

Τμήμα

Πληροφοριακά Συστήματα και Ψηφιακή Καινοτομία

Τίτλος

**«Βιβλιογραφική ανασκόπηση η μεταγραφή χειρογράφων του
Μεσαίωνα χρησιμοποιώντας σύγχρονες τεχνικές μάθησης»**

**Διατριβή η οποία υποβλήθηκε προς απόκτηση μεταπτυχιακού τίτλου
σπουδών στα Πληροφοριακά Συστήματα και τη Ψηφιακή Καινοτομία**

Ονοματεπώνυμο: Κώτσιας Χρήστος

Εποπτεύων εκπαιδευτικός: Κωνσταντίνος Ζαγόρης

2021

Πνευματικά δικαιώματα

Copyright © Κώτσιας Χρήστος, 2021

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διατριβής από το Πανεπιστήμιο Νεάπολις δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Πανεπιστημίου.

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας:

Τίτλος Μεταπτυχιακής Διατριβής:

.....

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:

Πρώτος επιβλέπων (Πανεπιστήμιο Νεάπολις Πάφος).....[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα, υπογραφή]

Ή ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Ο Κώστας Χρήστος, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Βιβλιογραφική ανασκόπηση η μεταγραφή χειρόγραφων του Μεσαίωνα χρησιμοποιώντας σύγχρονες τεχνικές μάθησης», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Ο/Η Δηλών /σα

Πίνακας συντομογραφιών

Παρακάτω παρατίθενται ορισμένα από τα πιο συχνά χρησιμοποιούμενα ακρωνύμια της παρούσας διπλωματικής εργασίας:

AI – Artificial Intelligence – Τεχνητή Νοημοσύνη

HTR – Handwriting Text Recognition – Αναγνώριση Χειρόγραφου Κειμένου

KWS – Key Word Spotting – Εντοπισμός με λέξεις κλειδιά

ML – Machine Learning – Μηχανική Μάθηση

NN – Neural Network – Νευρωνικό Δίκτυο

ANN – Artificial Neural Network – Τεχνητό Νευρωνικό Δίκτυο

DL – Deep Learning – Βαθιά Μάθηση

DNN – Deep Neural Network – Βαθιά Νευρωνικά Δίκτυα

CNN – Convolutional Neural Network – Συνελκτικά Νευρωνικά Δίκτυα

SVM – Support Vector Mechanism – Μηχανή Διανυσματικής Υποστήριξης.

BoVW – Bag of Visual Words – Σάκος των λέξεων.

Περίληψη

Βασικός στόχος της διατριβής είναι να μελετήσει έναν τομέα ο οποίος είναι ιδιαίτερα χρήσιμος σε επαγγελματίες ερευνητές ιστορίας αλλά και ψηφιακές βιβλιοθήκες. Αφορά τη δυνατότητα που μπορεί να δώσει η τεχνολογία σε εκατομμύρια χειρόγραφα που είναι είτε σε ψηφιοποιημένες εικόνες και απλά διατηρούνται ως ψηφιακά κειμήλια δίχως metadata χωρίς να μπορούν να μετεγγραφούν και να είναι προσβάσιμα για ανάγνωση και μελέτη. Πρόκειται για μια προσπάθεια βιβλιογραφικής ανασκόπησης της μετεγγραφής ψηφιοποιημένων ιστορικών χειρόγραφων βιβλίων. Σε αυτή την προσπάθεια θα γίνει μια περιγραφή των προσπαθειών που έχουν υπάρξει μέχρι σήμερα αλλά και η αποτελεσματικότητά τους.

Επίσης βασικό στοιχείο της μελέτης αποτελούν οι μέθοδοι χρήσης που ακολουθούνται προκειμένου να γίνει η καλύτερη δυνατή μετεγγραφή των χειρογράφων καθώς εδώ αναφέρονται μια σειρά από προβλήματα τα οποία και θα αναφερθούν όπως για παράδειγμα οι διαφορές που υπάρχουν ανά χρονική περίοδο των τρόπων που γράφονται οι χαρακτήρες, οι μεταβολές που γίνονται κατά την πάροδο των χρόνων ανά περιοχή σε θέματα διαλέκτου και ερμηνείας, οι συντομογραφίες οι ιδιαιτερότητες ανά συγγραφέα που χρησιμοποιούν τον δικό τους ιδιαίτερο τρόπο γραφής καθώς και πιθανοί λόγοι κατά τους οποίους το μελάνι απλώνει επάνω στο κείμενο και το υπολογιστικό σύστημα αναγνώρισης του κειμένου να έχει τη δυνατότητα να το κατανοήσει.

Ειδικότερα θα γίνει μια προσπάθεια περιγραφής από μια σειρά από πλατφόρμες αυτόματης αναγνώρισης χειρόγραφου υλικού και ο τρόπος λειτουργίας και χρήσης τους όπως η transkribus η οποία αποτελεί ένα τέτοιο εξαιρετικό εργαλείο. Περιλαμβάνει φιλικό περιβάλλον εργασίας στο διαδίκτυο.

Επιπλέον η συνεισφορά και η χρήση της τεχνητής νοημοσύνης και άλλων τεχνολογιών όπως η μηχανική μάθηση η βαθιά μάθηση κλπ για πρόσβαση σε κείμενα του μεσαίωνα είναι ιδιαίτερα σημαντική όπως και στην ανάλυση υπάρχοντων μεθόδων όπως τα λεγόμενα συστήματα Αναγνώρισης Κειμένου Χειρογράφων (HTR Systems) όπως επίσης και στον εντοπισμό λέξεων (KWS).

Τα παραπάνω είναι ιδιαίτερα σημαντικά για την κατανόηση της εποχής κατά την οποία τα χειρόγραφα αυτά μπορούν μέσω της μετεγγραφής να γίνουν προσβάσιμα τόσο σε ευρύτερο κοινό όσο και σε επαγγελματίες του χώρου ιστορικούς κλπ. Πρόκειται για μοναδική δυνατότητα που προσφέρεται μέσω της τεχνολογίας, όχι μόνο στη διάσωση του κοινού πολιτιστικού και ιστορικού μας παρελθόντος μέσω της ψηφιοποίησης αλλά και εν συνεχεία μέσω τεχνικών αναγνώρισης κειμένου της ανάδειξής τους, της πρόσβασης και διάθεσής τους μέσω της αναζήτησής του διαδικτύου παρέχοντας τα στο ευρύτερο κοινό.

