 48 48 49 48 48 48 48 48 48 48 48 48 0 0 36 40 42
44 44 46 47 47 48 48 48 48 49 49 49 48 48 47 47 46 45 42
41 4 5 5 47 48 48 47 47 47 47 48 47 47 47 47 47 47 22
22 46 47 47 47 47 47 47 46 46 46 47 47 47 47 46

(β) Διάνυσμα Κάτω Προφίλ

Σχήμα 3.12: Παραδείγματα Διανυσμάτων για Πάνω και Κάτω Προφίλ

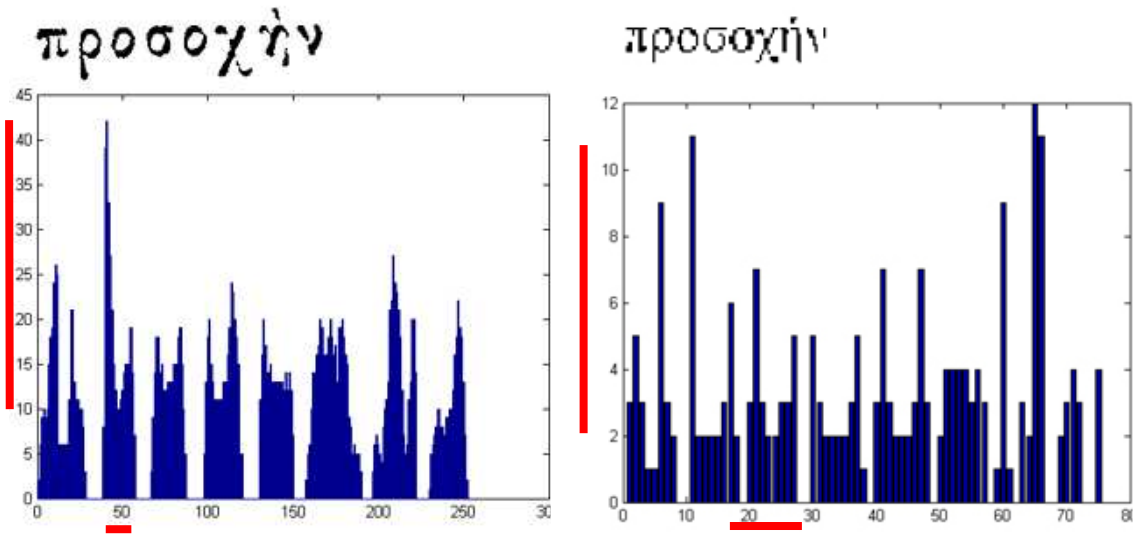
3.2.3 Κανονικοποίηση κατά ύψος

Οι εικόνες λέξεις οι οποίες εισάγονται στο υποσύστημα που προτείνουμε, μπορεί να προέρχονται από διαφορετικά ιστορικά έγγραφα, να είναι χειρόγραφες ή απλά τυπωμένο κείμενο. Αυτό σημαίνει ότι ίδιες λέξεις θα διαφέρουν όχι μόνο στο τρόπο με τον οποίο είναι γραμμένες, αλλά και στο ύψος. Άρα και οι τιμές των εξαγόμενων χαρακτηριστικών για διαφορετικές εμφανίσεις της ίδια λέξης δεν θα είναι ίδιες ή έστω κοντινές.

Στο σχήμα 3.13α βλέπουμε, για παράδειγμα, μια λέξη από ιστορικό κείμενο και το αντίστοιχο κάθετο ιστόγραμμά της, ενώ στο σχήμα 3.13β βλέπουμε μια διαφορετική εμφάνιση της ίδιας λέξης από τυπωμένο κείμενο. Το ύψος της πρώτης είναι *43 pixels*, ενώ της δεύτερης *12 pixels*.

Παρατηρούμε ότι στο κάθετο ιστόγραμμα της πρώτης εικόνας στη θέση όπου αναπαρίσταται το γράμμα *ρ* (ύπαρξη μεγάλου αριθμού μαύρων pixels σε αυτήν την στήλη, λόγω της ουράς) οι τιμές του ιστογράμματος κυμαίνονται από 10 έως 43 περίπου (κόκκινη γραμμή στον y-άξονα της α). Αντίθετα για την δεύτερη εικόνα οι τιμές που αναπαριστούν το *ρ* κυμαίνονται από 2 έως 11 περίπου (κόκκινη γραμμή στον y-άξονα της β). Αντίστοιχα και για τα υπόλοιπα σημεία της λέξης.

Αν συγκριθούν αυτές οι τιμές το αποτέλεσμα θα είναι η κατηγοριοποίηση των λέξεων αυτών σε διαφορετικές κλάσεις. Το ίδιο συμβαίνει και με τις τιμές των πάνω και κάτω προφίλ.



(α) (β)

Σχήμα 3.13: Διαφορετικά Κάθετα Ιστογράμματα για Διαφορετικές Εμφανίσεις Ίδιων λέξεων

Για το λόγο αυτό απαραίτητη η κανονικοποίηση των τιμών των εξαγόμενων χαρακτηριστικών κατά ύψος. Δηλαδή κάθε τιμή του κάθετου ιστογράμματος θα διαιρεθεί με το ύψος της εικόνας και κατά αντιστοιχία και οι τιμές του πάνω και κάτω προφίλ.

Αποτέλεσμα αυτού του σταδίου λοιπόν θα είναι η παραγωγή κανονικοποιημένων διανυσμάτων των χαρακτηριστικών, που ουσιαστικά θα είναι διανύσματα των οποίων οι τιμές θα είναι σχετικές το ύψος της κάθε εικόνας. Τώρα η σύγκριση των τιμών θα είναι ανάλογη και άρα τα λάθη της αναγνώρισης των λεξεων θα μειωθούν.

Για παράδειγμα το κανονικοποιημένο διάνυσμα του σχήματος 3.12α φαίνεται στο σχήμα 3.14.

Normalized UpperProfle (which)

0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409
0.1409 0.1409 0.1409 0.1477 0.2483 0.1409 0.1409 0.1409 0.1409 0.1409
0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.1409 0.2282 0.2349
0.1409 0.1342 0.1342 0.1342 0.1342 0.1342 0.1342 0.1342 0.1342 0.1409
0 0.0134 0.0134 0.0134 0.0134 0.0134 0.0134 0.0201 0.0201 0.0201
0.0201 0.0201 0.0201 0.1611 0.1544 0.1544 0.1477 0.1409 0.1409 0.1409
0.1409 0.1409 0.1342 0.1409 0.1409 0.1409 0.1477 0.1544 0.1611 0.2215
0.3020 0.3020 0.3087 0.3087 0 0 0.1409 0.0336 0.0268 0.0134
0.0134 0.0134 0.0134 0.0134 0.0134 0.0201 0.0268 0.3020 0.3020 0.3154
0 0 0.2215 0.1946 0.1812 0.1678 0.1544 0.1409 0.1342 0.1342 0.1342
0.1275 0.1275 0.1275 0.1275 0.1275 0.1342 0.1342 0.1342 0.1409 0.1409
0.1544 0.1611 0.1745 0.2685 0.0201 0.0134 0.0134 0.0134 0.0134 0.0134
0.0134 0.0134 0.0134 0.0067 0.0067 0.0067 0.0134 0.1544 0.1477 0.1409
0.1409 0.1342 0.1342 0.1342 0.1275 0.1275 0.1275 0.1342 0.1342 0.1409
0.1409 0.1477 0.1611 0.2953 0.2953 0.2953 0.2953 0.3020

Σχήμα 3.14: Παράδειγμα Κανονικοποιημένου Κατά Ύψος Διανύσματος Πάνω Προφίλ

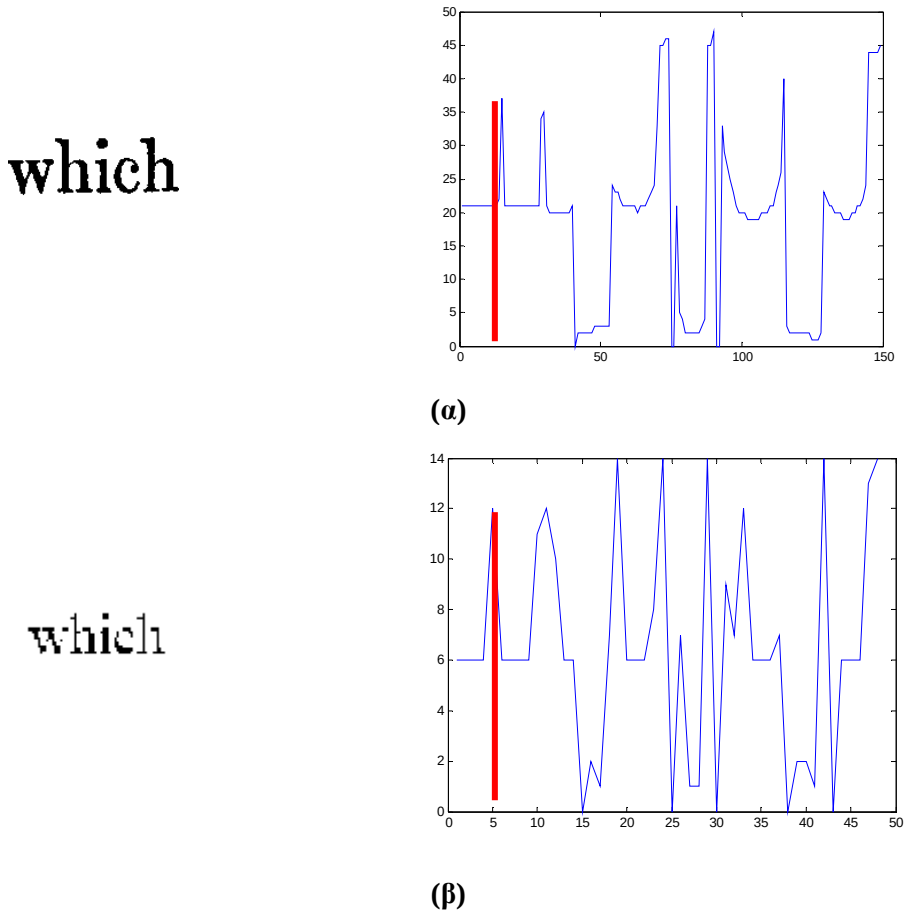
3.2.4 Κανονικοποίηση κατά μήκος με Παρεμβολή

Όπως αναφέρθηκε και στην παράγραφο 3.2.3, οι εικόνες λέξεις διαφέρουν μεταξύ τους, ακόμα κι αν μιλάμε για διαφορετικές εμφανίσεις της ίδιας λέξης. Εκεί μιλήσαμε για διαφορετικό ύψος της λέξης, αλλά δεν είναι η μόνη διαφορά που επηρεάζει την σύγκριση. Διαφορετικές εμφανίσεις ίδιων λέξεων διαφέρουν και ως προς το μήκος.

Για παράδειγμα στο σχήμα 3.15α βλέπουμε μια λέξη από ιστορικό κείμενο και τη γραφική παράσταση του αντίστοιχου διανύσματος του πάνω προφίλ της, ενώ στο σχήμα 3.15β παρατηρούμε μια διαφορετική εμφάνιση της ίδιας λέξης από τυπωμένο κείμενο και την αντίστοιχη γραφική παράσταση του πάνω προφίλ της.

Η πρώτη λέξη έχει *Μήκος=149 pixels*, ενώ η άλλη εμφάνιση της λέξης έχει διαφορετικό μέγεθος, *Μήκος=59 pixels*. Ενώ, λοιπόν, πρόκειται για την ίδια λέξη, οι τιμές του πάνω προφίλ, και αντίστοιχα και των υπόλοιπων χαρακτηριστικών, θα είναι διαφορετικές.

Για παράδειγμα στο πάνω προφίλ του σχήματος 3.15α το πρώτο ύψωμα το συναντάμε στη θέση 19 περίπου (κόκκινη γραμμή στο α), ενώ στην δεύτερη εικόνα στη θέση 5 περίπου (κόκκινη γραμμή στο β). Το ίδιο συμβαίνει και για τα υπόλοιπα σημεία των πάνω προφίλ των εικόνων .



(α)

(β)

Σχήμα 3.15: Πάνω Προφίλ Διαφορετικών Εμφανίσεων Ίδιων Λέξεων

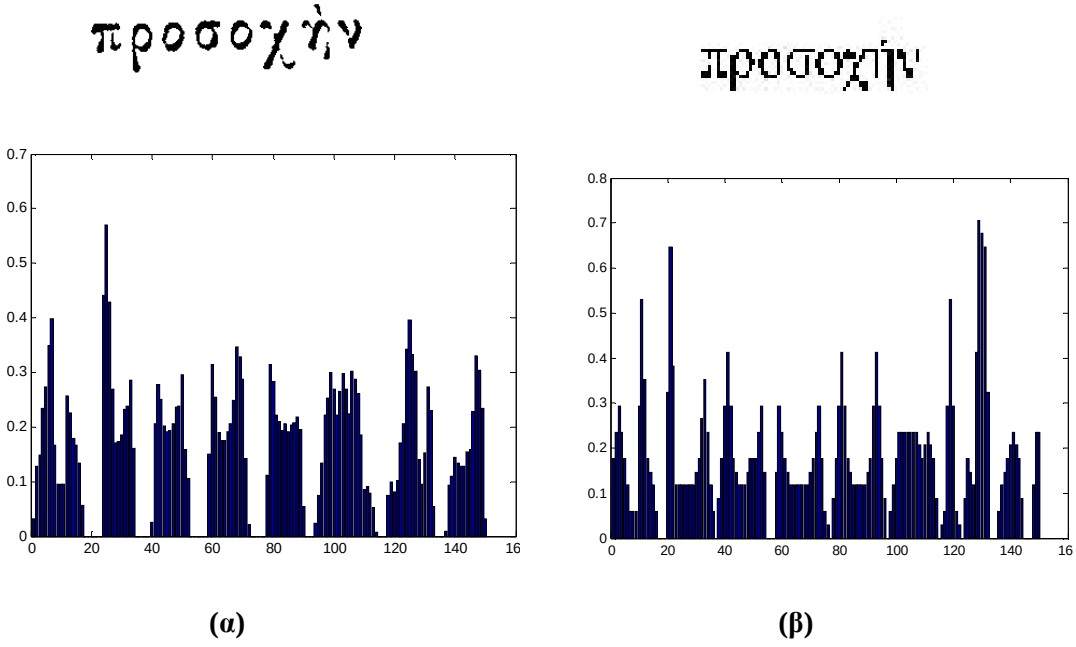
Αν παρατηρήσουμε επίσης το κάθετο ιστόγραμμα του σχήματος 3.13α, θα δούμε ότι το γράμμα *ρ* περιγράφεται τις θέσεις από 40 έως 60 περίπου (κόκκινη γραμμή στο x-άξονα α), ενώ στο σχήμα 3.13β καταλαμβάνει τις θέσεις από 11 έως 18 περίπου (κόκκινη γραμμή στο x-άξονα β). Το ίδιο θα παρατηρήσουμε να συμβαίνει και για τα κάτω προφίλ των εικόνων λέξεων.

Τη λύση στο πρόβλημα αυτό έδωσε η χρήση της μεθόδου της παρεμβολής *(interpolation method)*, με τέτοιο τρόπο ώστε το μήκος όλων των χαρακτηριστικών των εικόνων-λέξεων να γίνει ίδιο.

Σε αντίθεση οι [10], χρησιμοποιούν τα ίδια μεν χαρακτηριστικά με το δικό μας υποσύστημα, χωρίς όμως να τα μετατρέπουν στο ίδιο μήκος. Χρησιμοποίησαν δε για το σκοπό αυτό τον Διακριτό Μετασχηματισμό Fourier (DFT-Discrete Fourier Transform), που έχει μεγάλη υπολογιστική πολυπλοκότητα, χάνοντας όμως έτσι κάποιες από τις λεπτομέρειες των αρχικών χαρακτηριστικών.

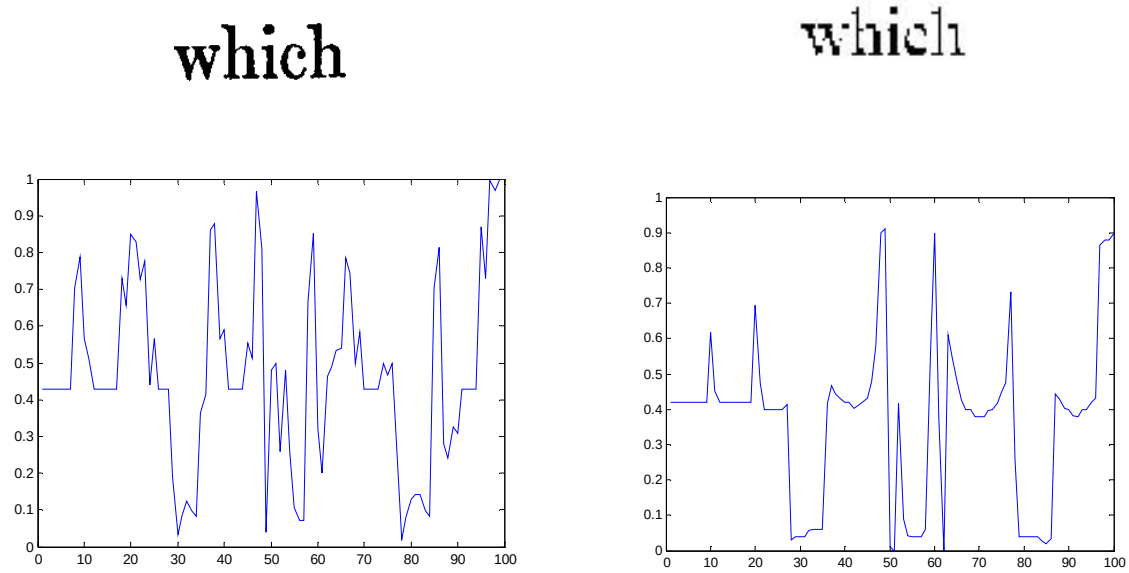
Στο προτεινόμενο υποσύστημα χρησιμοποιήθηκε ο αλγόριθμος παρεμβολής που παρουσιάστηκε στην παράγραφο 2.4. Ο αλγόριθμος αυτός δέχεται σαν είσοδο δυο ορίσματα, το διάνυσμα και το νέο μέγεθος στο οποίο θέλουμε να το ανάγουμε. Το πιο θα είναι αυτό το μέγεθος υπολογίζεται πειραματικά στο επόμενο κεφάλαιο.

Στο σχήμα 3.16 βλέπουμε τα δυο κάθετα ιστογράμματα του σχήματος 3.13 μετά την εφαρμογή της παρεμβολής με μήκος 150 και την κανονικοποίηση κατά ύψος, ενώ στο σχήμα 3.17 τα αντίστοιχα δυο πάνω προφίλ του σχήματος 3.15 κανονικοποιημένα κατά ύψος και μήκος, με μήκος 100.



(α) **(β)**

Σχήμα 3.16: Εφαρμογή Παρεμβολής στο Κάθετο Ιστόγραμμα μεγέθους 150



(α) **(β)**

Σχήμα 3.17: Εφαρμογή Παρεμβολής στο Πάνω Προφίλ μεγέθους 100

Τώρα η σύγκριση ανάμεσα στα εξαγόμενα διανύσματα των εικόνων-λέξεων μπορεί να πραγματοποιηθεί, αφού όλες οι τιμές είναι σχετικές του μεγέθους τους, ενώ ένα παράδειγμα τιμών κανονικοποημένου (κατά μήκος και ύψος) διανύσματος, έστω του πάνω προφίλ του σχήματος 3.17α φαίνεται στο παρακάτω σχήμα (3.18).