Λέξεις κλειδιά: Αναγνώριση χειρόγραφου κειμένου, εντοπισμός λέξεων, μέθοδοι και τεχνικές, τεχνητή νοημοσύνη, βαθιά μάθηση, μηχανική μάθηση, μοντέλα μηχανικής μάθησης, τεχνητά νευρωνικά δίκτυα, Συνελικτικά δίκτυα, οπτικοποίηση υπολογιστών.

Abstract

The main goal of the dissertation is to study an area that is particularly useful to professional history researchers as well as digital libraries. It is about the ability that technology can give to millions of manuscripts that are either in digitized images and are simply preserved as digital heirlooms without metadata and also without being able to be copied and accessible for reading and study. This is an attempt at a bibliographic review of the transcription of digitized historical manuscripts. In this effort, a description will be made of the efforts that have taken place so far and their effectiveness.

Also a key element of the study are the methods used in order to make the best possible transcription of the manuscripts as here arise a number of problems which will be reported such as for example the differences that exist over time of the ways in which the characters are written, the changes over time by region in terms of dialect and interpretation, abbreviations, peculiarities by author using their own particular way of writing as well as possible reasons why ink spreads on text and computer text recognition be able to understand it.

In particular, an attempt will be made to describe a number of platforms for automatic recognition of handwritten material and how they work and use such as transkribus which is such an excellent tool. Includes user friendly internet interface.

In addition, the contribution and use of artificial intelligence Deep learning Machine Learning Big Data technologies to access medieval texts is particularly important as in the analysis of existing methods such as the so-called Handwriting Recognition Systems (HTR Systems) and Key Word Spotting providing an evolution of great importance with extremely remarkable results.

The above are especially important for understanding the era in which these manuscripts can be made accessible to both the general public and professionals in the field of history through transcription, etc. This is a unique opportunity offered through technology, not only in rescue of our common cultural and historical past through digitization but also subsequently through text recognition techniques of their promotion, access and availability through internet search providing them to the general public

Key Words: Handwriting recognition, Key word spotting, Artificial Intelligence, Machine Learning, Deep Learning, Artificial Neural Network, Convolutional Neural Networks, Computer vision.

Ευχαριστίες

Με την ολοκλήρωση της εργασίας ένας προσωπικός στόχος μετά από αρκετούς κόπους επιτυγχάνεται και επιβραβεύεται. Δεν μένει η κούραση και ο κόπος αλλά η γλύκα της επίτευξης του στόχου και της ολοκλήρωσης αυτού του ταξιδιού.

Θα ήθελα να ευχαριστήσω από καρδιάς τον επιβλέποντα καθηγητή μου Κωνσταντίνο Ζαγόρη για τις πολύτιμες συμβουλές, και την καθοδήγηση του όλο αυτό το διάστημα. Από την πρώτη στιγμή αντιλήφθηκε τις δυσκολίες που είχα τόσο στη διαχείριση της εργασίας από πλευράς γνώσεων αλλά και οικονομίας χρόνου λόγω επαγγελματικών και οικογενειακών υποχρεώσεων που καθιστούσαν δύσκολο το έργο μου. Η ανοχή του, η υπομονή του και οι συμβολή του ήταν καθοριστικής σημασίας.

Επίσης θα ήθελα να ευχαριστήσω ονομαστικά τους συναδέλφους μου, Χρίστο, Κική και Μαργαρίτα για την ανοχή και την στήριξή τους.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου τη σύζυγό μου Χριστιάνα και τα παιδιά μου που αποτέλεσαν φάρο έμπνευσης για την υπέρβαση των όποιων δυσκολιών, όπως και στους αγαπημένους μου γονείς Βασίλη και Ζαφειρία που αν και μακριά τους αισθανόμουν πάντα δίπλα μου αρωγούς και συμπαραστάτες.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1°	13
1. Εισαγωγή. Περιγραφή της αξίας της μεταγραφής χειρογράφων του Μεσαίωνα για τις Βιβλιοθήκες τα Αρχεία και τους Ερευνητές.	13
1.1 Αυτόματη μεταγραφή χειρογράφων. Βασικά στάδια. Προκλήσεις.....	14
1.1.2 Ιστορική αναδρομή-Μέθοδοι αυτόματης μεταγραφής χειρογράφων-Εξέλιξη.....	16
1.2. Αυτόματη μεταγραφή χειρογράφων και επιστήμη της Παλαιογραφίας.....	20
1.3. Κρυφά μοντέλα MARKOV HIDDEN MARKOV MODELS (HMM).....	21
1.4. Νευρωνικά δίκτυα (NN).....	23
1.5. Εύρεση λέξεων κλειδιών (KWS).....	24
Κεφάλαιο 2°	28
2. Σύγχρονες εξελίξεις-τάσεις-μέθοδοι-Στάδια επεξεργασίας από την εικόνα στην μετεγγραφή και ανάκτηση κειμένου.....	28
2.1 Προκλήσεις-Τεχνικές.....	31
2.2 Φάσεις επεξεργασίας κειμένου και εικόνας.....	32
2.2.1 Είδη αρχειακών εγγράφων-Προβλήματα-Προκλήσεις.....	34
2.3 Αναγνώριση χειρογράφων-Επεξεργασία.....	39
2.3.1 Στάδια προ-επεξεργασίας	40
2.4 Τεχνικές αναγνώρισης χειρογράφου.....	43
2.5 Χρήση εργαλείων αναγνώρισης χειρόγραφου κειμένου.....	46
2.6 Βάσεις δεδομένων.....	47
2.7 Τάσεις σε Έργα ψηφιοποίησης.....	51
Κεφάλαιο 3ο Σύγχρονες κατευθύνσεις για περεταίρω έρευνα.....	56
3.Κατευθύνσεις για περεταίρω έρευνα.....	56
3.1 Συνελκτικτικά Νευρωνικά Δίκτυα CNN.....	57
3.1.1 Τρόπος λειτουργίας.....	58
3.2. Συνελκτικτικά νευρωνικά δίκτυα στην αναγνώριση κειμένου.....	59
3.3 Μηχανική μάθηση.....	60
3.4 Βαθιά μάθηση και συγκροτήματα νευρωνικών δικτύων.....	61