Interpolated UpperProfil (which)

0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.6170
0.4520 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.4200 0.6938
0.4760 0.4000 0.4000 0.4000 0.4000 0.4000 0.4148 0.0308 0.0400 0.0400
0.0400 0.0562 0.0600 0.0600 0.0600 0.4170 0.4672 0.4426 0.4276 0.4200
0.4200 0.4018 0.4116 0.4200 0.4312 0.4790 0.5772 0.9000 0.9104 0.0092
0 0.4158 0.0896 0.0412 0.0400 0.0400 0.0400 0.0586 0.5556 0.9000
0.3760 0 0.6104 0.5452 0.4744 0.4260 0.4000 0.4000 0.3800 0.3800
0.3800 0.3958 0.4000 0.4154 0.4496 0.4750 0.7328 0.2598 0.0400 0.0400
0.0400 0.0400 0.0400 0.0266 0.0200 0.0330 0.4428 0.4274 0.4024 0.4000
0.3820 0.3800 0.3984 0.4000 0.4200 0.4310 0.8640 0.8800 0.8800 0.9000

Σχήμα 3.18: Παράδειγμα Κανονικοποιημένου Διανύσματος κατά μήκος Πάνω Προφίλ

3.2.5 Επιπλέον βελτίωση

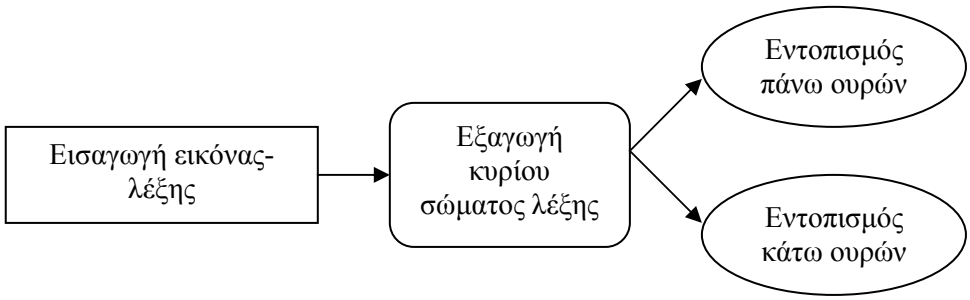
Πέρα από τα μορφολογικά χαρακτηριστικά (το κάθετο ιστόγραμμα, πάνω και κάτω προφίλ), που περιγράφουν το σχήμα και την μορφή της εικόνας-λέξης, σημαντική βελτίωση σε ολόκληρο το υποσύστημα της αναγνώρισης λέξης θα προσέφερε και η καταγραφή των ουρών που εξέχουν πάνω (*ascenders*) και κάτω (*descenders*) από την εικόνα.

Γενικά η καταγραφή της θέσης σημείων που προεξέχουν στη βιβλιογραφία είναι συχνή, αφού είναι βασικό χαρακτηριστικό των λέξεων, όπως στους [24], που χρησιμοποιείται

για κλάδεμα (pruning) των λέξεων που είναι πιθανό να μην ταιριάζουν. Μάλιστα σε πολλές περιπτώσεις, ειδικά όταν πρόκειται για αναγνώριση αρχαίων χειρόγραφων κειμένων, όπως στους [25], γίνονται προσπάθειες να αναγνωριστούν κοιλότητες και ανοίγματα (*open/closed cavities*) πάνω και κάτω από τη λέξη, ώστε να γίνει ο διαχωρισμός σε χαρακτήρες.

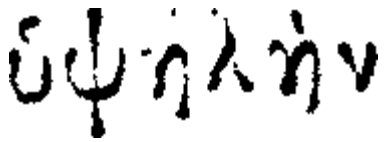
Θεωρώντας λοιπόν ότι η εξαγωγή αυτών των χαρακτηριστικών εξειδικεύει ακόμα περισσότερο την περιγραφή μιας λέξης, αποφασίστηκε να χρησιμοποιηθούν και στο δικό μας σύστημα και να ενσωματωθούν στο διάνυσμα με τα υπόλοιπα χαρακτηριστικά.

Ο εντοπισμός των πάνω και κάτω ουρών σε μια εικόνα λέξη γίνεται με βάση το σχήμα 3.19.



Σχήμα 3.19: Διαδικασία Εντοπισμού Πάνω και Κάτω Ουρών

Όπως παρατηρούμε, για να μπορέσουν να εντοπιστούν οι ουρές που προεξέχουν στην λέξη, πρέπει πρώτα να καθοριστούν τα όρια πάνω και κάτω από τα οποία θεωρούμε ότι κάποιο σημείο εξέχει. Τα όρια αυτά ορίζονται από το κυρίως σώμα μιας λέξης (*main body*). Πρόκειται δηλαδή για το τμήμα της λέξης χωρίς τις ουρές των χαρακτήρων. Στο σχήμα 3.20α βλέπουμε μια λέξη και στο σχήμα 3.20β το κυρίως σώμα της.



(α) Λέξη

(β) Κυρίως Σώμα

Σχήμα 3.20: Παράδειγμα Κυρίους Σώματος Λέξης

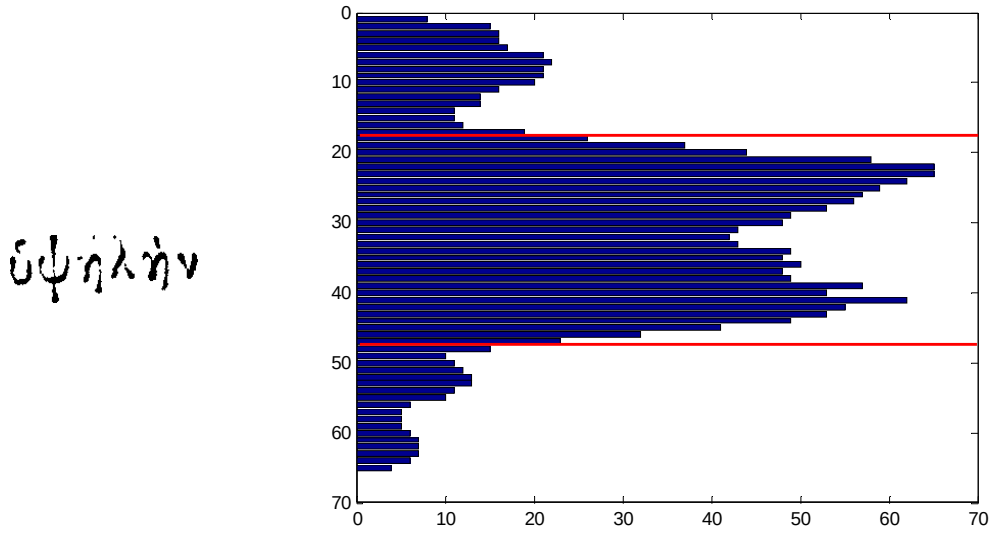
Άρα απαραίτητος είναι ο υπολογισμός του κυρίου σώματος της λέξης και για το σκοπό αυτό υλοποιήθηκε και χρησιμοποιήθηκε η συνάρτηση *mainbody*, που φαίνεται στο παρακάτω σχήμα (3.21) κατασκευασμένη στο MATLAB.

```
function g=mainbody(image)
%Βρίσκει το κυρίως σώμα της λέξης
imgHorHist=horhist(f);
Rows=size(imgHorHist);
imgMax=max(imgHorHist);
newImage=ones(Rows);
for i=1:Rows
   if imgHorHist (i)>=1/3*imgMax
       newimage(i)= imgHorHist (i);
   else
       newimage(i)=0;
   end
end
g=newimage;
```

Σχήμα 3.21: Συνάρτηση Υπολογισμού Κυρίους Σώματος Λέξης

Σύμφωνα με την παραπάνω συνάρτηση αρχικά υπολογίζεται το οριζόντιο ιστόγραμμα της λέξης, που είναι ουσιαστικά το πλήθος των μαύρων pixels σε κάθε γραμμή της εικόνας. Ως κυρίως σώμα της λέξης ορίζονται οι τιμές του οριζόντιου ιστογράμματος που ξεπερνούν το 1/3 της μέγιστης τιμής του. Τιμές που είναι μικρότερες του ορίου αυτού αντιστοιχούν σε ουρές που προεξέχουν. Το όριο αυτό είναι ικανοποιητικό, αφού έχει χρησιμοποιηθεί με επιτυχία και στους [13].

Στο σχήμα 3.22 βλέπουμε το οριζόντιο ιστόγραμμα της λέξης του σχήματος 3.20 και με κόκκινη γραμμή περικλείονται τα όρια του κυρίου σώματος. Η μέγιστη τιμή του είναι τα 65 και το 1/3 της το 21.66. Άρα απορρίπτονται οι τιμές που είναι μικρότερες του 21.66.



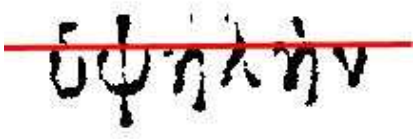
Σχήμα 3.22: Όρια Κυρίου Σώματος Λέξης στο Οριζόντιο Ιστόγραμμα

Αφού υπολογιστεί το κυρίως σώμα της λέξης είμαστε έτοιμοι να αναζητήσουμε τις ουρές που εξέχουν πάνω και κάτω.

3.2.5.1 Πάνω ουρές

Οι ουρές πάνω από το κυρίως σώμα της εικόνας είναι συνήθως είτα τόνοι στα γράμματα, είτε χαρακτηριστικό των ίδιων των γραμμάτων. Για παράδειγμα το ελληνικό γράμμα *δ* ή το λατινικό *h* κτλ.

Για να βρούμε ποιες ουρές έχει κάθε εικόνα λέξη κρατάμε από την εικόνα το τμήμα εκείνο που βρίσκεται πάνω από το κυρίως σώμα της εικόνας (σχήμα 3.23 – το πάνω τμήμα της λέξης βρίσκεται πάνω από την κόκκινη γραμμή). Η συνάρτηση που χρησιμοποιούμε για τη διαδικασία αυτή (συνάρτηση *upimage*), φαίνεται στο σχήμα 3.24 και είναι ανάλογη με την συνάρτηση εξαγωγής του κυρίους σώματος (σχήμα 3.21), με τη διαφορά ότι κρατάμε από το οριζόντιο ιστόγραμμα μόνο τις πρώτες τιμές που είναι μικρότερες από το 1/3 της μέγιστης τιμής του.



Σχήμα 3.23: Πάνω τμήμα Λέξης

```
function g=upimage(image)
%βρίσκει το τμήμα της εικόνας
%πάνω από το κυρίως σώμα
imgHorHist=horhist(f);
Rows=size(imgHorHist);
imgMax=max(imgHorHist);
newImage=ones(Rows);
for i=1:Rows
   if imgHorHist (i)<1/3*imgMax
       newimage(i)= imgHorHist (i);
   else
       break;
   end
end
g=image(1:i-1,1:end);
```

Σχήμα 3.24: Συνάρτηση Υπολογισμού Πάνω Τμήματος Λέξης

Στη συνέχεια ελέγχουμε στο κομμάτι αυτό της εικόνας που συναντάμε μαύρα pixels. Στην ουσία αυτά θα είναι και οι ουρές που εξέχουν από πάνω. Η συνάρτηση που χρησιμοποιήθηκε, υλοποιημένη σε MATLAB, φαίνεται στο σχήμα 3.25

```
function g=findascenders(image)
%Υπολογισμός ascenders
upimg=upimage(image);
vHist=verhist(upimg);
[Rows,Columns]=size(image);
ascenders=zeros(1,Columns);
isfirst=1;
for i=1:Columns
    if vHist(i)>0 & isfirst==1
          ascenders(i)=i/Rows;
          isfirst=0;
      elseif vHist(i)==0
          isfirst=1;
    end
end
g=ascenders;
```

Σχήμα 3.25: Συνάρτηση Εντοπισμού Πάνω Ουρών Λέξης

Όπως βλέπουμε αφού πάρουμε το πάνω τμήμα της εικόνας-λέξης, υπολογίζουμε το οριζόντιο ιστόγραμμά της. Στη συνέχεια το διατρέχουμε κατά μήκος (σε όλες τις στήλες) και ελέγχουμε αν θα βρούμε μαύρο pixel. Σε όποια στήλη βρεθεί κοιτάμε αν είναι πρώτο (αν πριν υπήρχαν άσπρα pixels) κι αν ναι καταγράφουμε τη στήλη την οποία βρέθηκε σε σχέση με το ύψος της εικόνας (δηλαδή διαιρούμενο με το πλήθος των γραμμών). Έτσι παίρνουμε ένα διάνυσμα που έχει 0 για τις στήλες που δεν αποτελούν αρχή ουράς (ascender) και μη μηδενικές τιμές για κάθε πρώτη στήλη που εμφανίζονται ουρές. Ένα παράδειγμα τέτοιου διανύσματος για την λέξη στο σχήμα 3.23 βλέπουμε στο σχήμα 3.26.

Ascenders(υψηλήν)

0 0 0 0 0 0 0 0.1231 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0.6154 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1.2615 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1.5692 0 0 0 0 0 1.6615 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.1692 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Σχήμα 3.26: Παράδειγμα Διανύσματος Πάνω Ουρών

Φυσικά καταλαβαίνουμε ότι αφού εξαχθεί και αυτό το χαρακτηριστικό της εικόνας θα πρέπει να κανονικοποιηθεί κατά μήκος με παρεμβολή για τους ίδιους λόγους που αναφέραμε στην παράγραφο 3.2.4.

3.2.5.2 Κάτω ουρές

Αντίστοιχα οι ουρές κάτω από την εικόνα-λέξη (*descenders*) είναι χαρακτηριστικά των ίδιων των γραμμάτων, όπως το γράμμα *γ* του ελληνικού αλφαβήτου ή το γράμμα *g* του λατινικού αλφαβήτου.

Για να βρούμε ποιες ουρές έχει κάθε εικόνα λέξη κρατάμε από την εικόνα το τμήμα εκείνο που βρίσκεται κάτω από το κυρίως σώμα της εικόνας (σχήμα 3.27 – το κάτω τμήμα της λέξης βρίσκεται κάτω από την κόκκινη γραμμή). Η συνάρτηση που χρησιμοποιούμε για τον υπολογισμό αυτό (συνάρτηση *downimage*), φαίνεται στο σχήμα 3.28 και όπως παρατηρούμε ουσιαστικά, αφού υπολογίσει το πάνω τμήμα της εικόνας και το κυρίως σώμα τα αφαιρεί από την αρχική εικόνα, με τμήμα της εικόνας που μένει να είναι το κάτω.

Σχήμα 3.27: Κάτω Τμήμα Λέξης

```
function g=downimage(image)
%βρίσκει το τμήμα της εικόνας
%κάτω από το κυρίως σώμα
%υπολογισμός πάνω τμήματος
upimg=upimage(image);
[Rows1,Columns1]=size(upimg);
%υπολογισμός κυρίου σώματος
mainimg=mainbody(image);
[Rows2,Columns2]=size(mainimg);

g=image(Rows1+Rows2+2:end,1:end);
```

Σχήμα 3.28: Συνάρτηση Υπολογισμού Κάτω Τμήματος Λέξης

Στη συνέχεια με τρόπο αντίστοιχο με αυτόν του εντοπισμού των πάνω ουρών, βρίσκουμε τις κάτω ουρές (descenders). Η συνάρτηση υλοποιημένη σε MATLAB φαίνεται στο σχήμα 3.29.

```
function g=finddescenders(image);
%Υπολογισμός descenders
downimg=downimage(image);
vhist=verhist(downimg);
[Rows,Columns]=size(image);
descenders=zeros(1,Columns);
isfirst=1;
for i=1:Columns
    if vhist(i)>0 & isfirst==1
          descenders(i)=i/Rows;
          isfirst=0;
      elseif vhist(i)==0
          isfirst=1;
    end
end
g=descenders;
```

Σχήμα 3.29: Συνάρτηση Εντοπισμού Κάτω Ουρών Λέξης

Στη συνάρτηση αυτή αφού εντοπίσουμε το κάτω τμήμα της εικόνας-λέξης, υπολογίζουμε το οριζόντιο ιστόγραμμά της. Το διατρέχουμε κατά μήκος (σε όλες τις στήλες) και ελέγχουμε αν θα βρούμε μαύρο pixel. Σε όποια στήλη βρεθεί κοιτάμε αν είναι πρώτο (αν πριν υπήρχαν άσπρα pixels) κι αν ναι καταγράφουμε τη στήλη την οποία βρέθηκε σε σχέση με το ύψος της εικόνας (δηλαδή διαιρούμενο με το πλήθος των γραμμών). Καταλήγουμε λοιπόν σε ένα διάνυσμα που έχει 0 για τις στήλες που δεν αποτελούν αρχή ουράς (descender) και μη μηδενικές τιμές για κάθε πρώτη στήλη που εμφανίζονται ουρές. Ένα παράδειγμα τέτοιου διανύσματος για την λέξη στο σχήμα 3.27 βλέπουμε στο σχήμα 3.30.

Descenders(υψηλήν)

0	*0.0308*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0.4923*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*1.1231*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
1.2923	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*1.5385*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*1.7846*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2.0923*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2.2923*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*

0 0 0 0 0 0 0 2.6769 0 0 0 0 0 0 0 0 0 0

Σχήμα 3.30: Παράδειγμα Διανύσματος Κάτω Ουρών

Και το διάνυσμα που προκύπτει από την παραπάνω διαδικασία πρέπει να κανονικοποιηθεί κατά μήκος με παρεμβολή για να είναι συγκρίσιμο.

3.2.5.3 Εξομάλυνση

Ένα βασικό θέμα της ανάλυσης δεδομένων (data analysis) είναι το ταίριασμα ενός μοντέλου σε δεδομένα, η μορφή του οποίου καθορίζεται από έναν μικρό αριθμό παραμέτρων. Η διαδικασία αυτή περιλαμβάνει δυο βήματα:

1. εντοπισμό του κατάλληλου μοντέλου
2. έλεγχο να δούμε αν το μοντέλο αυτό ταιριάζει στα δεδομένα

Τα μοντέλα αυτά μπορεί να είναι ευέλικτα, αλλά δεν ταιριάζουν σε όλα τα δεδομένα. Αν εναλλακτικά το μοντέλο αυτό μπορεί να εξομαλυνθεί το ίδιο και τα δεδομένα τότε μεγαλώνει το πεδίο των περιπτώσεων όπου αυτά ταιριάζουν.

Η εξομάλυνση (*smoothing*) αποτελεί συχνή και σημαντική διαδικασία στον τομέα της επεξεργασίας σημάτων.

Στην περίπτωσή μας προσπαθούμε να ταιριάξουμε τα χαρακτηριστικά που εξάγουμε από εικόνες-λέξεις, οι οποίες προέρχονται από διαφορετικά ιστορικά έγγραφα, από χειρόγραφο κείμενο και τυπωμένο κείμενο, όπως θα δούμε και στο κεφάλαιο 4 με τα πειράματα.

Έτσι χρησιμοποιώντας εξομάλυνση στις τιμές των χαρακτηριστικών (κάθετο ιστόγραμμα, πάνω και κάτω προφίλ) υπάρχει μεγάλη πιθανότητα να βελτιωθούν τα αποτελέσματα. Ειδικά σε χειρόγραφο κείμενο που προέρχεται από διαφορετικούς συγγραφείς, η βελτίωση είναι μεγάλη.

Ουσιαστικά με την εξομάλυνση επιτυγχάνεται μια γενικότερη περιγραφή της εικόνας-λέξης, που είναι πιθανότεροο να ταιριάζει στο σχήμα της πραγματικής λέξης. Βέβαια οι πειραματισμοί που θα αναλυθούν στο επόμενο κεφάλαιο περιλαμβάνουν εξομάλυνση στις τιμές των χαρακτηριστικών των δεδομένων εκπαίδευσης (training set) και των δεδομένων ελέγχου (testing set), ή σε ένα από αυτά, ανάλογα με την περίπτωση κάθε φορά.