3.5 Εντοπισμός Λέξεων και χρήση συγκροτημάτων νευρωνικών δικτύων.....	62
3.6 Μεταφορά μάθησης.....	63
3.7 Μοντέλα Αρχιτεκτονικής συγκροτημάτων νευρωνικών δικτύων.....	65
3.8 Βάσεις δεδομένων και χρήση τους από συνελκτικά νευρωνικά δίκτυα ImageNet CLaMM16.....	68
3.9 Συνελκτικά νευρωνικά δίκτυα και διαδικασίες τμηματοποίησης.....	69
3.10 Συγκροτήματα Επαναλαμβανόμενων Νευρωνικών Δικτύων CRNN.....	72
3.11 Πλήρως Συνελκτικά Νευρωνικά Δίκτυα FCNN.....	73
3.12 Συνελκτικά νευρωνικά δίκτυα και αναγνώριση κειμένου-σχολιασμοί.....	75
Κεφάλαιο 4°. Εφαρμογές πλατφόρμες αναγνώρισης κειμένου TRANSCRIBUS και μDoc.ts.....	78
4.1 Πλατφόρμα αυτόματης μεταγραφής χειρογράφων Transcribus.....	78
4.2 Ελληνική πλατφόρμα αυτόματης μεταγραφής χειρογράφων μDoc.ts.....	80
Κεφάλαιο 5°. Πίνακες και αποτελέσματα συγκριτικών μεθόδων και εφαρμογών σε τεχνικές εύρεσης λέξεων (word spotting, kws) και αναγνώρισης χειρόγραφου κειμένου (Handwriting text recognition, HTR).....	83
5.1 Εντοπισμός λέξεων σε χειρόγραφο κείμενο. Μέθοδοι και αποτελέσματα 2016.....	83
5.1.2 Εντοπισμός λέξεων σε χειρόγραφο κείμενο. Μέθοδοι και αποτελέσματα 2015.....	89
5.1.3 Εντοπισμός λέξεων σε χειρόγραφο κείμενο. Μέθοδοι και αποτελέσματα 2015.....	94
5.2. Πίνακες και αποτελέσματα συγκριτικών μεθόδων και εφαρμογών σε τεχνικές αναγνώρισης κειμένου (Handwriting Text Recoginion, HTR).....	100
5.2.1. Πίνακες και αποτελέσματα συγκριτικών μεθόδων και εφαρμογών σε τεχνικές αναγνώρισης κειμένου (Handwriting Text Recoginion, HTR). Μέθοδοι και αποτελέσματα 2016.....	102
5.2.2 Πίνακες και αποτελέσματα συγκριτικών μεθόδων και εφαρμογών σε τεχνικές αναγνώρισης κειμένου (Handwriting Text Recoginion, HTR). Μέθοδοι και αποτελέσματα 2014.....	108
Κεφάλαιο 6°. Συμπεράσματα.....	109
Βιβλιογραφία.....	111

ΚΕΦΑΛΑΙΟ 1^ο: ΕΙΣΑΓΩΓΗ

1.ΕΙΣΑΓΩΓΗ. Περιγραφή της αξίας της μεταγραφής χειρογράφων του Μεσαίωνα για τις Βιβλιοθήκες τα Αρχεία και τους Ερευνητές

Στην παρούσα εργασία θα γίνει μια προσπάθεια βιβλιογραφικής ανασκόπησης πάνω στις νέες τεχνολογικές εφαρμογές έρευνες και σύγχρονες τεχνικές μάθησης που έχουν ως στόχο την μεταγραφή χειρογράφων του Μεσαίωνα.

Η αυτόματη μετεγραφή ιστορικών κειμένων είναι καθοριστικής αξίας και σημασίας για την δημιουργία των ψηφιακών βιβλιοθηκών. Η αυτόματη μεταγραφή μεσαιωνικών κειμένων αποτελεί μια τεράστια πρόκληση. Οι ψηφιακές Βιβλιοθήκες και τα αρχεία αποτελούν πηγή πληθώρας πληροφοριών εμπλούτισης και αξιοποίησής τους. Πολλές έχουν αναλάβει μεγάλης κλίμακας προγράμματα ψηφιοποίησης επεκτείνοντας τις ψηφιακές τους συλλογές και παρέχοντας σε χρήστες τους υπερ-πολύτιμα δεδομένα.

Συνήθως σε ότι αφορά τα χειρόγραφα ιστορικά κείμενα αυτά είναι συνήθως ψηφιοποιημένες εικόνες δίχως να ακολουθούνται από την μεταγραφή τους. Ως γεγονός αυτού του τύπου τα αρχεία δε βοηθούν τους ερευνητές στο έργο τους καθώς αποτελούν σοβαρό εμπόδιο για τη συνέχεια της έρευνας. Πρόκειται για αρχεία τύπου απλών εικόνων στα οποία ναι μεν υπάρχει απρόσκοπτη πρόσβαση στο υλικό αλλά δυσκολεύει στην κατανόηση και ερμηνεία του κειμένου καθώς απαιτεί όχι μόνο γνώσεις παλαιογραφίας αλλά και στις περισσότερες περιπτώσεις τη συνδρομή της τεχνολογίας για να γίνουν ευδιάκριτα τα γραφόμενα.