Συγκεκριμένα χρησιμοποιήσαμε εξομάλυνση 3, 5, 7 και 9 σημείων, για να πειραματιστούμε. Αυτές ορίζονται ως εξής:

- *Εξομάλυνση 3 σημείων*: Διατρέχουμε το διάνυσμα και η τιμή κάθε σημείου, έστω x, αντικαθίσταται από τον μέσο όρο της τιμής του ιδίου και των τιμών των δυο γειτονικών του, δηλαδή: *Νέα τιμή x= ((x-1)+x+(x+1))/3*
- *Εξομάλυνση 5 σημείων*: Διατρέχουμε το διάνυσμα και η τιμή κάθε σημείου, έστω x, αντικαθίσταται από τον μέσο όρο της τιμής του ιδίου και των τιμών των τεσσάρων γειτονικών του.
- *Εξομάλυνση 7σημείων*: Διατρέχουμε το διάνυσμα και η τιμή κάθε σημείου, έστω x, αντικαθίσταται από τον μέσο όρο της τιμής του ιδίου και των τιμών των έξι γειτονικών του.
- *Εξομάλυνση 9 σημείων*: Διατρέχουμε το διάνυσμα και η τιμή κάθε σημείου, έστω x, αντικαθίσταται από τον μέσο όρο της τιμής του ιδίου και των τιμών των οχτώ γειτονικών του.

Στο σχήμα 3.31α βλέπουμε την συνάρτηση της εξομάλυνσης 3 σημείων και δίπλα στο σχήμα 3.31β την εξομάλυνση 5 σημείων, υλοποιημένες σε MATLAB, με βάση τους παραπάνω ορισμούς. Αντίστοιχες είναι και οι συναρτήσεις για την εξομάλυνση 7 και 9 σημείων.

```
function g=smoothing3(Vector)
[Rows,Cols]=size(Vector);
newVector(1)=Vector(1);
for i=2:Cols-1
  newVector(i)=(Vector(i-1)
+ Vector(i)+Vector(i+1))/3;
end
newVector(Cols)=Vector(Cols);
g=newVector;
```

(α) Εξομάλυνση 3

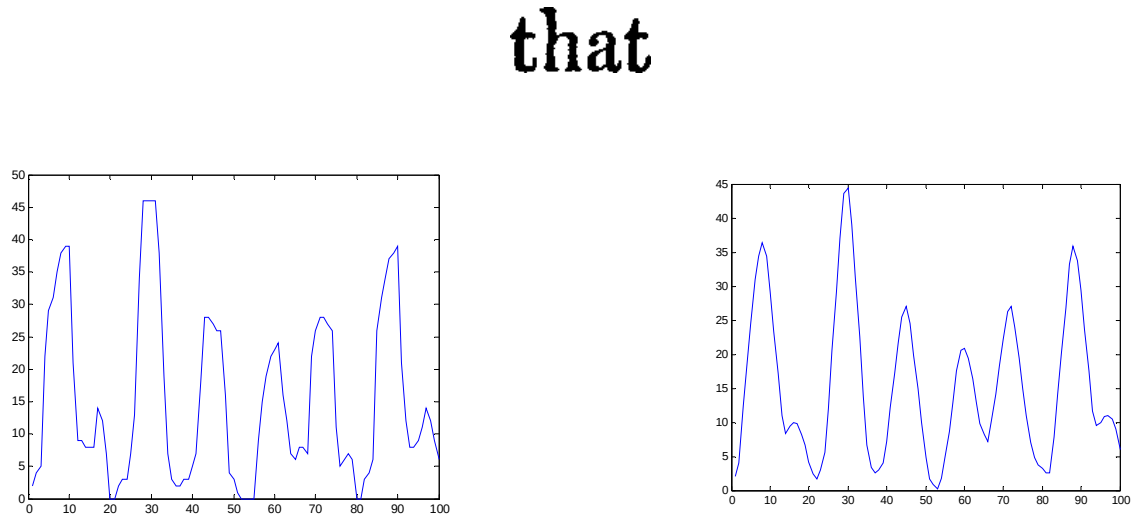
```
function g=smoothing5(Vector)
[Rows,Cols]=size(Vector);
newVector(1)=Vector(1);
newVector(2)=f(2);
for i=3:Cols-2
    newVector(i)=(Vector(i-2)
    +Vector(i-1) +Vector(i)+
   Vector(i+1)+Vector(i+2))/5;
end
newVector(Cols-1)=Vector(Cols-
1);
newVector(C)=Vector(C);
g=newVector;
```

(β) Εξομάλυνση 5

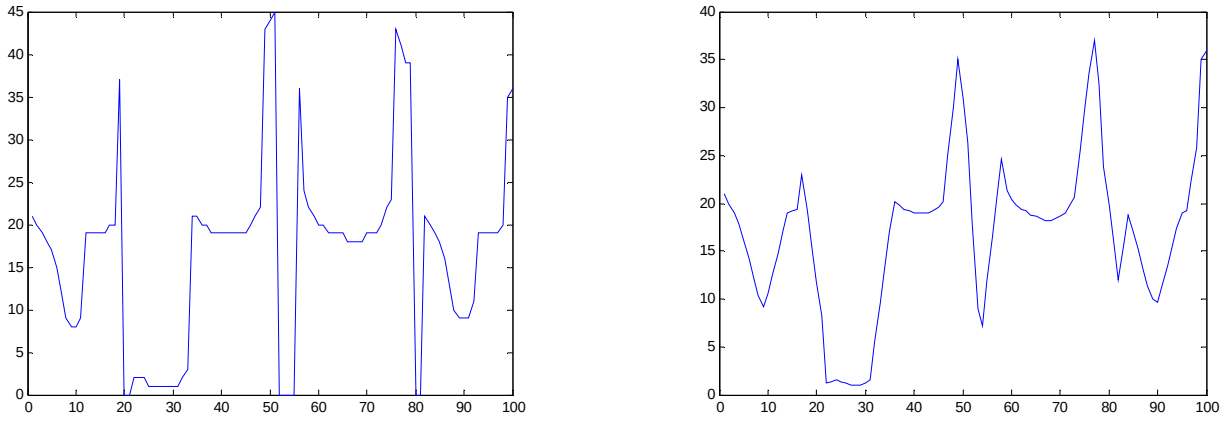
Σχήμα 3.31: Συναρτήσεις Εφαρμογής Εξομάλυνσης

Έτσι εφαρμόζονται στις διάφορες περιπτώσεις πειραμάτων για να δούμε ποια ταιριάζει καλύτερα ή όχι. Βέβαια η εφαρμογή γίνεται στα κανονικοποιημένα διανύσματα των χαρακτηριστικών, ώστε να εξομαλυνθούν και να δίνουν μια γενικότερη αναπαράσταση του σχήματος της λέξης, πριν συγκριθούν για τις διάφορες εικόνες λέξεις.

Στο σχήμα 3.32 βλέπουμε για μια εμφάνιση λέξης στη αριστερή πλευρά τις γραφικές παραστάσεις των χαρακτηριστικών της, κάθετο ιστόγραμμα, πάνω και κάτω προφίλ, ενώ στη δεξιά πλευρά τις εξομαλυσμένες γραφικές παραστάσεις τους αντίστοιχα 5 σημείων.



Κάθετο Ιστόγραμμα



(α) Απλά **(β) Εξομάλυνση 5**

Σχήμα 3.32: Χαρακτηριστικά Λέξεων Απλά και με Χρήση Εξομάλυνσης

Παρατηρούμε ότι οι γραφικές παραστάσεις στην δεξιά στήλη είναι πιο ομαλές από αυτές στην αριστερή, άρα μπορούν να αντιπροσωπεύουν περισσότερες εμφανίσεις της ίδιας λέξης, αφού απαλλάσσονται από κάποιες λεπτομέρειες.

Σε κάποιες όμως περιπτώσεις, όπως θα δούμε στα πειράματα η εξομάλυνση δεν βελτιώνει την απόδοση του συστήματος, ενώ η μεγάλη εξομάλυνση συνήθως την μειώνει και η αναγνώριση λέξεων γίνεται πιο δύσκολη, λόγω του ότι χάνεται σημαντική πληροφορία του σχήματος των λέξεων.

Πρέπει να τονιστεί ότι η εξομάλυνση δεν εφαρμόζεται στα διανύσματα των ουρών πάνω (ascenders) και κάτω (descenders) από τη λέξη. Ούτως ή άλλως οι στήλες που περιέχουν μη μηδενικές τιμές στα διανύσματα αυτά περικλείονται από μηδενικές τιμές (βλ σχήμα 3.26 και σχήμα 3.30), με αποτέλεσμα οι μη μηδενικές τιμές να τείνουν στο 0, χωρίς τελικά η εξομάλυνση να προσφέρει κάποια σημαντική αλλαγή.

3.2.6 Κατηγοριοποίηση με k-Means

Αφού υπολογιστούν οι τιμές των εικόνων-λέξεων για το κάθετο ιστόγραμμα, το πάνω και κάτω προφίλ, τις ουρές πάνω και κάτω, κανονικοποιηθούν κατά ύψος και μήκος, εφαρμοστεί εξομάλυνση ή όχι, ανάλογα με την περίπτωση, εξάγεται ένα συνολικό διάνυσμα. Για παράδειγμα αν εφαρμοστεί παρεμβολή 175 θέσεων το διάνυσμα αυτό θα είναι 875 τιμών (κάθε χαρακτηριστικό 175 τιμές). Η διαδικασία αυτή πραγματοποιείται για κάθε εικόνα-λέξη που εισάγεται στο σύστημά μας.

Το επόμενο βήμα είναι η κατηγοριοποίηση των εικόνων σε κλάσεις, ώστε διαφορετικές εμφανίσεις των ίδιων λέξεων να ανήκουν στην ίδια κλάση. Για το σκοπό αυτό χρησιμοποιήθηκε ο k-Means αλγόριθμος, όπως περιγράφηκε στην παράγραφο 2.5.

Όπως αναφέρθηκε, ο αλγόριθμος αυτός δέχεται τις πλήθος των κλάσεων στο οποίο θέλουμε να κατηγοριοποιήσουμε τις λέξεις, αρχικοποιεί τις κλάσεις, κατανέμει τα δεδομένα σε αυτές και υπολογίζει το κέντρο κάθε κλάσης. Στη συνέχεια τα δεδομένα κατηγοριοποιούνται στην κλάση που είναι πιο κοντά και η διαδικασία επαναλαμβάνεται.

Στο υποσύστημά μας χρησιμοποιήθηκε μια υλοποίηση του k-Means υλοποιημένη σε MATLAB από τον Kardi Teknomo, με μικρές αλλαγές. Η υλοποίηση αυτή φαίνεται στο σχήμα 3.33.

```
function [y,Centroids]=kMeansCluster(m,k,isRand)
if nargin<3,        isRand=0;   end
if nargin<2,        k=1;        end

[maxRow, maxCol]=size(m);
if maxRow<=k,
    y=[m, 1:maxRow];
else
     % initial value of centroid
    if isRand,
        p = randperm(size(m,1));      % random initialization
        for i=1:k
            c(i,:)=m(p(i),:);
        end
    else
        for i=1:k
           c(i,:)=m(i,:);        % sequential initialization
       end
    end
     temp=zeros(maxRow,1);   % initialize as zero vector
```

```
      while 1,
        d=DistMatrix2(m,c);% calculate objcets-centroid
                           %distances
        [z,g]=min(d,[],2);  % find group matrix g
        if g==temp,
            break;          % stop the iteration
        else
            temp=g;  % copy group matrix to temporary variable
        end
        for i=1:k
            f=find(g==i);
            if f    % only compute centroid if f is not empty
                c(i,:)=mean(m(find(g==i),:),1);
            end
        end
      end
      y=[m,g];
    Centroids=c;
end
```

Σχήμα 3.33: Συνάρτηση Αλγορίθμου k-Means

Όπως είπαμε η συνάρτηση αυτή έχει ως σκοπό την κατηγοριοποίηση των δεδομένων σε κλάσεις, με βάση κάποια χαρακτηριστικά. Το βασικό μέτρο σύγκρισης στην προκειμένη περίπτωση είναι η ελαχιστοποίηση της Ευκλείδειας απόστασης ανάμεσα στα κεντρικά σημεία των κλάσεων και τα σημεία των δεδομένων.

Ο *kMeansCluster* δέχεται σαν είσοδο ένα πίνακα δεδομένων *m*, όπου κάθε δεδομένο καταλαμβάνει μια γραμμή στον πίνακα και οι τιμές των χαρακτηριστικών είναι οι στήλες του. Ένα παράδειγμα του πίνακα εισαγωγής για τέσσερις εικόνες λέξεις είναι:

m=[Κάθετο Ιστόγραμμα1 Πάνω Προφίλ1 Κάτω Προφίλ1 Πάνω Ουρές1 Κάτω Ουρές1

Κάθετο Ιστόγραμμα2 Πάνω Προφίλ2 Κάτω Προφίλ2 Πάνω Ουρές2 Κάτω Ουρές2

Κάθετο Ιστόγραμμα3 Πάνω Προφίλ3 Κάτω Προφίλ3 Πάνω Ουρές3 Κάτω Ουρές3

Κάθετο Ιστόγραμμα4 Πάνω Προφίλ4 Κάτω Προφίλ4 Πάνω Ουρές4 Κάτω Ουρές4]

Στη συνέχεια αυτό που κάνει είναι να κατηγοριοποιεί τις εικόνες-λέξεις σε τόσες κλάσεις όσες δείχνει η μεταβλητή *k*, που δέχεται ως δεύτερο όρισμα, χρησιμοποιώντας είτε τυχαία αρχικοποίηση των κεντρικών σημείων των κλάσεων, αν το τρίτο όρισμα *isRand* είναι 1, είτε σε οποιαδήποτε άλλη περίπτωση αναθέτει ως κεντρικά σημεία τα *k* πρώτα δεδομένα.

Η συνάρτηση, αφού κάνει όλη την εργασία της κατανομής σε κλάσεις με βάση την Ευκλείδεια απόσταση, εύρεση κεντρικών σημείων, κατανομή στις νέες κλάσεις κοκ,

επιστρέφει τον ίδιο πίνακα δεδομένων *m* με την προσθήκη μιας στήλης στο τέλος. Αυτή δείχνει την κλάση στην οποία ανήκει το κάθε δεδομένο, δηλαδή στην περίπτωση μας την κλάση που ανήκει η κάθε λέξη. Επίσης επιστρέφει και τις τιμές των κεντρικών σημείων της τελικής κατηγοριοποίησης, στο διάνυσμα *Centroids*. Έτσι για παράδειγμα αν ο πίνακας *m* έχει τις τιμές:

m = [1 1; 2 1; 4 3; 5 4] με k=2

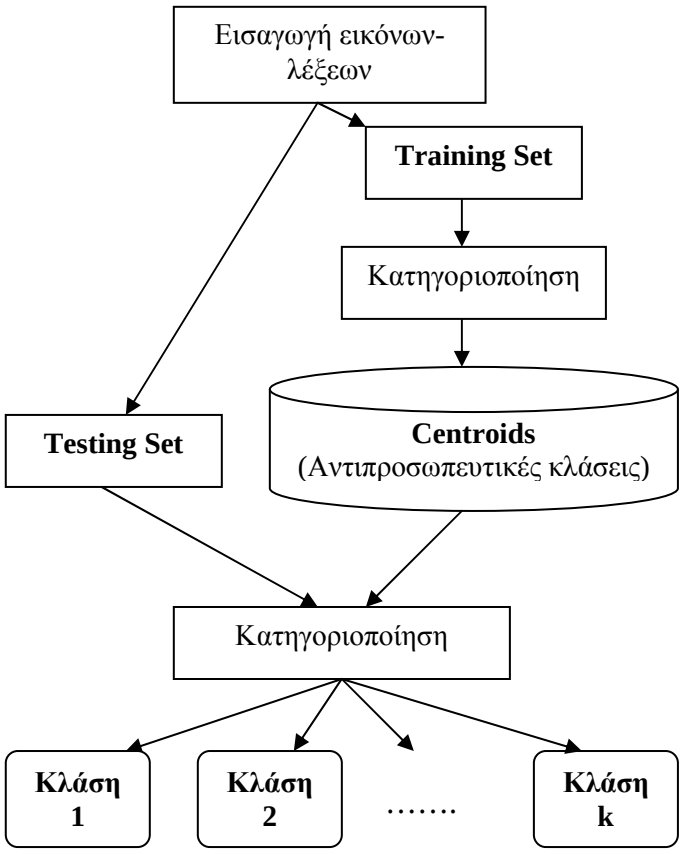
ο αλγόριθμος θα παράγει την παρακάτω κατηγοριοποίηση:

m = [1 1 1; 2 1 1; 4 3 2; 5 4 2]

που σημαίνει ότι τα δυο πρώτα δεδομένα με τιμές χαρακτηριστικών [1 1] και [2 1] θα ανήκουν στην πρώτη κλάση, ενώ τα δύο τελευταία με τιμές [4 3] και [5 4] θα ανήκουν στην δεύτερη κλάση.

3.3 Το υποσύστημά μας

Το υποσύστημα που προτείνουμε συγκεντρώνει όλα τα παραπάνω στάδια. Ουσιαστικά η διαδικασία της κατηγοριοποίησης εκτελείται δύο φορές. Μία για την εκπαίδευση των δεδομένων και μία για τις εισαγόμενες εικόνες λέξεις, όπως φαίνεται στο σχήμα 3.34.



Σχήμα 3.34: Διαδικασία Κατηγοριοποίησης στο Σύστημά μας

Όπως βλέπουμε παραπάνω, κάποιες εικόνες λέξεις θα αποτελέσουν το σύνολο των δεδομένων εκπαίδευσης και κάποιες θα είναι τα δεδομένα ελέγχου.

- ✓ **Δεδομένα Εκπαίδευσης (Training Set)**

  Διαβάζουμε τις λέξεις, που αποτελούν τα δεδομένα εκπαίδευσης, εξάγουμε τα χαρακτηριστικά τους, τα κανονικοποιούμε και με εφαρμογή του k-Means τα κατανέμουμε σε κλάσεις. Από αυτές θα πάρουμε τα κεντρικά σημεία (Centroids), που θα χρησιμοποιηθούν για μετέπειτα κατηγοριοποίηση

- ✓ **Δεδομένα Ελέγχου (Testing Set)**

  Διαβάζουμε τις λέξεις που έρχονται στο σύστημα για αναγνώριση, εξάγουμε τα χαρακτηριστικά τους, τα κανονικοποιούμε και στη συνέχεια τα εισάγουμε στον k-Means μαζί με τα κεντρικά σημεία του προηγούμενου βήματος. Τα σημεία αυτά είναι οι αντιπροσωπευτικές κλάσεις και για κάθε νέα εικόνα-λέξη θα υπολογίζεται η Ευκλείδεια απόσταση της από αυτές και θα γίνεται η τελική κατηγοριοποίηση τους.

Στόχος είναι οι τελικές κλάσεις να περιέχουν εμφανίσεις των ίδιων λέξεων, ανεξαρτήτως των εγγράφων προέλευσή τους, με όσο το δυνατόν λιγότερες λάθος κατηγοριοποιήσεις.

Διάφορα είδη δεδομένων εκπαίδευσης και δεδομένων ελέγχου θα δούμε στο επόμενο κεφάλαιο με τα πειράματα που έχουμε πραγματοποιήσει.

Κεφάλαιο 4

Πειράματα

4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα δούμε κάποια πειράματα που έγιναν με σκοπό την αξιολόγηση του παραπάνω υποσυστήματος αναγνώρισης λέξης. Πραγματοποιήθηκαν λοιπόν, διάφοροι τύποι πειραμάτων με εικόνες-λέξεις που προήλθαν από διάφορα έγγραφα. Αντίθετα οι [10] αξιολόγησαν το σύστημά τους μόνο με 20 σελίδες χειρόγραφου κειμένου του George Washington από τη βιβλιοθήκη του Κογκρέσου τόσο για εκπαίδευση όσο και για έλεγχο.