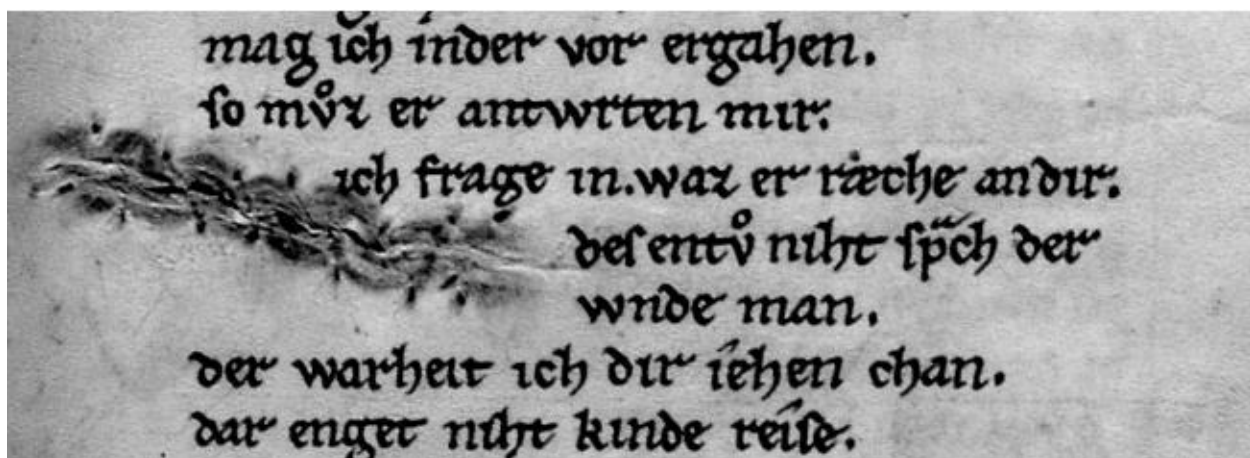
Σκοπός των ερευνητών σε συνδυασμό με την αλματώδη εξέλιξη της τεχνολογίας είναι να βρεθούν εκείνες οι εφαρμογές ή βελτιώσεις σε ήδη υφιστάμενες υποσχόμενες, που θα είναι σε θέση να μετουσιώσουν εικόνες μεγάλης αξίας και χρησιμότητας ιστορικών αρχείων με δυνατότητες αναζήτησης ανάκτησης και μεταγραφής σε υψηλής ποιότητας και πιστότητας σύγχρονα κείμενα. Η μέθοδος αναγνώρισης χειρογράφων ψηφιοποιημένων αρχείων ή φωτογραφικού τύπου εικόνες κειμένων είναι ένα διαδεδομένο ζητούμενο στην επιστημονική κοινότητα και στην επιστήμη της πληροφορικής.

Ειδικότερα δε τα τελευταία χρόνια έχει γίνει τεράστιες προσπάθειες για αναγνώριση χειρογράφων σε ιστορικά κείμενα. Μάλιστα προκειμένου να διασωθούν και να προστατευθούν ανεκτίμητης αξίας ιστορικά αρχεία, τεράστιος όγκος από αυτά έχουν ψηφιοποιηθεί. Ως αποτέλεσμα της δημιουργίας τεράστιου όγκου συλλογής αυτού του είδους έχει προκύψει μια άνευ προηγουμένου ζήτηση και ενδιαφέρον από ερευνητές στην ιστορία, τη φιλολογία, την ιατρική όπως επίσης και διάφορα πεδία της επιστήμης της πληροφορικής τα οποία θα φανούν χρήσιμα στην επεξεργασία και πρόσβαση του υλικού προς τις ψηφιακές βιβλιοθήκες. Ως συνέπεια των παραπάνω μια βάση που μπορεί να λειτουργήσει ως ευρετήριο περιεχομένων καθίσταται απαραίτητη για αναζήτηση και ανάκληση δεδομένων από εικόνες.

1.1 ΑΥΤΟΜΑΤΗ ΜΕΤΑΓΡΑΦΗ ΧΕΙΡΟΓΡΑΦΩΝ (HTR). ΒΑΣΙΚΑ ΣΤΑΔΙΑ. ΠΡΟΚΛΗΣΕΙΣ

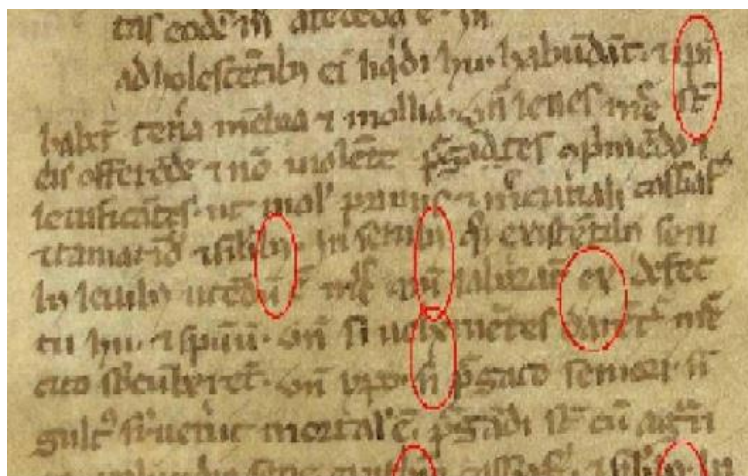
Ένα από εκείνα τα ερωτήματα που απασχολούν πολλούς από τους ερευνητές του χώρου είναι κατά πόσο ακριβής και ποιοτική είναι η όποια χρησιμοποιούμενη μέθοδος για αυτόματη μετεγγραφή χειρόγραφου κειμένου.

Μεταγραφή και αποκωδικοποίηση deciphering κειμένων του μεσαίωνα είναι μια πολύπλοκη και προβληματική διαδικασία, συνήθως λόγω συγκεκριμένων χαρακτηριστικών μεσαιωνικής γραφής καλλιγραφίας, τρόπου γραφής και συντήρησης. Σε ότι αφορά τα ζητήματα της καλλιγραφίας, υπάρχει μεγάλη ποικιλία τύπων καλλιγραφίας, χαρακτήρες κεφαλαία γράμματα και χρήση της λατινικής γλώσσας που είναι μια νεκρή πλέον γλώσσα. [MALAGON C., 2012]



Εικ. 1. Ραφή που προκύπτει από προσπάθεια συντήρησης επιδιόρθωσης περγαμνής.

Επιπλέον παράγοντας δυσκολίας είναι οι συντομογραφίες (abbreviations), ο εντοπισμός τους και η αναγνώρισή τους. Αυτά εμφανίζονται σε πληθώρα με διαφορετικούς τύπους σχήματα και θέσεις μέσα σε κείμενα συνήθως πιο εξειδικευμένα όπως στην ιατρική κλπ. Πρόβλημα αποτελεί επίσης η περιορισμένη πρόσβαση σε τέτοιου τύπου σπάνια εύθραυστα ντοκουμέντα.



Εικ. 2. Επικαλυπτόμενοι γραφικοί χαρακτήρες.

Επίσης έχει παρατηρηθεί διαφοροποίηση σε διαφορετικές χρονικές περιόδους τόσο των χαρακτήρων, οι οποίοι μεταβάλλονται και τροποποιούνται όπως επίσης θέματα με τις διαλέκτους και με τις ερμηνείες στη χρήση της γλώσσας. Βαθμός δυσκολίας αποτελεί και ο τρόπος γραφής ανά συγγραφέα οι οποίο χρησιμοποιούν το δικό τους ιδιαίτερο τρόπο γραφής καθώς και ζητήματα με το άπλωμα του μελανιού το οποίο καθιστά δυσδιάκριτα τα γράμματα αλλά και τις λέξεις με αποτέλεσμα να δυσκολεύει το αυτόματο υπολογιστικό σύστημα στην αναγνώριση και μετέπειτα στην μεταγραφή. (MALAGON,C et all, 2012)

Τέλος ζήτημα αποτελεί και η προσπάθεια για κατάτμηση του κειμένου ο διαχωρισμός των σειρών του κειμένου αλλά και των λέξεων καθώς πολλές φορές είναι δύσκολο τόσο να διαχωριστούν σωστά προκειμένου να αναγνωριστεί η δομή και εν συνεχεία η αλληλουχία των σειρών και των λέξεων που θα αποδώσουν ένα ακριβές και κατανοητό κείμενο.



Εικ. 3. Τμηματοποίηση γραμμής γραφικών χαρακτήρων.



Εικ. 4. Τμηματοποίηση λέξεων.

1.1.2 ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ-ΜΕΘΟΔΟΙ ΑΥΤΟΜΑΤΗΣ ΜΕΤΑΓΡΑΦΗΣ ΧΕΙΡΟΓΡΑΦΩΝ-ΕΞΕΛΙΞΗ

Η χρήση της τεχνολογίας Handwritten Recognition HTR μπορεί να επεκτείνει την υπάρχουσα ερευνητική υποδομή τόσο στους τομείς των αρχείων των Βιβλιοθηκών αλλά και των ανθρωπιστικών επιστημών. Πρόκειται για ενεργό τομέα της έρευνας και τεχνολογίας στην επιστήμη των υπολογιστών η οποία έχει την αφετηρία της στα μέσα του εικοστού αιώνα [Dimond 1957]. Αρχικά η τεχνολογία HTR ήταν σε συνάρτηση με την ανάπτυξη της τεχνολογίας OCR οπτική αναγνώριση κατά την οποία οι σαρωμένες εικόνες του τυπωμένου κειμένου ή του χειρόγραφου κειμένου μετατρέπονται σε κείμενο με βάση την κωδικοποίηση μηχανής, ή γενικότερα συγκρίνοντας μεμονωμένους χαρακτήρες με υπάρχοντα πρότυπα.

Το HTR εξελίχτηκε σε μια ερευνητική περιοχή από μόνο του λόγω της μεταβλητότητας διαφορετικών τρόπων γραφής και αντιμετώπισης των προβλημάτων και της υπολογιστικής πολυπλοκότητας της διαδικασίας. Σταδιακά τις στατιστικές εξελίξεις στην δεκαετία του 80' και την προηγμένη αναγνώριση προτύπων σε συνδυασμό με την τεχνητή νοημοσύνη την επόμενη δεκαετία του 1990, ήρθαν να ενισχύσουν η ανάπτυξη των βαθύς νευρωνικών δικτύων στις δεκαετίες του 2000 και του 2010.

Ο συνδυασμός των παραπάνω εξελίξεων, με την διαθεσιμότητα αυξημένης ισχύος επεξεργασίας υπολογιστών, έχει οδηγήσει σε βελτιώσεις στην αναγνώριση χειρόγραφων ιστορικών εγγράφων όπως αναφύεται σε έγκριτες μελέτες άρθρα και συνέδρια που αναφέρονται στις εξελίξεις στον τομέα της αναγνώρισης χειρογράφων.

Αρχικά οι ερευνητές ανέπτυξαν την πιο πάνω τεχνολογία, έχοντας υπόψιν τους χειρόγραφα υλικά, η οποία είναι ευρέως διαδεδομένη στον τομέα της πληροφορικής υπό τα αρχικά HTR. Ωστόσο, η ίδια τεχνολογία, δύναται να χρησιμοποιηθεί και να εφαρμοστεί εξίσου και σε πρώιμα τυπωμένα κείμενα που είναι αρκετά πιο περίπλοκα για επαρκή επεξεργασία μέσω των τεχνικών οπτικής αναγνώρισης OCR.

Πέρα από την αναγνώριση κειμένων του παρελθόντος πολλές από τις εφαρμογές της τεχνολογίας HTR ήταν στον εμπορικό (π.χ διερμηνεία ταχυδρομικών διευθύνσεων) στον χρηματοοικονομικό τομέα (επεξεργασία τραπεζικών επιταγών), επαλήθευση υπογραφής και ταυτοποίηση βιομετρικών στοιχείων συγγραφέων. Ωστόσο, οι πρόσφατες επιτυχίες της τεχνολογίας HTR συνίστανται στην διαθεσιμότητα προσιτής και μάλιστα υψηλής ποιότητας τεχνολογίας ψηφιακής απεικόνισης, σχετικών διαδικτυακών συστημάτων για τη φιλοξενία εικόνων και επακόλουθων προγραμμάτων μαζικής ψηφιοποίησης που συντελούνται από τις μεγαλύτερες Βιβλιοθήκες και αρχεία σε παγκόσμια κλίμακα προκειμένου να αυξήσουν τις δυνατότητες πρόσβασης στις συλλογές τους για το ευρύτερο κοινό.