Εμείς πειραματιστήκαμε με διάφορα σύνολα λέξεων, τα οποία τα χρησιμοποιήσαμε μόνο για εκπαίδευση ή μόνο για έλεγχο. Επίσης στα πειράματα που αφορούν χειρόγραφο κείμενο, εμείς χρησιμοποιήσαμε εικόνες-λέξεις από διαφορετικούς συγγραφείς, σε αντίθεση με τους [10], που ασχολήθηκαν μόνο με χειρόγραφο ενός συγγραφέα.

4.2 Πειραματικά Δεδομένα

Τα δεδομένα τα οποία χρησιμοποιήσαμε στα πειράματά μας, όπως είπαμε, προέρχονται από διαφορετικά έγγραφα.

Όσον αφορά την προέλευση των εικόνων-λέξεων ιστορικού κειμένου χρησιμοποιήθηκαν τρία ιστορικά βιβλία. Αυτά είναι:

1. *Travels in Italy, Greece and the Ionian Islands, H.W. Williams, Edinburgh 1820*
2. *Απομνημονεύματα επί της σύγχρονης ιστορίας, Σπυρίδωνος Μαλάκη, Εν Αθήναις 1895*
3. *Σκιαγραφία Αργοστολίου του 1821, Σάββα Αννίνος*

Από το πρώτο πήραμε αγγλικές λέξεις (σχήμα 4.1α), ενώ το δεύτερο μας και το τρίτο μας βοήθησε με πειράματα σε ελληνικό ιστορικό κείμενο, του οποίου η γραφή είναι πολυτονική (σχήμα 4.1 β).

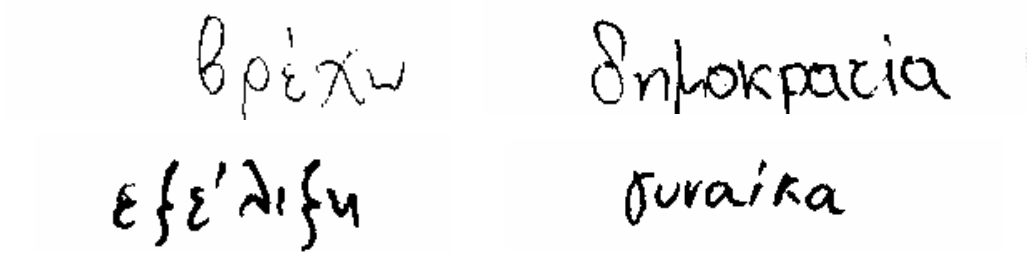
(α) Αγγλικές Λέξεις

(β) Ελληνικές Λέξεις

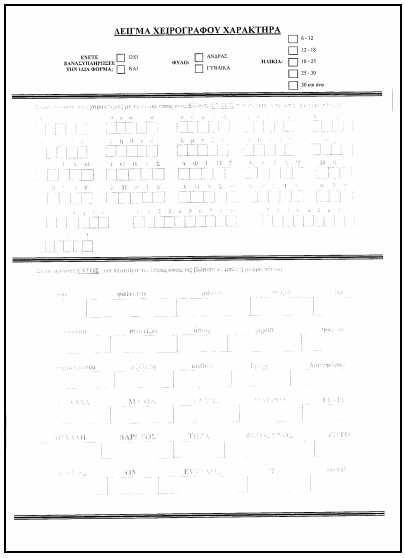
Σχήμα 4.1: Παραδείγματα Λέξεων που Εισάγονται στο Σύστημα

Το τυπωμένο κείμενο, που χρησιμοποιήθηκε κυρίως για εκπαίδευση, είναι γραμματοσειρά Times New Roman με μέγεθος 12.

Ενώ τέλος, το χειρόγραφο κείμενο είναι 15 διαφορετικές ελληνικές λέξεις που προήλθαν από 15 διαφορετικούς συγγραφείς, σύνολο 225 λέξεις (σχήμα 4.2). Οι λέξεις αυτές προήλθαν από τη βάση δεδομένων GCDB, που κατασκεύασαν οι Μαργαρώνης Ιωάννης και Χρήστου Μηνάς, για το Πανεπιστήμιο Αιγαίου-Τμήμα Μηχανικών Η/Υ και Πληροφοριακών και Επικοινωνιακών Συστημάτων με τη συμπλήρωση από διαφορετικούς συγγραφείς φόρμας της μορφής του σχήματος 4.3.



Σχήμα 4.2: Λέξεις από Ελληνική Βάση



Σχήμα 4.3: Παράδειγμα Φόρμας Ελληνικής Βάσης

4.3 Παράμετροι

Στα πειράματα που ακολουθούν θεωρείται απαραίτητη η μεταβολή κάποιων παραμέτρων, ώστε να δούμε σε ποιες περιπτώσεις εμφανίζονται καλύτερα αποτελέσματα. Όπως είναι φυσικό ανάλογα με τα δεδομένα εκπαίδευσης και ελέγχου που χρησιμοποιούμε, καλύτερα αποτελέσματα επιτυγχάνονται κάθε φορά με διαφορετικό συνδυασμό παραμέτρων.

Τέτοιες παράμετροι είναι αρχικά το πλήθος των δεδομένων με τα οποία γίνεται η εκπαίδευση, το μέγεθος της παρεμβολής που επιλέγεται και φυσικά το ποσοστό της εξομάλυνσης.

Στη συνέχεια θα τα δούμε πιο αναλυτικά.

4.3.1 Πλήθος Δεδομένων Εκπαίδευσης

Σε πειράματα κατηγοριοποίησης σημαντικό ρόλο παίζει ο αριθμός των δεδομένων που χρησιμοποιούνται για εκπαίδευση. Αντικειμενικός σκοπός είναι να επιτυγχάνονται καλύτερα αποτελέσματα με όσο το δυνατό μικρότερο αριθμό δεδομένων εκπαίδευσης και μεγαλύτερο αριθμό δεδομένων ελέγχου. Τότε θα θεωρείται ότι ο αλγόριθμος αναγνώρισης λέξεων που έχουμε χρησιμοποιήσει είναι καλός.

Στην περίπτωση μας, στα περισσότερα πειράματα που ακολουθούν, το σύνολο αυτό των δεδομένων είναι μικρό.

4.3.2 Μέγεθος Παρεμβολής

Μια πολύ σημαντική παράμετρος στο σύστημά μας προκύπτει κατά τη δημιουργία του διανύσματος των τιμών των χαρακτηριστικών των εικόνων-λέξεων και δεν είναι άλλη από το μέγεθος της παρεμβολής (interpolation).

Το ποιο είναι το κατάλληλο μέγεθος εξαρτάται κάθε φορά από τα δεδομένα που έχουμε. Οι δοκιμές που θα φανούν στη συνέχεια χρησιμοποιούν τέσσερα μεγέθη παρεμβολής.

- *175 pixels*: Είναι και το αρχικό μέγεθος που επιλέχτηκε δεδομένου ότι αντιστοιχεί σε ένα μέσο μέγεθος για λέξεις 6 έως 7 γραμμάτων, (αυτό θεωρήθηκε ότι είναι ένα μέσο μέγεθος λέξης που συναντάμε στην βιβλιογραφία)
- *67 pixels*: Αντιστοιχεί σε ένα μέσο μέγεθος λέξεων 2 έως 3 γραμμάτων
- *129 pixels*: Αντιστοιχεί σε ένα μέσο μέγεθος λέξεων 4 έως 5 γραμμάτων
- *230 pixels*: Αντιστοιχεί σε ένα μέσο μέγεθος λέξεων 8 έως 9 γραμμάτων

Στα πειράματα παρακάτω θα δούμε αναλυτικά τα αποτελέσματα για κάθε μέγεθος παρεμβολής.

4.3.3 Εξομάλυνση

Στη συνέχεια όπως είδαμε και στην παράγραφο 3.2.5.3 προσπαθούμε να βελτιώσουμε τα αποτελέσματα εξομαλύνοντας τα εξαγόμενα χαρακτηριστικά των εικόνων-λέξεων. Υλοποιήσαμε τέσσερις μορφές εξομάλυνσης:

- *3 σημείων*
- *5 σημείων*
- *7 σημείων* και
- *9 σημείων*

Σε κάποιες περιπτώσεις τα αποτελέσματα δεν επηρεάζονται από την εξομάλυνση και μάλιστα μειώνεται το ποσοστό επιτυχίας, λόγω του ότι χάνεται απαραίτητη για το σύστημα πληροφορία του σχήματος των λέξεων. Σε άλλες όμως περιπτώσεις τα αποτελέσματα βελτιώνονται σημαντικά.

Έτσι στα πειράματα που ακολουθούν θα δούμε στα αποτελέσματα διάφορες μεταβολές της παραμέτρου αυτής. Επίσης σε κάποιες περιπτώσεις εφαρμόζεται εξομάλυνση μόνο στα δεδομένα ελέγχου κι όχι στα δεδομένα εκπαίδευσης.

4.3.4 Αριθμός Συγγραφέων

Ένας πολύ σημαντικός παράγοντας που επηρεάζει κατά πολύ τα εξαγόμενα αποτελέσματα είναι ο αριθμός των συγγραφέων. Το σύστημα αντιδρά καλύτερα αν τα δεδομένα ελέγχου προέρχονται από έναν μόνο συγγραφέα, όπως έκαναν και οι [10], γιατί είναι φυσικό οι λέξεις ίδιου γραφικού χαρακτήρα να μοιάζουν μεταξύ τους.

Εμείς προχωρήσαμε ακόμα παραπάνω και στις δοκιμές που ακολουθούν θα δούμε τα αποτελέσματα χρήσης λέξεων που προέρχονται από διαφορετικούς συγγραφείς, φτάνοντας σε κάποιες περιπτώσεις τους 15.

4.4 Μέτρο αξιολόγησης

Σαν μέτρο αξιολόγησης των πειραμάτων θεωρήθηκε το ποσοστό επιτυχίας, που δείχνει ουσιαστικά το ποσοστό των λέξεων που κατηγοριοποιήθηκε σωστά σε σχέση με το συνολικό αριθμό τους που εισήχθηκε στο υποσύστημα.

Παρακάτω θα δούμε το συνολικό ποσοστό επιτυχίας σε κάθε περίπτωση δεδομένων εκπαίδευσης και δεδομένων ελέγχου, με μεταβολή των διαφόρων παραμέτρων, αλλά και τα επιμέρους ποσοστά επιτυχίας ανά κλάση. Αυτό το τελευταίο αναφέρεται στο ποσοστό των λέξεων μιας κλάσης που κατηγοριοποιήθηκαν σωστά σε σχέση με τον συνολικό αριθμό των λέξεων αυτής της κλάσης που εισήχθησαν αρχικά στο σύστημα.

4.5 Περιγραφή Πειραμάτων

Στην παράγραφο αυτή θα γίνει αναλυτική περιγραφή των πειραμάτων που πραγματοποιήθηκαν.

4.5.1 Εκπαίδευση και Έλεγχος σε Ιστορικό Ενός Συγγραφέα

Στην περίπτωση αυτή οι λέξεις τόσο της εκπαίδευσης όσο και του ελέγχου προέρχονται από το βιβλίο *Travels in Italy, Greece and the Ionian Islands* του *H.W. Williams, Edinburgh 1820.* Επιλέχτηκαν 360 λέξεις που κατηγοριοποιήθηκαν σε 13 κλάσεις, χρησιμοποιώντας μια εμφάνιση κάθε λέξης ως το σύνολο δεδομένων εκπαίδευσης.

Τα χαρακτηριστικά των λέξεων που χρησιμοποιήθηκαν ήταν το κάθετο ιστόγραμμα, το πάνω και κάτω προφίλ και η παρεμβολή είχε μέγεθος 175 pixels, χωρίς να εφαρμοστεί εξομάλυνση.

Τα αποτελέσματα φαίνονται στο πινακάκι 4.1

Κλάση	Λέξεις	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	the	92	89	96,73913%
2	and	41	38	92,68293%
3	of	106	106	100%
4	that	19	19	100%
5	than	4	4	100%
6	his	32	32	100%
7	her	13	13	100%
8	with	16	16	100%
9	which	9	9	100%
10	from	5	4	100%
11	it	7	6	85,71429%
12	at	13	13	100%
13	Spain	3	3	100%
	Σύνολο	**360**	**353**	
Ποσοστό Επιτυχίας:			**97.77778%**	

Πίνακας 4.1: Κατηγοριοποίηση με χρήση κάθετου ιστογράμματος, πάνω και κάτω προφίλ

Στη συνέχεια προσθέτουμε στα χαρακτηριστικά τις πάνω και κάτω ουρές σε σχέση με τη θέση τους στην εικόνα. Τα διαμορφωμένα αποτελέσματα φαίνονται στον πίνακα 4.2.

Κλάση	Λέξεις	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	the	92	89	96,73913%
2	and	41	38	92,68293%
3	of	106	106	100%
4	that	19	19	100%
5	than	4	4	100%
6	his	32	32	100%
7	her	13	13	100%
8	with	16	16	100%
9	which	9	9	100%
10	from	5	4	100%
11	it	7	6	85,71429%
12	at	13	13	100%
13	Spain	3	3	100%
	Σύνολο	**360**	**353**	
Ποσοστό Επιτυχίας:			**97.77778%**	

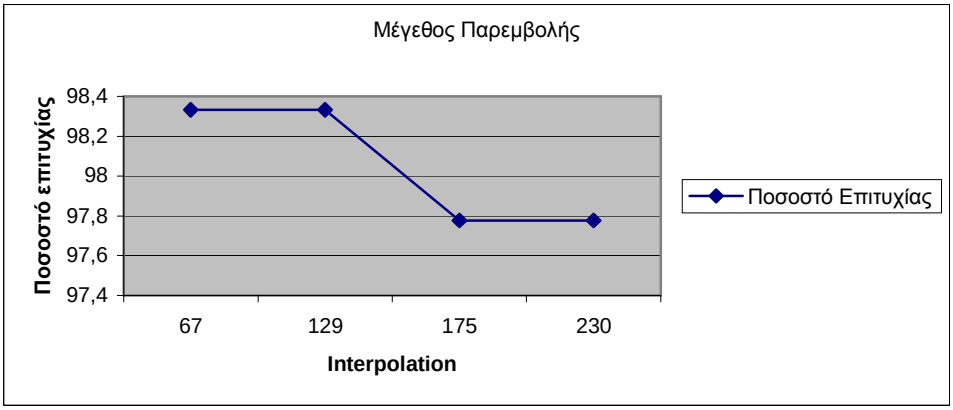
Πίνακας 4.2: Κατηγοριοποίηση με χρήση επιπλέον πάνω και κάτω ουρών

Όπως βλέπουμε δεν υπάρχουν διαφορές στα αποτελέσματα. Αφού ήδη η κατηγοριοποίηση είναι πολύ καλή, χωρίς να ξεχνάμε ότι οι εικόνες-λέξεις δεν έχουν υποστεί ουσιαστικά καμιά προεπεξεργασία.

Στη συνέχεια θα δούμε το ποσοστό επιτυχίας σε σχέση με το μέγεθος της παρεμβολής (πίνακας 4.3 και αντίστοιχο διάγραμμα 4.1), αφαιρώντας από τα χαρακτηριστικά τις πάνω και κάτω ουρές, αφού δεν προσφέρουν ουσιαστική βελτίωση.

	Interpolation 67	Interpolation 129	Interpolation 175	Interpolation 230
Ποσοστό Επιτυχίας	98,33333%	98,33333%	97,77778%	97,77778%

Πίνακας 4.3: Μεταβολή του μεγέθους της παρεμβολής



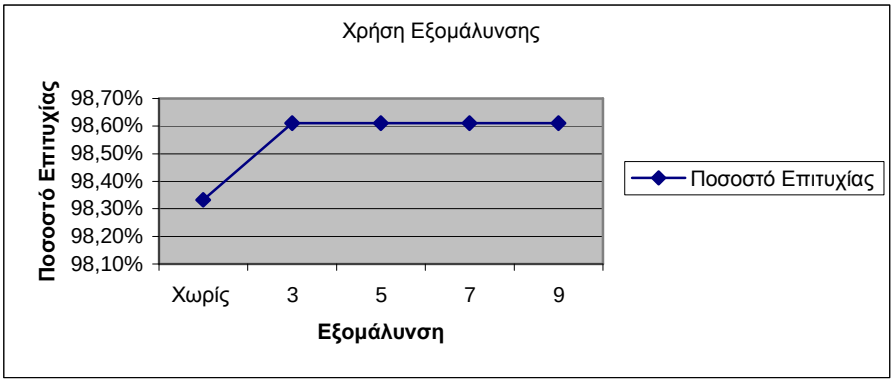
Διάγραμμα 4.1: Μεταβολή του μεγέθους της Παρεμβολής

Παρατηρούμε ότι καλύτερα ποσοστά έχουμε όταν η παρεμβολή έχει μικρό μέγεθος, δηλαδή όταν παρεμβάλλουμε σημεία στα χαρακτηριστικά ώστε να ανήκουν σε λέξεις 2 έως 3 γραμμάτων ή 4 έως 5 γραμμάτων, ενώ πέφτει για μεγαλύτερη παρεμβολή. Βέβαια η διαφορά είναι πολύ μικρή 0,6% και οφείλεται στο γεγονός ότι οι λέξεις που έχουν εισαχθεί στο σύστημα είναι οι περισσότερες 2 έως 4 γραμμάτων.

Τέλος δοκιμάζουμε να εφαρμόσουμε εξομάλυνση στα παραπάνω χαρακτηριστικά. Στον πίνακα 4.4 βλέπουμε τα ποσοστά επιτυχίας σε κάθε περίπτωση εξομάλυνσης για χαμηλή παρεμβολή (67) αφού αυτή είχε καλύτερα αποτελέσματα και το αντίστοιχο διάγραμμα (4.2).

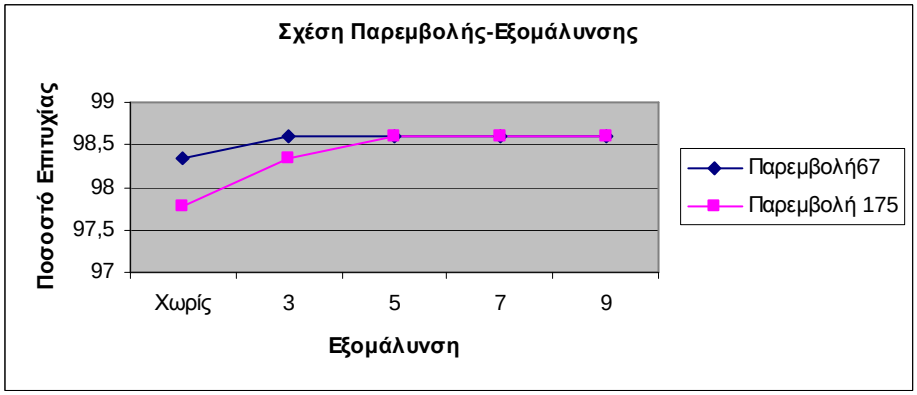
	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	98,33333%	98,61111%	98,61111%	98,61111%	98,61111%

Πίνακας 4.4: Εφαρμογή διαφορετικού βαθμού εξομάλυνσης



Διάγραμμα 4.2: Εφαρμογή διαφορετικού βαθμού εξομάλυνσης

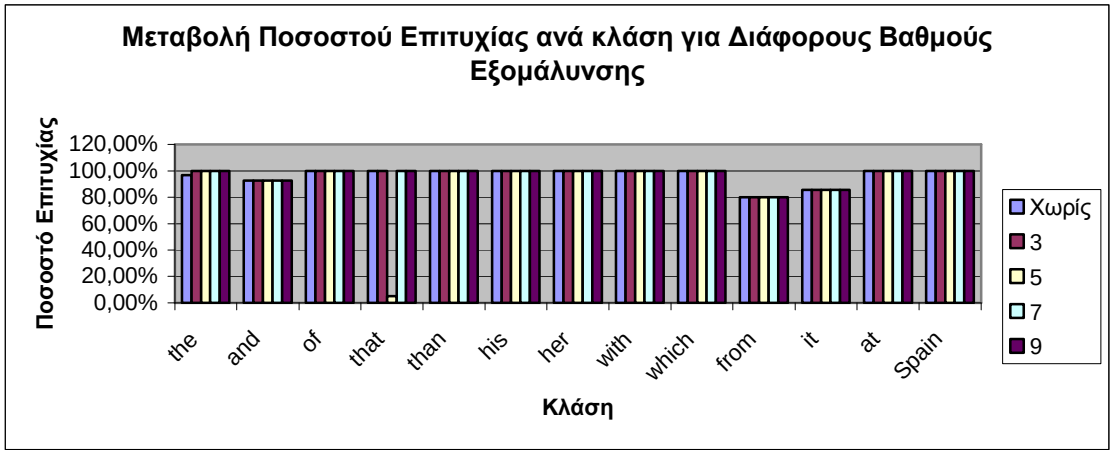
Όπως παρατηρούμε η εξομάλυνση 3 αρκεί για να επιφέρει βελτίωση στο σύστημα 98,61%, ενώ η επιπλέον εξομάλυνση δεν προσφέρει τίποτα παραπάνω. Στο παρακάτω διάγραμμα (4.3) βλέπουμε την σχέση ανάμεσα στην εξομάλυνση και την παρεμβολή.