Ένα από τα διαπιστωμένα προβλήματα κατά τη διάρκεια αυτής της διαδικασίας ήταν πως σε σημαντικό βαθμό σαρωμένων χειρόγραφων, οι τρέχουσες τεχνικές αναγνώρισης OCR και χειρογράφου εμφανίζουν μεγάλες δυσκολίες μεταγραφής καθώς τα συστήματα δεν έχουν εκπαιδευτεί ή δεν έχουν τις κατάλληλες βάσεις δεδομένων ή άλλως βιβλιοθήκες για ανάλογα σενάρια με βάση τα οποία γράφονται αυτά τα χειρόγραφα. Επομένως, έγγραφα αυτής της κατηγορίας προερχόμενα από πεφωτισμένα μεσαιωνικά χειρόγραφα, χειρόγραφα γράμματα έως πρώιμα έντυπα έργα πολλές φορές δεν υπάρχει η δυνατότητα αξιόπιστης μεταγραφής, με αποτέλεσμα αυτά και οι χρήσιμες πληροφορίες που εμπεριέχουν παραμένουν μη αναζητήσιμες.

Η μαζική ψηφιοποίηση ιστορικού υλικού χάρη στην χρήση της τεχνολογίας HTR σε συνδυασμό με τους παραδοσιακούς καταλόγους αρχειοθέτησης και εύρεσης βοηθημάτων, διευρύνει την πρόσβαση σε χρήσιμες πηγές και συλλογές εγγράφων. Η αυτοματοποιημένη μεταγραφή και αναζήτηση ψηφιοποιημένων κειμένων προχωρά εξελίσσεται και επεκτείνει τις υπάρχουσες δυνατότητες ιστορικής έρευνας για μελετητές ιδρύματα εμπορικούς παρόχους και όποιους άλλους δυνητικούς χρήστες.

Η επιτυχής ανάπτυξη του HTR θα βελτιώσει και θα αυξήσει την πρόσβαση στις συλλογές, επιτρέποντας στους χρήστες όχι μόνο στο να εντοπίζουν γρήγορα και αποτελεσματικά τα θέματα που τους ενδιαφέρουν καθώς και να απομονώνουν λέξεις, άτομα, ονόματα, μέρη και γεγονότα σε έγγραφα, αλλά και να αλλάζουν την κατανόηση του περιβάλλοντος και να πολλαπλασιάζουν τις ερευνητικές δυνατότητες.

Η δυνατότητα και δημιουργία αναγνώσιμων κειμένων αυτόματα και μηχανικά, θα παρέχει τη βάση για προηγμένης ικανότητας σημασιολογική, γλωσσική και γεω-χωρική υπολογιστική ανάλυση του πρωτογενούς ιστορικού υλικού. Σύμφωνα με δε με πρόσφατες μελέτες [Estill, 2016] τα ερευνητικά ερωτήματα που μπορούν να ζητηθούν από ιστορικά χειρόγραφα αλλάζουν, όπως επίσης ο τρόπος με τον οποίο θα παραδοθούν και θα παρουσιαστούν από τα ιδρύματα αρχειακό υλικό θα είναι αντίστοιχος.

Η εμπορική χρήση στο χώρο της ψηφιοποίησης έχει λάβει τέραστιες διαστάσεις και οι πάροχοι που δραστηριοποιούνται σε αυτό το χώρο αναλαμβάνοντας την ψηφιοποίηση διαφόρων ιδρυμάτων επιτρέπουν άδειες πρόσβασης στους πόρους που προκύπτουν από την επεξεργασία.

Αξίζει να αναφερθεί μια σχετικά πρόσφατη εξέλιξη της τεχνολογίας HTR προ 2 ετών ο Adam Matthew Digital ο οποίος αναφέρεται ως «ο μόνος εκδότης που χρησιμοποιεί τεχνητή νοημοσύνη η οποία παρέχει Χειρόγραφο Αναγνώριση Κειμένου για τις χειρόγραφες συλλογές της». (Adam Matthew Digital, 2018). Παρέχει το ίδιο λογισμικό με το Transkribus (αναφέρονται πληροφορίες στη συνέχεια της εργασίας για την εν λόγω πλατφόρμα), επιτρέποντας αναζήτηση με βάση το HTR σε πολλές από τις θεματικές ψηφιακές συλλογές αρχείων της και διαθέτει το

HTR ως μέρος μιας υπηρεσίας διαχείρισης συλλογών μέσω της πλατφόρμας Quartex. [MUEHLBERGER G., 2019]

Η εμπορική χρήση της τεχνολογίας HTR παρότι περιορίζει την χρήση της σε ιδρύματα και Οργανισμούς οι οποίοι την χρηματοδοτούν έχει ως μειονέκτημα και το γεγονός ότι ερευνητές και άλλοι χρήστες δεν θα μπορούν να αλληλεπιδράσουν και να συνεσιφέρουν στην εξέλιξη ανάπτυξη και εφαρμογή της τεχνολογίας. Η μηχανική μάθηση δεν αποτελεί σε καμία περίπτωση την λύση για όλα τα προβλήματα και η κριτική αξιολόγηση της εκπαιδευτικής διαδικασίας και τα βασικά δεδομένα που αυτή παρέχει θα πρέπει να κρίνονται απαραίτητα στο αν αυτή η τεχνολογία θα μπορεί και με ποια διαδικασία να ενσωματωθεί στην αρχειακή πρακτική και στην επιστημονική έρευνα με τρόπο ρεαλιστικό.

Σε επόμενο στάδιο θα ακολουθήσουν αναφορές και θα αναδειχθούν οι δυσκολίες και η ακρίβεια των εφαρμογών και συστημάτων επάνω στη μεταγραφή χειρογράφων και θα εστιαστούν σε σημεία και τρόπους με τα οποία τα παραπάνω θα μπορούν να εξελιχθούν και να βελτιωθούν παρέχοντας αποτελέσματα με τη μέγιστη δυνατή αξιοπιστία.