Διάγραμμα 4.3: Σχέση Παρεμβολής και Εξομάλυνσης

Βλέπουμε ότι για εξομάλυνση 5 τα ποσοστά επιτυχίας για μικρότερη ή μεγαλύτερη παρεμβολή τείνουν στην ίδια τιμή 98,61%.

Σε κάποιες από τις πιο πάνω περιπτώσεις τα ποσοστά επιτυχίας παραμένουν τα ίδια. Δεν είναι όμως τα ίδια και ανά κλάση. Έτσι στην καλύτερη περίπτωση, για παρεμβολή 67, τα ποσοστά επιτυχίας ανά κλάση και διάφορους βαθμούς εξομάλυνσης. φαίνονται στο διάγραμμα 4.4.



Διάγραμμα 4.4: Μεταβολή ποσοστού επιτυχίας ανά κλάση για Διάφορους Βαθμούς Εξομάλυνσης

4.5.2 Εκπαίδευση Τυπωμένο και Έλεγχος σε Ιστορικό

Αντικειμενικός σκοπός του συστήματος είναι η αναγνώριση των εικόνων-λέξεων και η μετατροπή τους σε κείμενο. Στην κατηγορία αυτή των πειραμάτων ουσιαστικά πραγματοποιείται αυτό το πράγμα. Καταφέρνοντας να κατηγοριοποιήσουμε τις εικόνες-λέξεις ιστορικού κειμένου με βάση τυπωμένο κείμενο πετυχαίνουμε τον σκοπό μας.

Στην περίπτωση αυτή χρησιμοποιήθηκε για εκπαίδευση τυπωμένο κείμενο γραμματοσειράς Times New Roman και μεγέθους 12, από μια εμφάνιση για κάθε κατηγορία-λέξη, ενώ για εκπαίδευση χρησιμοποιήθηκαν δύο ομάδες λέξεων: αγγλικής και ελληνικής γλώσσας.

Στη συνέχεια θα δούμε τα αποτελέσματα αναλυτικά για καθεμία περίπτωση.

4.5.2.1 Ιστορικό Αγγλικό

Οι λέξεις προέρχονται, όπως και στην παράγραφο 4.4.1 από το βιβλίο *Travels in Italy, Greece and the Ionian Islands* του *H.W. Williams, Edinburgh 1820.* Αυτή τη φορά χρησιμοποιήθηκαν 108 εμφανίσεις λέξεων (13 κλάσεις) και τα αποτελέσματα της κατηγοριοποίησης με αρχικό μέγεθος παρεμβολής 67 και χαρακτηριστικά κάθετο ιστόγραμμα, πάνω και κάτω προφίλ φαίνονται στον πίνακα 4.5.

Κλάση	Λέξη	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	and	10	10	100%
2	that	10	10	100%
3	the	10	0	0%
4	which	9	9	100%
5	with	10	10	100%
6	at	10	10	100%
7	from	5	4	80%
8	her	10	10	100%
9	his	10	0	0%
10	it	7	6	85,71429%
11	of	10	10	100
12	Spain	3	0	0%
13	than	4	0	0%
	Σύνολο	**108**	**79**	
Ποσοστό Επιτυχίας			**73,14815 %**	

Πίνακας 4.5: Κατηγοριοποίηση με χρήση κάθετου ιστογράμματος, πάνω και κάτω προφίλ

Στη συνέχεια στον πίνακα 4.6 βλέπουμε τα αποτελέσματα της κατηγοριοποίησης αν επιπλέον βάλουμε στα χαρακτηριστικά πάνω και κάτω ουρές.

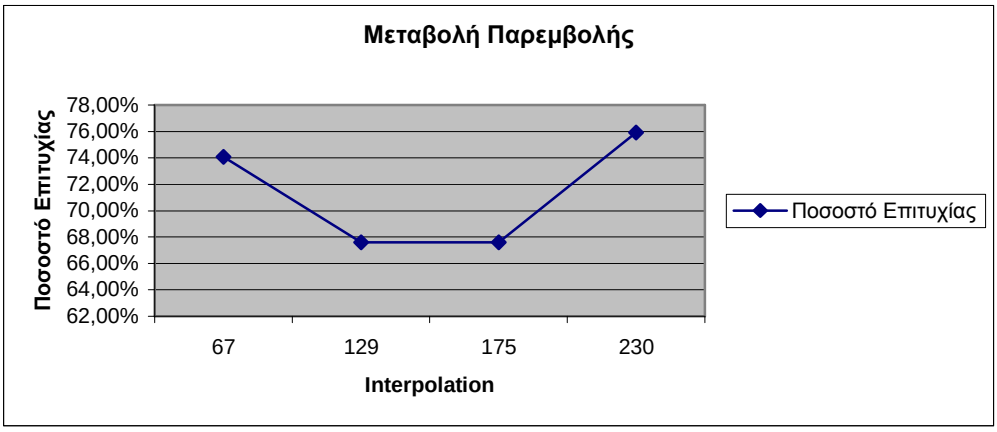
Κλάση	Λέξη	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	and	10	10	100%
2	that	10	10	100%
3	the	10	0	0%
4	which	9	9	100%
5	with	10	10	100%
6	at	10	10	100%
7	from	5	5	100%
8	her	10	10	100%
9	his	10	0	0%
10	it	7	6	85,71429%
11	of	10	10	100%
12	Spain	3	0	0%
13	than	4	0	0%
	Σύνολο	**108**	**80**	
Ποσοστό Επιτυχίας			**74,07407%**	

Πίνακας 4.6: Κατηγοριοποίηση με χρήση επιπλέον πάνω και κάτω ουρών

Τα αποτελέσματα βελτιώνονται λίγο παραπάνω με την χρήση των ουρών. Στη συνέχεια θα δούμε τα ποσοστά επιτυχίας της κατηγοριοποίησης σε σχέση με την μεταβολή του μεγέθους της παρεμβολής (πίνακας 4.7 και αντίστοιχο διάγραμμα 4.5), κρατώντας τα χαρακτηριστικά των πάνω και κάτω ουρών, αφού προσφέρουν βελτίωση στην απόδοση του συστήματος μας.

	Interpolation 67	Interpolation 129	Interpolation 175	Interpolation 230
Ποσοστό Επιτυχίας	74,07407%	67,59259%	67,59259%	75,92593%

Πίνακας 4.7: Μεταβολή του μεγέθους της παρεμβολής



Διάγραμμα 4.5: Μεταβολή του μεγέθους της παρεμβολής

Παρατηρούμε ότι για μικρές τιμές της παρεμβολής το ποσοστό επιτυχίας είναι αρκετά καλό, για μεσαίες τιμές πέφτει και κορυφώνεται για μεγαλύτερες τιμές (230) και φτάνει στο 75,92%. Βεβαία η διαφορά είναι μικρή και αυτό μπορεί να είναι τυχαίο.

Τέλος εφαρμόζουμε εξομάλυνση στα παραπάνω χαρακτηριστικά. Στον πίνακα 4.8α και 4.8β βλέπουμε τα ποσοστά επιτυχίας σε κάθε περίπτωση εξομάλυνσης για χαμηλή παρεμβολή (67) και υψηλή παρεμβολή (230) αντίστοιχα, αφού αυτές είχαν τα καλύτερα αποτελέσματα και τα αντίστοιχα διαγράμματα (4.6α και 4.6β).

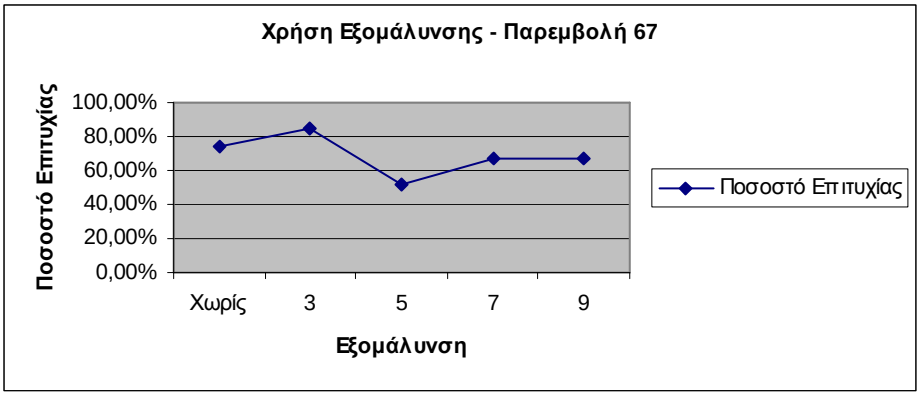
	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	74,07407%	85,18519%	51,85185%	66,66667%	66,66667%

(α): Παρεμβολή 67

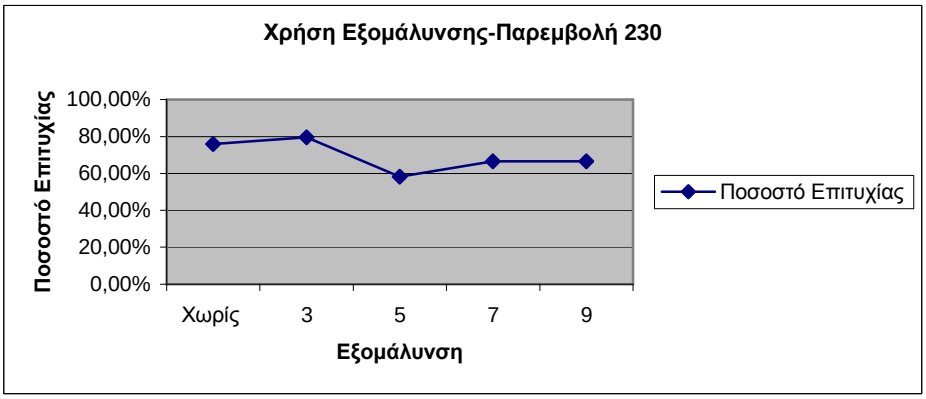
	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	75,92593%	79,62963%	58,33333%	66,66667%	66,66667%

(α): Παρεμβολή 230

Πίνακας 4.8: Εφαρμογή διαφορετικού βαθμού εξομάλυνσης



(α)

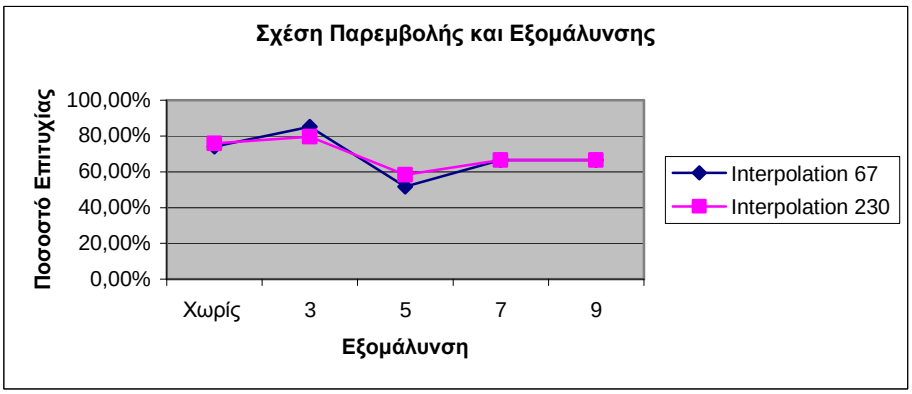


(β)

Διάγραμμα 4.6: Εφαρμογή Διαφορετικού Βαθμού Εξομάλυνσης

Τελικά τα καλύτερα αποτελέσματα τα έχουμε για μικρή παρεμβολή (67) και εξομάλυνση 3 και φτάνουμε στο 85,18% επιτυχία, ποσοστό που θεωρείται πολύ καλό.

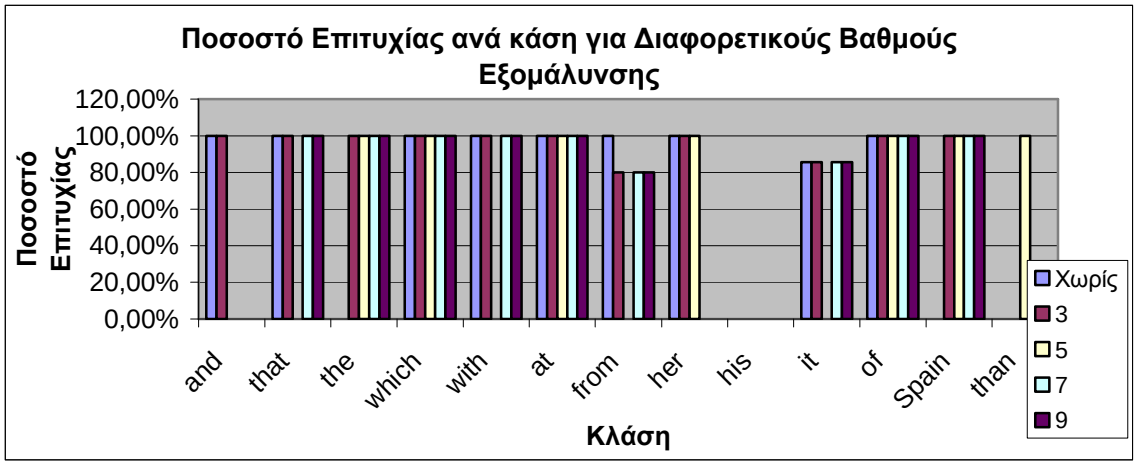
Στο διάγραμμα 4.7 βλέπουμε την σχέση εξομάλυνσης και παρεμβολής.



Διάγραμμα 4.7: Σχέση Παρεμβολής και Εξομάλυνσης

Ενώ κατά κύριο λόγο η παρεμβολή 230 έχει καλύτερα αποτελέσματα για εξομάλυνση 3 η παρεμβολή 67 την ξεπερνάει αρκετά κοντά 6%, ποσοστό που δεν είναι τυχαίο.

Όπως είπαμε και πριν οι κάποια ποσοστά επιτυχίας μπορεί να είναι ίδια, κάποιες όμως από τις κλάσεις όμως κατηγοριοποιούνται με μικρές διαφοροποιήσεις. Έτσι στο διάγραμμα 4.8 βλέπουμε αυτές τις διαφοροποιήσεις για μέγεθος παρεμβολής 67 και διάφορους βαθμούς εξομάλυνσης.



Διάγραμμα 4.8: Ποσοστό επιτυχίας ανά κλάση για Διαφορετικούς Βαθμούς Εξομάλυνσης

Βλέπουμε εδώ για παράδειγμα την λέξη *and* η οποία έχει διαφορετικά ποσοστά επιτυχίας, στην περίπτωση της μη ύπαρξης εξομάλυνσης είναι 100%, το ίδιο και για εξομάλυνση 3, ενώ για τους υπόλοιπους βαθμούς εξομάλυνσης είναι 0%.

4.5.2.2 Ιστορικό Ελληνικό

Στα πειράματα αυτά χρησιμοποιήθηκε πάλι ιστορικό κείμενο αλλά από την ελληνική βιβλιογραφία και συγκεκριμένα οι εικόνες-λέξεις προέρχονται από δύο βιβλία τα *Απομνημονεύματα επί της σύγχρονης ιστορίας, Σπυρίδωνος Μαλάκη, Εν Αθήναις 1895* και *Σκιαγραφία Αργοστολίου του 1821, Σάββα Αννίνος*.

Χρησιμοποιήθηκαν 46 εμφανίσεις λέξεων που κατηγοριοποιήθηκαν σε 21 κλάσεις. Εδώ πρέπει να τονιστεί ότι οι λέξεις του ιστορικού κειμένου είναι γραμμένες στο πολυτονικό σύστημα, ενώ οι λέξεις εκπαίδευσης (τυπωμένο Times New Roman 12) είναι στο μονοτονικό σύστημα, γεγονός που δυσκολεύει την αναγνώριση ακόμα περισσότερο.

Επίσης στα πειράματα που ακολουθούν ως χαρακτηριστικά των εικόνων λέξεων χρησιμοποιήσαμε και τα πέντε (το κάθετο ιστόγραμμα, το πάνω και κάτω προφίλ και τις πάνω και κάτω ουρές), αφού όπως είδαμε και στην προηγούμενη παράγραφο, τα αποτελέσματα είναι καλύτερα απ' ότι αν χρησιμοποιούσαμε μόνο τα τρία πρώτα.

Στον πίνακα 4.9 βλέπουμε μια πρώτη κατηγοριοποίηση των λέξεων με παρεμβολή 175.

Κλάση	Λέξεις	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	και	11	11	100%
2	ελθών	1	0	0%
3	οικαδε	1	1	100%
4	με	1	0	0%
5	κατεθλιμένην	1	0	0%

6	ψυχήν	1	0	0%
7	έγγραψα	1	0	0%
8	εις	2	2	100%
9	της	6	0	0%
10	την	3	0	0%
11	των	2	0	0%
12	προς	2	0	0%
13	δια	4	4	100%
14	υπό	2	2	100%
15	βαθμίδα	1	1	100%
16	ανελθών	1	1	100%
17	που	2	0	0%
18	επίθεσιν	1	1	100%
19	θυγατρός	1	0	0%
20	προσοχήν	1	0	0%
21	πάλιν	1	0	0%
	Σύνολα	**46**	**23**	
Ποσοστό επιτυχίας			**50%**	

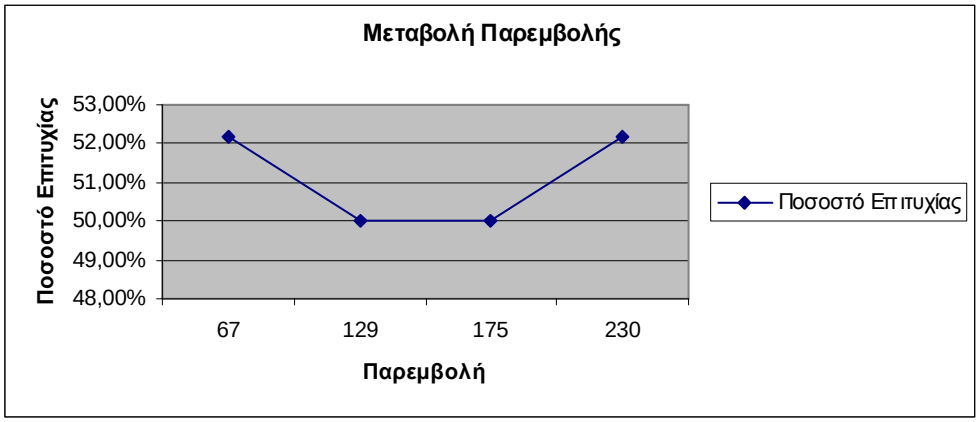
Πίνακας 4.9: Κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών

Το ποσοστό επιτυχίας είναι αρκετά μικρότερο σε σχέση με το ποσοστό της προηγούμενης παραγράφου στο αγγλικό κείμενο. Αυτό οφείλεται σε μεγάλο βαθμό στο πολυτονικό σύστημα των δεδομένων ελέγχου.