Η μεταγραφή χειρογράφων χωρίζεται σε online και offline. Η δε τεχνολογία η οποία είναι απαραίτητη για αναγνώριση όλο το φάσμα των στοιχείων σε ένα κείμενο, ήτοι προτάσεις, λέξεις, χαρακτήρες σαν ένα ενοποιημένο σύνολο αναφέρεται σαν off line HTR και πραγματοποιείται συνήθως μέσω OCR.

Η online Handwritting recognition ή διαδικτυακή αναγνώριση χειρόγραφου κειμένου, περιλαμβάνει την αυτόματη μετατροπή του, καθώς αυτό είναι περασμένο σε ειδικό ψηφιοποιητή ή άλλως το PDA. Εκεί ένας αισθητήρας λαμβάνει τις κινήσεις κινήσεις τις πένες πάνω στο χαρτί καθώς και την αλλαγή στη χρήση της από άλλη. Αυτό το είδος δεδομένων γνωστό ως ψηφιακή μελάνι μπορεί να θεωρηθεί ως ψηφιακή αναπαράσταση της γραφής. Το λαμβανόμενο σήμα μετατρέπεται σε κωδικούς γραμμάτων που μπορούν να χρησιμοποιηθούν σε εφαρμογές επεξεργασίας υπολογιστή και κειμένου.

Τα συστήματα αναγνώρισης γραφής εκτός σύνδεσης off line είναι λιγότερο ακριβή από εκείνα διαδικτυακής αναγνώρισης online, επειδή μόνο χωρικές πληροφορίες είναι διαθέσιμες για

συστήματα εκτός σύνδεσης, ενώ τόσο χωρικές όσο και χρονικές πληροφορίες είναι διαθέσιμες για τα online συστήματα αναγνώρισης γραφής.

1.2 ΑΥΤΟΜΑΤΗ ΜΕΤΑΓΡΑΦΗ ΧΕΙΡΟΓΡΑΦΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΗΣ ΠΑΛΑΙΟΓΡΑΦΙΑΣ

Η επιστήμη η οποία μελετά χειρόγραφα κείμενα και δη του Μεσαίωνα είναι η επιστήμη της Παλαιογραφίας. Ειδικότερα κείμενα παλαιογραφίας του Μεσαίωνα τα οποία δημιουργήθηκαν από τον 5^ο αι. περίπου μ. Χ μέχρι προ της ανακάλυψης της τυπογραφίας από τον Γουτεμβέργιο το 1453, έχουν ιδιαίτερα χαρακτηριστικά και η γνώση τους είναι απαραίτητη για να γίνει μεταγραφή παλαιογραφιών αλλά και η συμβολή αυτής της γνώσης στα ανάλογα προγράμματα και συστήματα πληροφορικής.

Ο συνδυασμός της γνώσης είναι απαραίτητη προϋπόθεση για τη συνεργασία μεταξύ ερευνητών και ειδικών σε τέτοιου είδους κείμενα προκειμένου να αναπτυχθούν τα τεχνολογικά εκείνα εργαλεία με τα οποία θα επιτευχθεί η καλύτερη δυνατή μεταγραφή και αυτόματη μετάφραση παλαιογραφικών κειμένων σε σύγχρονο κείμενο.

Τα μεσαιωνικά κείμενα αποτελούν μια τεράστια ποικιλία πληροφοριών τόσο από ιστορική θρησκευτική δημογραφική άποψη από τρόπο ζωής καθώς πρόκειται για ιστορικά κείμενα νοταριακά έγγραφα διαθήκες συμβολαιογραφικές πράξεις περιουσιακά στοιχεία γάμους βαφτίσεις κλπ καθώς και εκείνα που έχουν να κάνουν με την ιστορία της ιατρικής την επιστήμη της ιατρικής κατά το Μεσαίωνα. Η άντληση πληροφοριών μέσα από μια τέτοια πληθώρα κειμένων σημαντικής αξίας, με τη μεγαλύτερη δυνατή ακρίβεια και πιστότητα θα αποτελέσει ένα σημαντικό βήμα για την κατανόηση μια χρονικής περιόδου από τον Πρώιμο ως και τον Όρσιμο Μεσαίωνα. Το γεγονός ότι δεν έχει γίνει μεγάλη έρευνα σε θέματα αυτόματης μεταγραφής κειμένων παλαιογραφίας αφήνει πεδίο δόξης λαμπρό τόσο σε ερευνητικό όσο και τεχνολογικό επίπεδο. Τα ιστορικά ντοκουμέντα μετά από τη μακραίωνη διάσωσή τους σε φυσική μορφή πλέον στο στάδιο επεξεργασίας θα ξεκλειδώσουν πολλά από τα μυστικά της χειρόγραφης γραφής θα φανερώσουν και σημαντικές λεπτομέρειες του παρελθόντος

Σε αυτήν την χρονική περίοδο υπάρχουν μια σειρά από πολύ σημαντικά γεγονότα τα οποία διαμορφώνουν τον πολιτικό κοινωνικό δημογραφικό και θρησκευτικό χάρτη της Ευρώπης.