Στη συνέχεια θα δούμε τα αποτελέσματα της κατηγοριοποίησης για διάφορες τιμές παρεμβολής (πίνακας 4.10 και διάγραμμα 4.9)

	Interpolation 67	Interpolation 129	Interpolation 175	Interpolation 230
Ποσοστό Επιτυχίας	52,17391%	50%	50%	52,17391%

Πίνακας 4.10: Μεταβολή του μεγέθους της παρεμβολής



Διάγραμμα 4.9: Μεταβολή του μεγέθους της παρεμβολής

Όπως παρατηρούμε τα ποσοστά είναι μεγαλύτερη για πολύ μικρή ή πολύ μεγάλη παρεμβολή κι αυτό γιατί ο μεγαλύτερος αριθμός λέξεων είναι λίγων ή πολλών γραμμάτων και όχι ενός ενδιάμεσου αριθμού (4 έως 6 γραμμάτων).

Στη συνέχεια στον πίνακα 4.11 εφαρμόζουμε εξομάλυνση διαφορών σημείων στην προσπάθειά μας να βελτιώσουμε τα αποτελέσματα, για τα δυο μεγέθη παρεμβολής (α και β). Επίσης στο διάγραμμα 4.10 βλέπουμε την μεταβολή του ποσοστού επιτυχίας σε σχέση με την εξομάλυνση και την παρεμβολή.

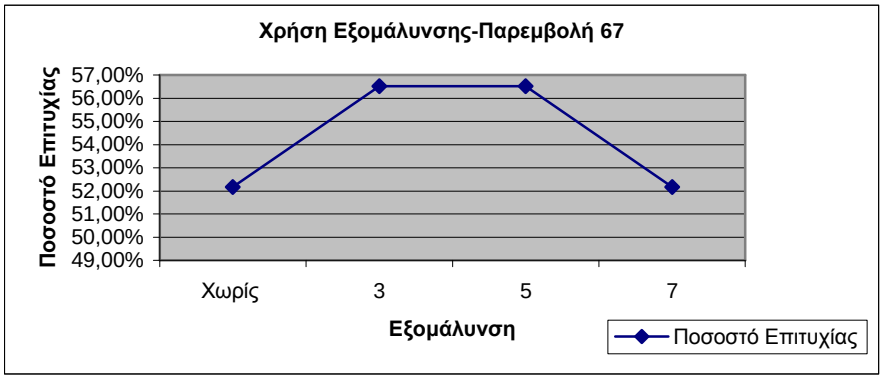
	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	52,17391%	56,52174%	56,52174%	52,17391%	54,34783%

(α): Παρεμβολή 67

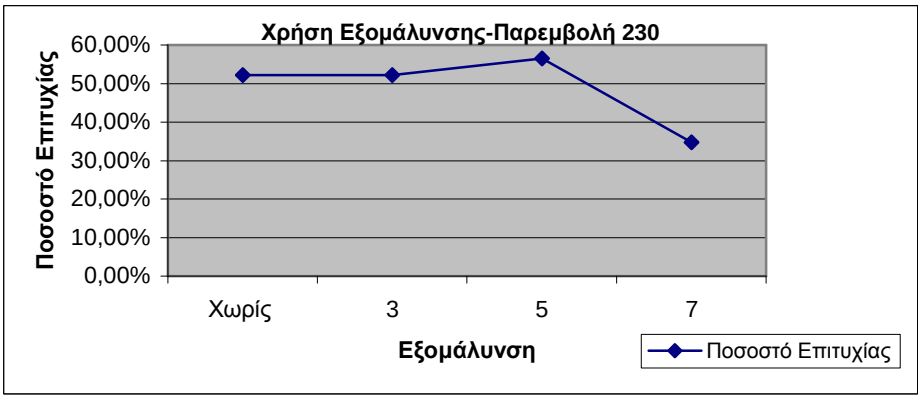
	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	52,17391%	52,17391%	56,52173%	34,7826087%	52,17391%

(α): Παρεμβολή 230

Πίνακας 4.11: Εφαρμογή διαφορετικού βαθμού εξομάλυνσης



(α)

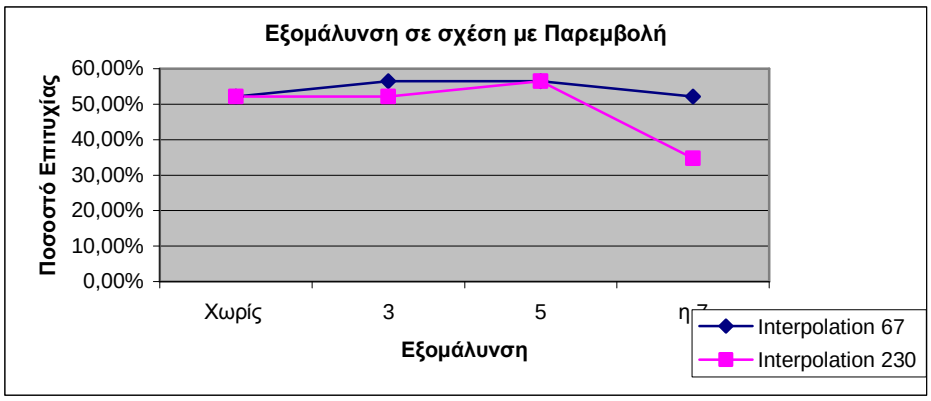


(β)

Διάγραμμα 4.10: Εφαρμογή διαφορετικού βαθμού εξομάλυνσης

Παρατηρούμε ότι καλύτερα αποτελέσματα έχουμε ή για μικρή παρεμβολή και εξομάλυνση 3 ή 5 ή για μεγάλη παρεμβολή και εξομάλυνση 5 και το ποσοστό επιτυχίας φτάνει το 56,52%.

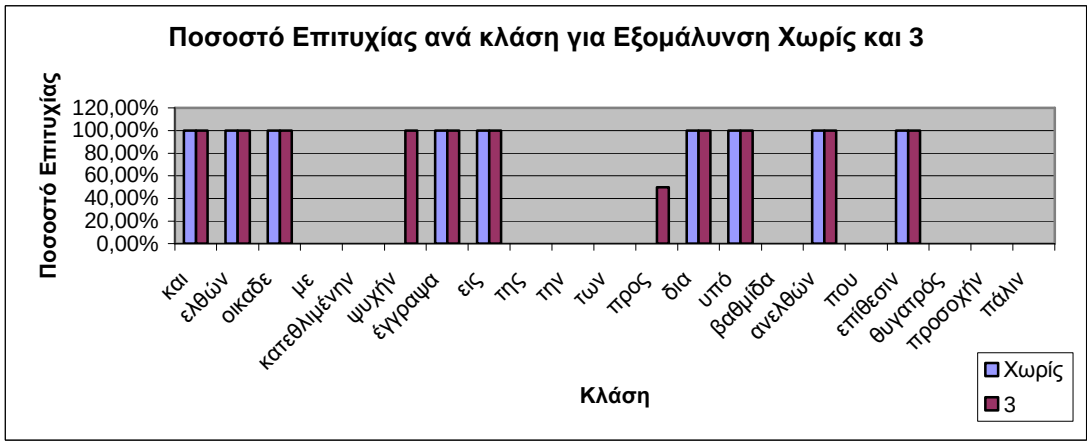
Στο διάγραμμα 4.11 βλέπουμε την εξομάλυνση σε σχέση με το μέγεθος της παρεμβολής.



Διάγραμμα 4.11: Σχέση Εξομάλυνσης με Παρεμβολή

Βλέπουμε ότι στο μεγαλύτερο μέρος του διαγράμματος καλύτερα αποτελέσματα έχουμε όταν η παρεμβολή είναι μικρή. Οι τιμές συμπίπτουν για εξομάλυνση 5, που σημαίνει ότι η περίπτωση αυτή είναι ανεξάρτητη της παρεμβολής.

Και σε αυτό το σημείο υπάρχουν διαφορετικές διαφοροποιήσεις της κατηγοριοποίησης των κλάσεων και στο διάγραμμα 4.12, τις βλέπουμε για κάθε κλάση και για βαθμούς εξομάλυνσης χωρίς και 3 και παρεμβολή 67.



Διάγραμμα 4.12: Ποσοστό Επιτυχίας ανά κλάση για Εξομάλυνση Χωρίς και 3

Τα αποτελέσματα όμως συνεχίζουν να μην είναι πολύ ψηλά σε σχέση με αυτά της παραγράφου 4.4.2.1. Αυτό είπαμε ότι οφείλεται στο πολυτονικό σύστημα των λέξεων ελέγχου. Στη συνέχεια δοκιμάζουμε να εξομαλύνουμε τα δεδομένα αυτά και όχι τα δεδομένα εκπαίδευσης ή δοκιμάζουμε και διαφορετικούς βαθμούς εξομάλυνσης για κάθε σύνολο λέξεων, μεγαλύτερο κάθε φορά για το δεύτερο σύνολο, προσπαθώντας έτσι να επηρεάζει λιγότερο η χρήση του πολυτονικού συστήματος.

Τα αποτελέσματα φαίνονται στον πίνακα 4.12, για μέγεθος παρεμβολής 67.

	Δεδομένα Εκπαίδευσης	**Δεδομένα Ελέγχου**	**Ποσοστό Επιτυχίας**
Εξομάλυνση	-	*3*	54,347826%
	-	*5*	54,347826%
	-	*7*	54,347826%
	-	*9*	54,347826%
	3	*5*	56,521739%
	3	*7*	56,521739%
	3	*9*	56,521739%
	5	*7*	54,347826%
	5	*9*	54,347826%

Πίνακας 4.12: Χρήση διαφορετικού βαθμού εξομάλυνση για training και testing δεδομένα Interpolation=67

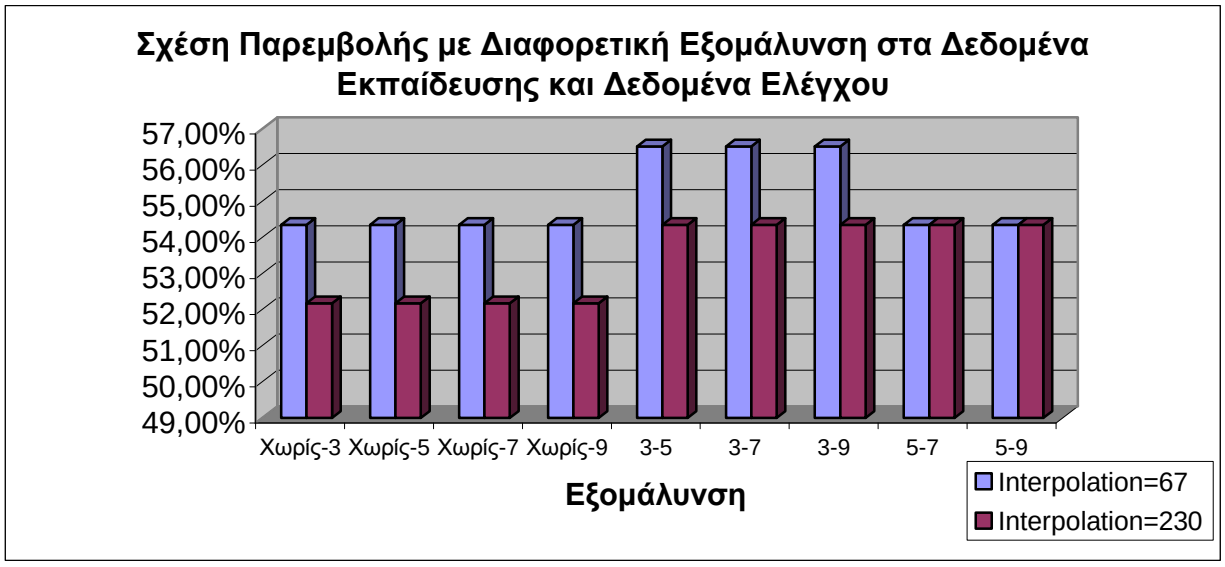
Όπως φαίνεται το ποσοστό επιτυχίας φτάνει μέχρι το 56,52%, το ίδιο δηλαδή που είχαμε για εξομάλυνση 3 και των δεδομένων εκπαίδευσης και των δεδομένων ελέγχου.

Δοκιμάζουμε το ίδιο πράγμα και για παρεμβολή 230, αφού και για αυτήν είχαμε παρόμοια ποσοστά επιτυχίας και παίρνουμε τον πίνακα 4.13.

	Δεδομένα Εκπαίδευσης	**Δεδομένα Ελέγχου**	**Ποσοστό Επιτυχίας**
Εξομάλυνση	-	*3*	52,173913%
	-	*5*	52,173913%
	-	*7*	52,173913%
	-	*9*	52,173913%
	3	*5*	54,347826%
	3	*7*	54,347826%
	3	*9*	54,347826%
	5	*7*	54,347826%
	5	*9*	54,347826%

Πίνακας 4.13: Χρήση διαφορετικού βαθμού εξομάλυνση για training και testing δεδομένα Interpolation=230

Εδώ παρατηρούμε ότι το ποσοστό επιτυχίας δεν αυξάνεται, αλλά μάλιστα μένει το ίδιο για ίδιο βαθμό εξομάλυνσης στο σύνολο των δεδομένων εκπαίδευσης. Στο διάγραμμα 4.13 βλέπουμε τις διακυμάνσεις του ποσοστού επιτυχίας σε σχέση με την παρεμβολή και τον βαθμό εξομάλυνσης.



Διάγραμμα 4.13: Σχέση Παρεμβολής με Διαφορετική Εξομάλυνση στα Δεδομένα Εκπαίδευσης και Δεδομένα Ελέγχου

Τελικά, όπως βλέπουμε στο διάγραμμα τα καλύτερα ποσοστά τα επιτυγχάνουμε για μέγεθος παρεμβολής 67.

4.5.3 Εκπαίδευση Τυπωμένο και Έλεγχος σε Χειρόγραφο Διαφορετικών Συγγραφέων

Η αναγνώριση χειρόγραφου κειμένου είναι πάρα πολύ δύσκολη. Ειδικά στην περίπτωση όπου το κείμενο προέρχεται όχι από έναν αλλά από 15 διαφορετικούς συγγραφείς. Σε κάποιες περιπτώσεις ο ίδιος ο άνθρωπος έχει αντικειμενικές δυσκολίες να το κάνει αυτό με 100% επιτυχία.

Εμείς χρησιμοποιώντας το σύστημα που περιγράψαμε, θα προσπαθήσουμε να εξάγουμε τα χαρακτηριστικά από 225 εικόνες-λέξεις χειρόγραφου κειμένου, που έχουμε πάρει από τη βάση δεδομένων GCDB, που κατασκεύασαν οι Μαργαρώνης Ιωάννης και Χρήστου Μηνάς, για το Πανεπιστήμιο Αιγαίου-Τμήμα Μηχανικών Η/Υ και Πληροφοριακών και Επικοινωνιακών Συστημάτων και στη συνέχεια χρησιμοποιώντας για δεδομένα εκπαίδευσης τυπωμένο κείμενο γραμματοσειράς Times New Roman, μεγέθους 12, να τις κατηγοριοποιήσουμε σε 15 κλάσεις. Ουσιαστικά έχουμε 15 εμφανίσεις κάθε λέξης, ενώ για τις λέξεις του συνόλου εκπαίδευσης έχουμε μόνο από μία εμφάνιση.

Όπως αναφέρθηκε και στην εισαγωγή αυτού του κεφαλαίου, οι χειρόγραφες λέξεις δεν έχουν υποστεί καμιά μεταβολή και πρέπει να λάβουμε υπόψη μας την κλίση των χαρακτήρων, που συναντάμε σε πολλούς γραφικούς χαρακτήρες ή το συνεχόμενο γράψιμο, χαρακτηριστικά που δεν βλέπουμε στο τυπωμένο κείμενο.

Τα χαρακτηριστικά που εξάγουμε είναι το κάθετο ιστόγραμμα, το πάνω και κάτω προφίλ, οι πάνω και κάτω ουρές. Στον πίνακα 4.14 βλέπουμε τα αποτελέσματα μιας πρώτης κατηγοριοποίησης χρησιμοποιώντας μέγεθος παρεμβολής 175.

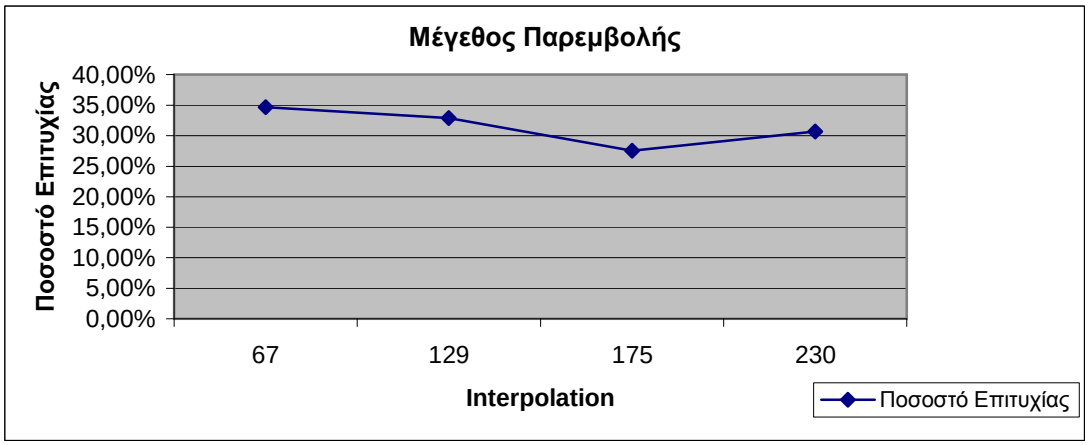
Κλάση	Λέξεις	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	βρέχω	15	6	40%
2	δημοκρατία	15	0	0%
3	δορυφόρος	15	4	26,66667%
4	εξέλιξη	15	0	0%
5	φαίνεται	15	3	20%
6	για	15	7	46,66667%
7	γυναίκα	15	8	53,33333%
8	και	15	11	73,33333%
9	καθώς	15	2	13,33333%
10	μάλλον	15	2	13,33333%
11	όπως	15	4	26,66667%
12	ψάχνω	15	7	46,66667%
13	σαστίζω	15	2	13,33333%
14	σώμα	15	5	33,33333%
15	ζημιά	15	1	6,666667%
	Σύνολο	**225**	**62**	
Ποσοστό Επιτυχίας			**27,55556%**	

Πίνακας 4.14: Κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών

Το ποσοστό επιτυχίας είναι αρκετά μικρό 27,55% και θα προσπαθήσουμε να το βελτιώσουμε αρχικά με μεταβολή του μεγέθους της παρεμβολής. Έτσι στον πίνακα 4.15 βλέπουμε την μεταβολή του ποσοστού επιτυχίας με την μεταβολή του μεγέθους της παρεμβολής και τα ίδια αποτελέσματα φαίνονται γραφικά στο διάγραμμα 4.14.

	Interpolation 67	Interpolation 129	Interpolation 175	Interpolation 230
Ποσοστό Επιτυχίας	34,66667%	32,88889%	27,55556%	30,66667%

Πίνακας 4.15: Μεταβολή του μεγέθους της παρεμβολής



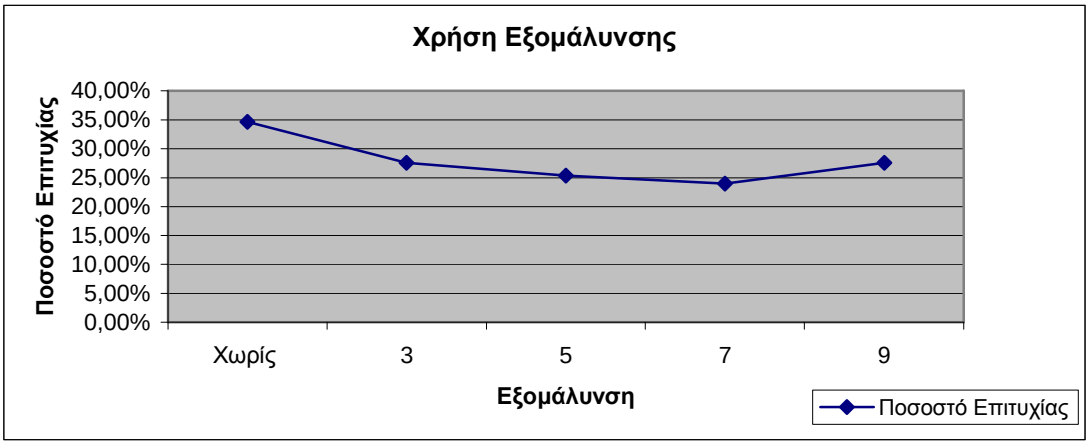
Διάγραμμα 4.14: Μεταβολή του μεγέθους της παρεμβολής

Βλέπουμε ότι καλύτερα αποτελέσματα έχουμε για μικρή παρεμβολή (μέγεθος 67) και φτάνει στο 34,66%. Όσο μεγαλώνει η παρεμβολή το ποσοστό επιτυχίας πέφτει, για να κάνει μια μικρή άνοδο για μέγεθος 230.

Στη συνέχεια εφαρμόζουμε διάφορους βαθμούς εξομάλυνσης και τα αποτελέσματα φαίνονται στον πίνακα 4.16 και διάγραμμα 4.15 με παρεμβολή 67, αφού για αυτήν είχαμε μακράν τα καλύτερα αποτελέσματα.

	Χωρίς Εξομάλυνση	Εξομάλυνση 3	Εξομάλυνση 5	Εξομάλυνση 7	Εξομάλυνση 9
Ποσοστό Επιτυχίας	34,66667%	27,55555%	25,33333%	24%	27,55556%

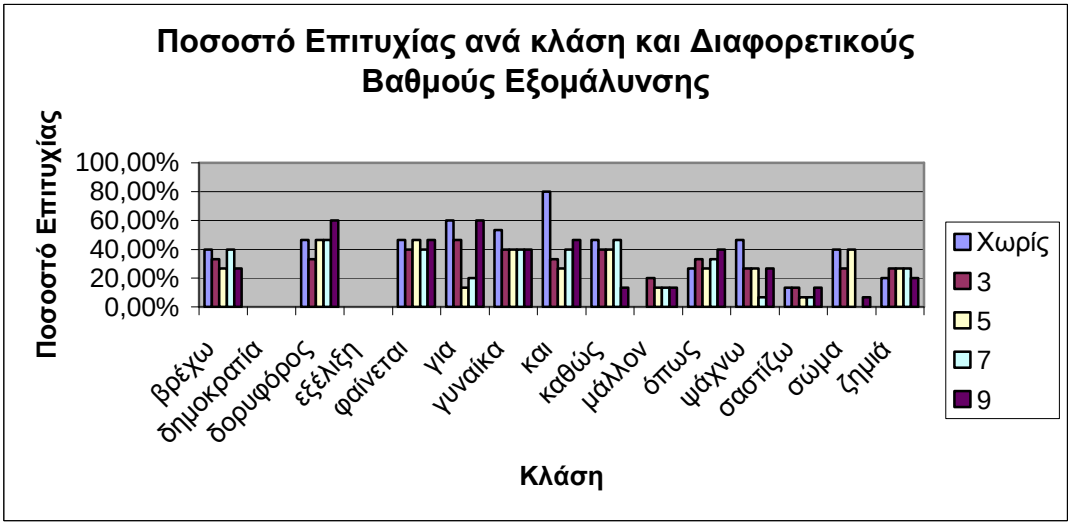
Πίνακας 4.16: Χρήση διαφορετικού βαθμού εξομάλυνσης



Διάγραμμα 4.15: Χρήση διαφορετικού βαθμού εξομάλυνσης

Βλέπουμε ότι καλύτερα αποτελέσματα έχουμε χωρίς εξομάλυνση, και αυτό συμβαίνει γιατί με την εξομάλυνση χάνεται σημαντικό ποσοστό της περιγραφής του σχήματος της λέξης όταν εφαρμόζεται και στα δυο σύνολα δεδομένων (λέξεις εκπαίδευσης και ελέγχου).

Στο διάγραμμα 4.16 βλέπουμε τις μεταβολές του ποσοστού επιτυχίας ανά κλάση για διάφορους βαθμούς εξομάλυνσης και παρεμβολή 67.

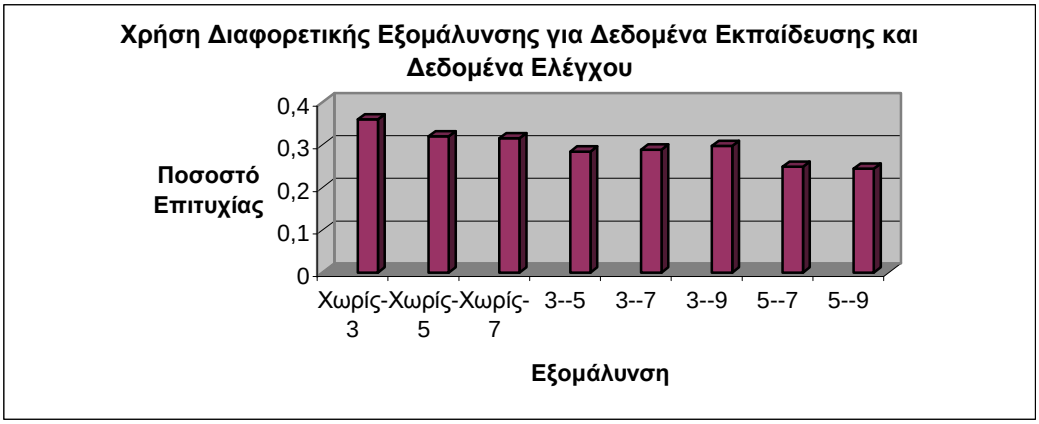


Διάγραμμα 4.16: Ποσοστό Επιτυχίας ανά κλάση για διαφορετικούς βαθμούς εξομάλυνσης

Στη συνέχεια θα εφαρμόσουμε παρεμβολή διαφορετικού βαθμού εξομάλυνση για τα δεδομένα εκπαίδευσης και ελέγχου. Τα αποτελέσματα φαίνονται στον πίνακα 4.17 και στο διάγραμμα 4.17.

	Δεδομένα Εκπαίδευσης	Δεδομένα Ελέγχου	Ποσοστό Επιτυχίας
Εξομάλυνση	-	*3*	36%
	-	*5*	32%
	-	*7*	31,55556%
	3	*5*	28,44444%
	3	*7*	28,88889%
	3	*9*	29,77778%
	5	*7*	24,88889%
	5	*9*	24,44444%

Πίνακας 4.17: Χρήση διαφορετικού βαθμού εξομάλυνσης στα δεδομένα εκπαίδευσης και εισόδου



Διάγραμμα 4.17: Χρήση διαφορετικού βαθμού εξομάλυνσης στα δεδομένα εκπαίδευσης και εισόδου

Παρατηρούμε ότι τα καλύτερα αποτελέσματα για την κατηγορία αυτή των πειραμάτων επιτυγχάνονται για δεδομένα εκπαίδευσης χωρίς εξομάλυνση και δεδομένα ελέγχου με εξομάλυνση 3 και είναι 36%.

Βέβαια όπως αναφέρθηκε και προηγουμένως οι εικόνες λέξεις προέρχονται από 15 διαφορετικούς συγγραφείς. Τα αποτελέσματα είναι καλύτερα για λιγότερους. Έτσι αν βάλουμε στο σύστημα από τρεις εμφανίσεις των λέξεων (3 διαφορετικούς συγγραφείς) θα έχουμε την παρακάτω κατηγοριοποίηση, για παρεμβολή 67 (πίνακας 4.18).

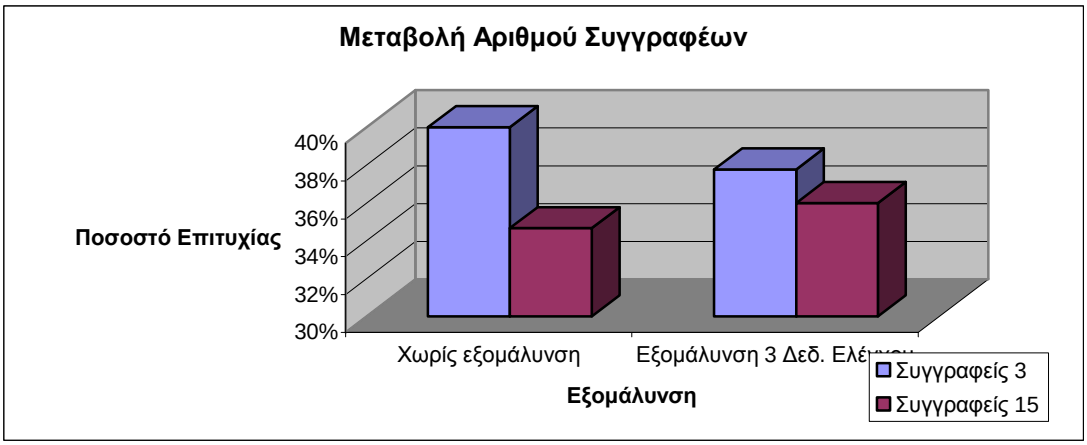
Κλάση	Λέξεις	Εμφανίσεις Λέξεων	Σωστή Κατηγοριοποίηση	Ποσοστό Επιτυχίας ανά κλάση
1	βρέχω	3	1	33,33333%
2	δημοκρατία	3	0	0%
3	δορυφόρος	3	2	66,66667%
4	εξέλιξη	3	0	0%

5	φαίνεται	3	2	66,66667%
6	για	3	1	33,33333%
7	γυναίκα	3	1	33,33333%
8	και	3	3	100%
9	καθώς	3	2	66,66667%
10	μάλλον	3	0	0%
11	όπως	3	2	66,66667%
12	ψάχνω	3	2	66,66667%
13	σαστίζω	3	0	0%
14	σώμα	3	1	33,33333%
15	ζημιά	3	1	33,33333%
	Σύνολο	**45**	**18**	
Ποσοστό Επιτυχίας			40%	

Πίνακας 4.18: Κατηγοριοποίηση για 3 συγγραφείς

Βλέπουμε ότι τα αποτελέσματα φτάνουν το 40%, ενώ σε πειράματα που έγιναν όσον αφορά το μέγεθος της παρεμβολής ή την εξομάλυνση δεν είχαμε επιπλέον βελτίωση. Συγκεκριμένα για δεδομένα εκπαίδευσης χωρίς εξομάλυνση και δεδομένα ελέγχου με εξομάλυνση 3, που στην προηγούμενη περίπτωση είχαμε ποσοστό επιτυχίας 37,77%.

Στο παρακάτω διάγραμμα (4.18) βλέπουμε την σχέση ανάμεσα στον αριθμό των συγγραφέων και το ποσοστό επιτυχίας για δεδομένα χωρίς εξομάλυνση (α) και για δεδομένα εκπαίδευσης χωρίς εξομάλυνση και δεδομένα ελέγχου με εξομάλυνση 3 (β).



Διάγραμμα 4.18: Μεταβολή του Αριθμού Συγγραφέων

Βλέπουμε ότι, όπως είναι φυσικό, για λιγότερους συγγραφείς έχουμε καλύτερα αποτελέσματα, αλλά ενώ στους λίγους συγγραφείς η εξομάλυνση δεν βελτιώνει τα αποτελέσματα, στους πολλούς έχουμε βελτίωση. Αυτό σημαίνει ότι γενικά η εξομάλυνση βοηθάει.

Κεφάλαιο 5

Συμπεράσματα

Στα προηγούμενα κεφάλαια παρουσιάσαμε ένα υποσύστημα αναγνώρισης ολόκληρης λέξης. Το υποσύστημά μας ενδείκνυται για περιπτώσεις μετατροπής ψηφιακών ιστορικών κειμένων με την μορφή εικόνας, σε κείμενο με τη χρήση λεξιλογίου, ώστε να είναι άμεσα επεξεργάσιμες, όπως επίσης και για περιπτώσεις αναζήτησης σε τέτοια έγγραφα. Το υποσύστημα αυτό αποτελείται από ένα στάδιο καθαρισμού των εικόνων-λέξεων, την εξαγωγή κάποιων χαρακτηριστικών των λέξεων αυτών, την κανονικοποίησή τους και τέλος την κατηγοριοποίηση τους σε κλάσεις.

Συγκεκριμένα είδαμε ότι οι εικόνες-λέξεις δεν υπέστησαν κάποια βασική προεπεξεργασία, εκτός από την απαλοιφή των κενών γραμμών πάνω και κάτω από την λέξη.

Δείξαμε ακόμα ποια είναι τα χαρακτηριστικά των λέξεων που επιλέχτηκαν: το κάθετο ιστόγραμμα, για να έχουμε μια ένδειξη του μελανιού που υπάρχει σε κάθε στήλη της εικόνας, το πάνω και κάτω προφίλ, που ουσιαστικά είναι η περιγραφή του σχήματος της λέξης και ο εντοπισμός των πάνω και κάτω ουρών από το κυρίως σώμα της λέξης για να έχουμε επιπλέον βελτίωση.

Επίσης δείξαμε την αναγκαιότητα της κανονικοποίησης των τιμών των εξαγόμενων χαρακτηριστικών τόσο κατά ύψος όσο και κατά μήκος, ώστε να μπορούμε να τις συγκρίνουμε για τις διαφορετικές εικόνες-λέξεις. Παρουσιάστηκε για το σκοπό αυτό η μέθοδος της παρεμβολής (για την κανονικοποίηση κατά μήκος), η οποία αποτελεί βασική διαφοροποίηση σε σχέση με την υπάρχουσα βιβλιογραφία σε αυτόν τον τομέα.

Παρουσιάστηκε ακόμα, η ανάγκη εξομάλυνσης κάποιων χαρακτηριστικών, ώστε να ταιριάζουν σε περισσότερες διαφορετικές εμφανίσεις των ίδιων εικόνων-λέξεων.

Τέλος χρησιμοποιήθηκε ο αλγόριθμος k-Means για την κατηγοριοποίηση των λέξεων σε κλάσεις, με βάση τα προηγούμενα χαρακτηριστικά τους. Όπως είδαμε η μέθοδος αυτή είναι αρκετά γρήγορη, χωρίς πολύπλοκους υπολογισμούς, όπως σε άλλες περιπτώσεις στην βιβλιογραφία που γίνεται η χρήση νευρωνικών δικτύων, dynamic time warping (DTW) [24], αλυσίδων Markov [26] [10] κτλ.

Τα πειράματα που περιγράφηκαν στο κεφάλαιο 4 με το προτεινόμενο υποσύστημα πραγματοποιήθηκαν στις εξής ομάδες εικόνων λέξεων:

- Δεδομένα εκπαίδευσης και δεδομένα ελέγχου από ιστορικό *αγγλικό κείμενο*
- Δεδομένα εκπαίδευσης τυπωμένο κείμενο και δεδομένα ελέγχου ιστορικό:
  - ✓ *Αγγλικού κειμένου*
  - ✓ *Ελληνικού κειμένου*
- Δεδομένα εκπαίδευσης τυπωμένο κείμενο και δεδομένα ελέγχου χειρόγραφο από τη βάση GCDB.

Σε κάθε περίπτωση οι δυσκολίες ήταν αρκετές. Το γεγονός ότι οι λέξεις δεν είχαν υποστεί καμιά ουσιαστική προεπεξεργασία, όπως διόρθωση γωνίας εκτροπής, διόρθωση κλίσης χαρακτήρων κτλ δυσκόλευε την αναγνώριση ακόμα παραπάνω.

Για παράδειγμα, στα πειράματα με λέξεις από ελληνικά ιστορικά κείμενα είδαμε ότι η χρήση του πολυτονικού συστήματος που ίσχυε την εποχή εκείνη, κάνει την αναγνώριση αυτών των λέξεων ακόμα πιο δύσκολη. Και αυτό συμβαίνει γιατί προσπαθούμε να τις κατηγοριοποιήσουμε με βάση τυπωμένο κείμενο του μονοτονικού συστήματος.

Τέλος το χειρόγραφο κείμενο είναι μια από τις δυσκολότερες περιπτώσεις στον τομέα της οπτικής αναγνώρισης χαρακτήρων. Οι χαρακτήρες μπορεί να έχουν περισσότερη κλίση από αντίστοιχο τυπωμένο κείμενο, μπορεί να είναι συνεχόμενοι χωρίς κενά, και γενικά να προκαλούν δυσκολία στην αναγνώριση ακόμα και από τον ίδιο τον άνθρωπο.

Σε κάθε περίπτωση πάντως γίνεται μεταβολή διαφόρων χαρακτηριστικών, όπως το μέγεθος της παρεμβολής που θα χρησιμοποιηθεί και ο βαθμός της εξομάλυνσης που θα εφαρμοστεί. Όπως είδαμε χρησιμοποιήσαμε τέσσερα μεγέθη παρεμβολής: 67, 129, 175 και 230, που αντιστοιχούν σε λέξεις 2 έως 3, 4 έως 5, 6 έως 7 και 8 έως 9 γραμμάτων και τρεις βαθμούς εξομάλυνσης: 3, 5, 7 και 9 σημείων.

Στην πρώτη περίπτωση πειραμάτων (δεδομένα εκπαίδευσης και ελέγχου λέξεις ιστορικού κειμένου) τα καλύτερα αποτελέσματα τα πετύχαμε για μικρό μέγεθος παρεμβολής, 67, και βαθμό εξομάλυνσης 3. Έτσι καταφέραμε να πετύχουμε κατηγοριοποίηση με ποσοστό 98,61% που θεωρείται πολύ καλό. Πρέπει να τονιστεί επίσης ότι οι λέξεις που δεν κατηγοριοποιήθηκαν στην σωστή κλάση είχαν κατά κύριο λόγο έντονες μορφές θορύβου, όπως μεγάλα μαύρα σημάδια.

Στα επόμενα πειράματα, που θεωρούνται πιο σημαντικά, για δεδομένα εκπαίδευσης όπως είδαμε χρησιμοποιήθηκε απλό τυπωμένο κείμενο γραμματοσειράς Times New Roman και μέγεθος 12 και δύο κατηγορίες δεδομένων ελέγχου.

- Στην πρώτη περίπτωση, με τις λέξεις αγγλικού ιστορικού κειμένου καλύτερα αποτελέσματα πετύχαμε για χαμηλή πάλι παρεμβολή, 67, και εξομάλυνση 3 και το ποσοστό επιτυχίας φτάνει το 85,18%.
- Στην επόμενη περίπτωση, ελληνικό ιστορικό κείμενο για δεδομένα ελέγχου, το ποσοστό επιτυχίας φτάνει το 56,25% και το πετυχαίνουμε κυρίως για μικρή παρεμβολή (μέγεθος 67) και βαθμό εξομάλυνσης 3 είτε και στα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου, είτε αν εφαρμοστεί μόνο στα δεδομένα εκπαίδευσης και εφαρμοστεί διαφορετική εξομάλυνση στα δεδομένα ελέγχου.

Τέλος έχουμε την περίπτωση του χειρόγραφου κειμένου, όπου το ποσοστό επιτυχίας φτάνει στο 36% για εικόνες-λέξεις που προέρχονται από 15 διαφορετικούς συγγραφείς και μικρή παρεμβολή, μέγεθος 67. Η εξομάλυνση που εφαρμόστηκε εδώ ήταν 3 σημείων μόνο για τα δεδομένα ελέγχου, ενώ τα δεδομένα εκπαίδευσης παρέμειναν όπως ήταν. Στην περίπτωση που χρησιμοποιήσαμε εικόνες-λέξεις από 3 μόνο συγγραφείς, το ποσοστό επιτυχίας έφτασε στο 40%, ενώ η εξομάλυνση δεν προσέφερε καμία βελτίωση ακόμα και

όταν εφαρμόστηκε με διαφορετικούς βαθμούς στα δυο σύνολο δεδομένων, εκπαίδευσης και ελέγχου.

Σε γενικές γραμμές πάντως μπορούμε να πούμε ότι καλύτερα αποτελέσματα είχαμε με μικρότερη παρεμβολή, μέγεθος 67 ή 129, αν και σε κάποιες περιπτώσεις φάνηκε ότι η μεγάλη παρεμβολή 230 είχε καλά αποτελέσματα. Στο σύνολό της όμως, άμα κρίνουμε από τα διάφορα διαγράμματα του κεφαλαίου 4, είχε χαμηλότερες τιμές από την παρεμβολή 67. Αυτό βέβαια οφείλεται στο γεγονός ότι οι λέξεις που χρησιμοποιήθηκαν στο σύστημα μας ήταν κατά κύριο λόγο μικρές σε μέγεθος και η παρεμβολή δεν τις αλλοίωνε πολύ.

Επίσης είδαμε ότι στις περισσότερες περιπτώσεις η εφαρμογή εξομάλυνσης 3 σημείων στις τιμές των χαρακτηριστικών βοήθησε στην βελτίωση των αποτελεσμάτων, ενώ αν εφαρμόζαμε επιπλέον εξομάλυνση χάνονταν πολύτιμη για το σχήμα των λέξεων πληροφορία, με αποτέλεσμα τα ποσοστά επιτυχίας να πέφτουν. Κρίνοντας πάλι από τα διαγράμματα η μεγαλύτερη εξομάλυνση έδινε καλύτερα αποτελέσματα σε συνδυασμό με μεγαλύτερη παρεμβολή.

Μπορούμε πάντως να πούμε ότι το υποσύστημα που προτείνουμε έχει πολύ καλά αποτελέσματα στην περίπτωση της κατηγοριοποίησης αγγλικού ιστορικού κειμένου με βάση το τυπωμένο κείμενο, που είναι ένα από τα βασικά ζητούμενα. Επιπλέον βελτιώσεις, πρέπει πάντως να γίνουν στην κατεύθυνση του χειρόγραφου κειμένου.

Κεφάλαιο 6

Μελλοντική Εργασία

Το προτεινόμενο υποσύστημα αναγνώρισης ολόκληρης λέξης παρουσιάζει καλά αποτελέσματα για κάποιες κατηγορίες λέξεων, ενώ σε άλλες όπως το χειρόγραφο κείμενο δεν είναι τόσο καλά. Πρέπει να γίνουν επιπλέον βελτιώσεις για την περίπτωση αυτή, όπως η ενσωμάτωση επιπλέον χαρακτηριστικών που θα βοηθούν την αναγνώριση. Ως τώρα τα χαρακτηριστικά που χρησιμοποιήσαμε περιγράφουν το εξωτερικό σχήμα της λέξης. Ίσως αυτό που χρειαζόμαστε είναι ένα χαρακτηριστικό που θα μας δίνει στοιχεία και για το εσωτερικό της σχήμα.

Βελτίωση πάντως μπορεί να προκύψει και με μια παραλλαγή της εξαγωγής των χαρακτηριστικών των πάνω και κάτω ουρών. Αντί να κρατάμε την πρώτη στήλη στην οποία εμφανίζεται μια ουρά, μπορούμε να κρατάμε την στήλη η οποία εμφανίζει την μεγαλύτερη συγκέντρωση μελανιού (το *peak*) σε σχέση με τις υπόλοιπες στήλες που περιγράφουν την ουρά. Η μεταβολή αυτή είναι πιθανό να έχει καλύτερα αποτελέσματα στο χειρόγραφο κείμενο, όπου ο κάθε συγγραφέας γράφει τις ουρές με το δικό του τρόπο και το δικό του μέγεθος και μπορεί να διαφέρουν ακόμα και αν προέρχονται από το ίδιο άτομο.

Το υποσύστημα πάντως που προτείναμε έχει ένα βασικό μειονέκτημα: μπορεί να αναγνωρίσει λέξεις που ανήκουν μόνο στο λεξιλόγιο εκπαίδευσης. Πρέπει λοιπόν να γίνουν βελτιώσεις και σε αυτόν τον τομέα, όπως για παράδειγμα το λεξιλόγιο να δημιουργείται δυναμικά κατά την εκπαίδευση του.

Κεφάλαιο 7

Βιβλιογραφία

[1] T. Steihhertz, E. Rivlin, N. Intrator, "Off-line cursive word recognition –A Survey", International Journal on Document Analysis and Recognition, Vol 2, Issue 2-3, pp 90-110, 1999

[2] E. Kavallieratou, N. Fakotakis, G. Kokkinakis "Handwritten character recognition based on structural characteristics" 16th International Conference on Pattern Recognition, 2002, pp 139-142

[3] Min Soo Kim , Kyu Tae Cho , Hee Kue Kwag and Jin Hyung Kim "Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents" Book: Document Analysis Systems VI, Volume 3163/2004, pp 114-124

[4] A. Vinciarelli "A survey on off-line cursive word recognition", Pattern Recognition, Vol 35, no 7, pp 1433-1446, 2002

[5] J. Rocha and T. Pavlidis. "New method for word recognition without segmentation." In *Proceedings of SPIE*, volume 1906, page 76, 1993

[6] S. Madhvanath, V. Govindaraju, "The role of holistic paradigms in handwritten word recognition", Trans. On Pattern Analysis and Machine Intelligence 23:2, pp 149-164, 2001

[7] K. Ntzios, B.Gatos, I.Pratikakis, T.Konidaris, S.J.Perantonis, "An Old Greek Handwritten OCR System", IEEE 2005, pp 64-68

[8] V.Faber, "Clustering and the Continuous k-Means Algorithm", Los Alamos Science, No 22, pp138-144, 1994

[9] T. Kanungo, N.S.Netanyahu, A.Y.Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol24, No7, pp 881-892 , 2002

[10] V.Lavrenko, T.M.Rath, R.Manmatha: "Holistic Word Recognition for Handwritten Historical Documents", DIAL, pp 278-287,2004

[11] T.Pavlidis, J.Zhou, "Page Segmentation by White Streams" ICDAR, Int Assoc. Pattern Recognition, pp 945-953,1991

[12] Peake, Tan "A General Algorithm for Document Skew Angle Estimation", IEEE Int. Conf. On Image Processing, vol2, pp230-233,1997

[13] E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, "Un Off-line Unconstrained Handwritting Recognition System", International Journal of Document Analysis and Recognition, no 4, pp. 226-242, 2002

[14] Heute, Paquetet, Moreau, Lecourtier, Olivier, "A structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, vol 19, pp 629-641, 1998

[15] T.Blu, P.Thevenaz, M.Unser, "How a Simple Shift can Significalty Improve the Performance of Linear Interpolation", IEEE ICIP pp 377-380,2002

[16] P. Thevenaz, T. Blu, and M. Unser, "Image interpolation and resampling," in *Handbook of Medical Imaging,Processing and Analysis*, I.N. Bankman, Ed., pp.393–420. Academic Press, San Diego CA, USA, 2000.

[17] Meijering, Erik. "A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing." *Proceedings of the IEEE*. vol. 90, no. 3, pp. 319-42. March 2002.

[18] G. Wolberg, *Digital Image Warping*, IEEE Computer Society Press, Los Alamitos, CA, 1990.

[19] O' Gorman, "The document spectrum for page layout analysis", IEEE trans On Pattern Analysis and Machine Intelligence, vol 15, pp 1162-1173, 1993

[20] E.Kavallieratou, N.Dromazou, N.Fakotakis, G.Kookinakis, " An Integrated System for Handwritten Document Image Processing", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 17, No. 4 ,pp 617-636, 2003

[21] M. Sarfraz, A.Zidouri, S.A. Shahab, "A novel approach for skew estimation of document images in OCR system" , International Conference on Computer Graphics, Imaging and Vision, pp 175-180, 2005

[22] A.Senior, A.Robinson, "An off-line Cursive Handwritting Recognition System", IEEE Trans. On Pattern Analysis and Machine Intelligence, vol 20, No3, pp 309-321, 1998

[23] B. Zhang, S. Shrihari, C.Huang, "Word Image Retrieval using Binary Features", SPIE, Document Recognition and Retrieval XI, San Jose, California, USA, 2004

[24] T. M. Rath, R. Manmatha, “Word Image Matching Using Dynamic Time Warping.” CVPR (2) pp. 521-527, 2003

[25] K.Nitzios, B.Gatos, I.Pratikakis, T.Konidaris, S.J. Peaantonis, “An Old Greek Handwritten OCR System”, IEEE pp 64-68, 2005

[26] S.Kuo, O.E. Agazzi, “Keyword spotting in poorly printed documents using pseudo 2-d hidden Markov models”, IEEE Trans. Pattern Analysis and Machine Intelligence 16, pp 842-848, 1994

[27] M. Shridhar, A. Badreldin, “Recognition of isolated and connected handwritten numerals”, Proc. IEEE International Conference on Systems, Man and Cybernaics, pp 142,146, 1984

[28] V.K. Govindan, A.P. Shivaprasad, “Character Recognition-A Review”, Pattern Recognition, Vol 23, No 7, pp 671-683, 1990

[29] N.B.Amor, N.E.B.Amara, “Combining a hybridic approach for feature selection and Hidden Markov Models in Multifont Arabic Characters Recognition”, DIAL’ 06, pp 103-114, 2006

[30] X.L. Xie, M.Suk, “On Machine Recognition of Handprinted Chinese Characters, by feature relaxation”, Pattern Recognition 21, pp 1-7, 1988

[31] A.K. Jain , “Tutorial on Statistical Pattern Recognition” , In Proc. 10th Interanational Conference on Pattern Recognition, pp 8-20, 1990

Κατάλογος Πινάκων

Πίνακας 4.1:	Κατηγοριοποίηση με χρήση κάθετου ιστογράμματος, πάνω και κάτω προφίλ	51
Πίνακας 4.2:	Κατηγοριοποίηση με χρήση επιπλέον πάνω και κάτω ουρών	52
Πίνακας 4.3:	Μεταβολή του μεγέθους της παρεμβολής	52
Πίνακας 4.4:	Εφαρμογή διαφορετικού βαθμού εξομάλυνσης	53
Πίνακας 4.5:	Κατηγοριοποίηση με χρήση κάθετου ιστογράμματος, πάνω και κάτω προφίλ	55
Πίνακας 4.6:	Κατηγοριοποίηση με χρήση επιπλέον πάνω και κάτω ουρών	55
Πίνακας 4.7:	Μεταβολή του μεγέθους της παρεμβολής	56
Πίνακας 4.8:	Εφαρμογή διαφορετικού βαθμού εξομάλυνσης (α) Παρεμβολή 67, (β) Παρεμβολή 230	56
Πίνακας 4.9:	Κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών	59
Πίνακας 4.10:	Μεταβολή του μεγέθους της παρεμβολής	59
Πίνακας 4.11:	Εφαρμογή διαφορετικού βαθμού εξομάλυνσης (α) Παρεμβολή 67 (β) Παρεμβολή 230	60
Πίνακας 4.12:	Χρήση διαφορετικού βαθμού εξομάλυνση για training και testing δεδομένα, Interpolation=67	62
Πίνακας 4.13:	Χρήση διαφορετικού βαθμού εξομάλυνση για training και testing δεδομένα, Interpolation=230	62
Πίνακας 4.14:	Κατηγοριοποίηση με χρήση όλων των χαρακτηριστικών	64
Πίνακας 4.15:	Μεταβολή του μεγέθους της παρεμβολής	64
Πίνακας 4.16:	Χρήση διαφορετικού βαθμού εξομάλυνσης	65
Πίνακας 4.17:	Χρήση διαφορετικού βαθμού εξομάλυνσης στα δεδομένα εκπαίδευσης και εισόδου	66
Πίνακας 4.18:	Κατηγοριοποίηση για 3 συγγραφείς	67

Κατάλογος Σχημάτων και Διαγραμμάτων

Σχήμα 1.1:	Διάφορα είδη εγγράφων	2
Σχήμα 1.2:	Γενικό Σύστημα Οπτικής Αναγνώρισης Χαρακτήρων	4
Σχήμα 1.3:	Μέθοδος Αναγνώρισης με Κατάτμηση	5
Σχήμα 1.4:	Μέθοδος Αναγνώρισης χωρίς Κατάτμηση	5
Σχήμα 2.1:	(α) Λέξη, (β) Κάθετο Ιστόγραμμα, (γ) Οριζόντιο Ιστόγραμμα	10
Σχήμα 2.2:	Διάφορα Είδη Προφίλ Λέξεων (α) Λέξη, (β) Πάνω Προφίλ, (γ) Κάτω Προφίλ	12
Σχήμα 2.3:	Αλγόριθμος Μεθόδου Παρεμβολής	14
Σχήμα 2.4:	Εφαρμογή Αλγορίθμου Παρεμβολής σε Διάνυσμα	15
Σχήμα 2.5:	Αλγόριθμος k-Means	17
Σχήμα 2.6:	Παράδειγμα Κατηγοριοποίησης με τη χρήση του k-Means	17
Σχήμα 3.1:	Γενικό Σύστημα Αναγνώρισης Ολόκληρης Λέξης	20
Σχήμα 3.2:	Στάδια Προτεινόμενου Υποσυστήματος	21
Σχήμα 3.3:	Ατέλειες Εγγράφων (α) Έγγραφο με Γωνία Εκτροπής, (β) Λέξη με Κλίση Χαρακτήρων, (γ) Θόρυβος	22
Σχήμα 3.4:	Κάθετο Ιστόγραμμα Λέξης πριν τον Καθαρισμό	23
Σχήμα 3.5:	Παραδείγματα Λέξεων Πριν και Μετά τον Καθαρισμό	24
Σχήμα 3.6:	Εξαγωγή Χαρακτηριστικών Λέξεων	25
Σχήμα 3.7:	Διαφορετικές Εμφανίσεις Ίδιων Λέξεων και τα αντίστοιχα Κάθετα Ιστογράμματα τους	26
Σχήμα 3.8:	Παράδειγμα Διανύσματος Κάθετου Ιστογράμματος	26
Σχήμα 3.9:	Συνάρτηση Υπολογισμού Πάνω Προφίλ Λέξης	27
Σχήμα 3.10:	Συνάρτηση Υπολογισμού Κάτω Προφίλ Λέξης	28
Σχήμα 3.11:	Διαφορετικές Εμφανίσεις Ίδιων Λέξεων και τα αντίστοιχα Πάνω και Κάτω Προφίλ τους	28
Σχήμα 3.12:	Παραδείγματα Διανυσμάτων για Πάνω και Κάτω Προφίλ (α) Διάνυσμα Πάνω Προφίλ (β)Διάνυσμα Κάτω Προφίλ	29
Σχήμα 3.13:	Διαφορετικά Κάθετα Ιστογράμματα για Διαφορετικές Εμφανίσεις Ίδιων λέξεων	30
Σχήμα 3.14:	Παράδειγμα Κανονικοποιημένου Κατά Ύψος Διανύσματος Πάνω Προφίλ	30
Σχήμα 3.15:	Πάνω Προφίλ Διαφορετικών Εμφανίσεων Ίδιων Λέξεων	31
Σχήμα 3.16:	Εφαρμογή Παρεμβολής στο Κάθετο Ιστόγραμμα μεγέθους 150	32
Σχήμα 3.17:	Εφαρμογή Παρεμβολής στο Πάνω Προφίλ μεγέθους 100	33
Σχήμα 3.18:	Παράδειγμα Κανονικοποιημένου Διανύσματος	

	κατά μήκος Πάνω Προφίλ	33
Σχήμα 3.19:	Διαδικασία Εντοπισμού Πάνω και Κάτω Ουρών	34
Σχήμα 3.20:	Παράδειγμα Κυρίους Σώματος Λέξης (α) Λέξη (β) Κυρίως Σώμα	34
Σχήμα 3.21:	Συνάρτηση Υπολογισμού Κυρίους Σώματος Λέξης	35
Σχήμα 3.22:	Όρια Κυρίου Σώματος Λέξης στο Οριζόντιο Ιστόγραμμα	35
Σχήμα 3.23:	Πάνω Τμήμα Λέξης	36
Σχήμα 3.24:	Συνάρτηση Υπολογισμού Πάνω Τμήματος Λέξης	36
Σχήμα 3.25:	Συνάρτηση Εντοπισμού Πάνω Ουρών Λέξης	37
Σχήμα 3.26:	Παράδειγμα Διανύσματος Πάνω Ουρών	38
Σχήμα 3.27:	Κάτω Τμήμα Λέξης	38
Σχήμα 3.28:	Συνάρτηση Υπολογισμού Κάτω Τμήματος Λέξης	38
Σχήμα 3.29:	Συνάρτηση Εντοπισμού Κάτω Ουρών Λέξης	39
Σχήμα 3.30:	Παράδειγμα Διανύσματος Κάτω Ουρών	40
Σχήμα 3.31:	Συναρτήσεις Εφαρμογής Εξομάλυνσης (α) Εξομάλυνση 3 (β) Εξομάλυνση 5	41
Σχήμα 3.32:	Χαρακτηριστικά Λέξεων Απλά και με Χρήση Εξομάλυνσης	42
Σχήμα 3.33:	Συνάρτηση Αλγορίθμου k-Means	44
Σχήμα 3.34:	Διαδικασία Κατηγοριοποίησης στο Σύστημά μας	45
Σχήμα 4.1:	Παραδείγματα Λέξεων που Εισάγονται στο Σύστημα (α) Αγγλικές Λέξεις, (γ) Ελληνικές Λέξεις	48
Σχήμα 4.2:	Λέξεις από Ελληνική Βάση	48
Σχήμα 4.3	Παράδειγμα Φόρμας Ελληνικής Βάσης	49
Διάγραμμα 4.1:	Μεταβολή του μεγέθους της Παρεμβολής	52
Διάγραμμα 4.2:	Εφαρμογή διαφορετικού βαθμού εξομάλυνσης	53
Διάγραμμα 4.3:	Σχέση Παρεμβολής και Εξομάλυνσης	53
Διάγραμμα 4.4:	Μεταβολή ποσοστού επιτυχίας ανά κλάση για Διάφορους Βαθμούς Εξομάλυνσης	54
Διάγραμμα 4.5:	Μεταβολή του μεγέθους της παρεμβολής	56
Διάγραμμα 4.6:	Εφαρμογή Διαφορετικού Βαθμού Εξομάλυνσης	57
Διάγραμμα 4.7:	Σχέση Παρεμβολής και Εξομάλυνσης	57
Διάγραμμα 4.8:	Ποσοστό επιτυχίας ανά κλάση για Διαφορετικούς Βαθμούς Εξομάλυνσης	58
Διάγραμμα 4.9:	Μεταβολή του μεγέθους της παρεμβολής	59
Διάγραμμα 4.10:	Εφαρμογή διαφορετικού βαθμού εξομάλυνσης	60
Διάγραμμα 4.11:	Σχέση Εξομάλυνσης με Παρεμβολή	61
Διάγραμμα 4.12:	Ποσοστό Επιτυχίας ανά κλάση για Εξομάλυνση Χωρίς και 3	61

Διάγραμμα 4.13: Σχέση Παρεμβολής με Διαφορετική Εξομάλυνση στα Δεδομένα Εκπαίδευσης και Δεδομένα Ελέγχου 63
Διάγραμμα 4.14: Μεταβολή του μεγέθους της παρεμβολής 64
Διάγραμμα 4.15: Χρήση διαφορετικού βαθμού εξομάλυνσης 65
Διάγραμμα 4.16: Ποσοστό Επιτυχίας ανά κλάση για διαφορετικούς βαθμούς εξομάλυνσης 65
Διάγραμμα 4.17: Χρήση διαφορετικού βαθμού εξομάλυνσης στα δεδομένα εκπαίδευσης και εισόδου 66
Διάγραμμα 4.18: Μεταβολή του Αριθμού Συγγραφέων 67