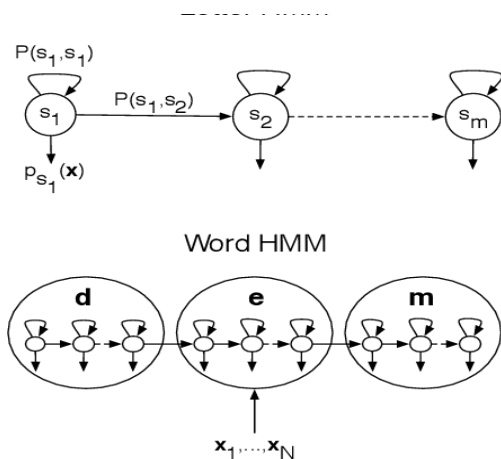
Ειδικότερα δε, θα δωθεί φώς σε περιόδους κατά τις οποίες δεν έχουμε επαρκή πληροφόρηση για τις ιστορικές και άλλες συνθήκες ώστε να κατανοήσουμε τις αιτίες και τα γεγονότα του μετασχηματισμού της Ηπείρου μας τα οποία οδήγησαν στη συνέχεια στη σημερινή διαμόρφωση. Στόχος είναι το τεράστιο αυτό υλικό να καταστεί προσβάσιμο στο ευρύτερο κοινό και δίχως τους σημερινούς περιορισμούς. Είναι κάτι παραπάνω από προφανές ότι η επιστήμη της ιστορίας θα επηρεαστεί σε σημαντικό βαθμό τόσο ως προς την καταγραφή της αλλά κυρίως από την άποψη και τον τρόπο της άντλησης αυτών των νέων δεδομένων.

Γνώσεις παλαιογραφίας, αρχειοθέτησης, αλλά και βιβλιοθηκονομίας κρίνονται πολύ σημαντικές καθώς τα μεταδεδομένα αποτελούν κύρια εργαλεία για την διάσωση των χειρογράφων σπάνιας αξίας, αλλά και μέσω των ψηφιακών βιβλιοθηκών όπου πέρα από τις εφαρμογές αναγνώρισης με αυτά θα είναι εφικτό να αναζητηθεί και να ανακτηθεί αυτό το υλικό σπάνιας αξίας.

1.3 ΚΡΥΦΑ ΜΟΝΤΕΛΑ MARKOV. HIDDEN MARKOV MODELS (HMM)

Ένα από τα πλέον γνωστά συστήματα αναγνώρισης χειρογράφου κειμένου είναι τα Hidden Markov Models, τα οποία στην πάροδο του χρόνου εξελίχτηκαν και αντικαταστάθηκαν από τα λεγόμενα Neural Networks.

Τα κρυφά μοντέλα Markov πήραν το όνομά τους από τον Ρώσο μαθηματικό Andrey Andreyevitch Markov ο οποίος και ανέπτυξε σημαντικό μέρος της σχετικής στατιστικής θεωρίας η οποία εισάχθηκε και μελετάται πίσω στις αρχές της δεκαετίας του 70'. Πρόκειται για στατιστικά μοντέλα για τη λήψη κρυφών πληροφοριών από παρατηρήσιμα διαδοχικά σύμβολα. Οι επαναλαμβανόμενες φανερές ακολουθίες χρησιμοποιούνται για να εντοπιστούν οι κρυφές.



Εικ. 5. Διάγραμμα Hidden Markov Model.

Για πάρα πολλά χρόνια τα HMM ήταν εξαιρετικά αποτελεσματικά στον τομέα της αυτόματης μεταγραφής κειμένων. Πρόκειται για ένα γραφικό μοντέλο πιθανοτήτων που χρησιμοποιείται συνήθως στην αναγνώριση και ταξινόμηση στατιστικών προτύπων. Έχει εφαρμοστεί με αρκετή επιτυχία στην αναγνώριση και την επανάληψη χρονικών μοτίβων στην περίπτωση μας τα χειρόγραφα την αποσαφήνιση της ερμηνείας λέξεων.

Αποτελείται από δύο μέρη. Κάθε HMM εμπεριέχει μια σειρά διακριτών, χρονικά ομόλογων αλυσίδων Markov πρώτης τάξης Markov Chains. Τα MC είναι μια αλυσίδα στην οποία υπάρχει διακριτή χρονική διαδικασία και πρόβλεψη για την επόμενη κατάσταση η οποία εξαρτάται από τα δεδομένα της τρέχουσας κατάστασης. Η κάθε κατάσταση έχει διακριτή ή συνεχή κατανομή πιθανοτήτων είτε για εκπομπές είτε για εξόδους. Η δημιουργία των εξόδων συμβαίνει κατά την μετάβαση από τη μία κατάσταση στην άλλη. Η πιθανότητα μετάβασης αφορά την πιθανότητα μετακίνησης από τη μία κατάσταση στην άλλη μέσω ενός συνδεδεμένου άκρου και η πιθανότητα εκπομπής ενός συγκεκριμένου συμβόλου γραμματος λέξης κλπ σε μια κατάσταση. Οι ακολουθίες των καταστάσεων μέσω των οποίων περνά το μοντέλο είναι κρυμμένες και δεν μπορούν να παρατηρηθούν, από αυτό το βασικό χαρακτηριστικό πήρε και το όνομά του κρυφό μοντέλο Markov. [FISCHER , 2009]

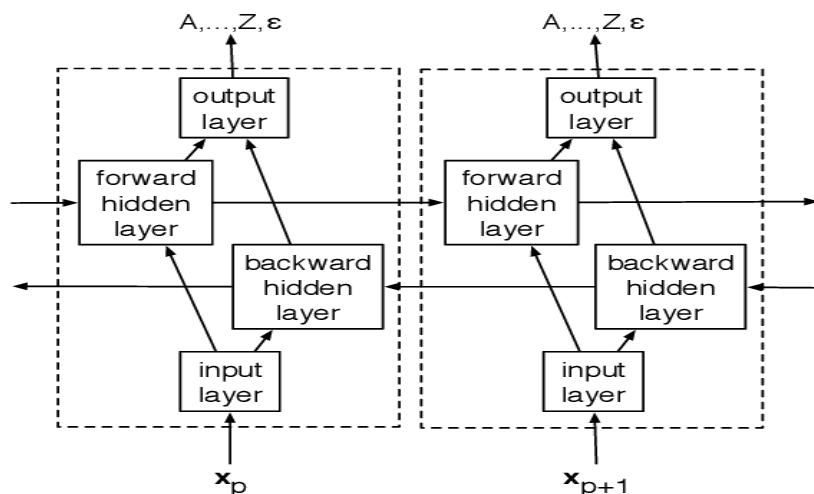
Το μαθηματικό μοντέλο των πιθανοτήτων και των ακολουθιών που χρησιμοποιείται είναι πως η πιθανότητα οποιασδήποτε ακολουθίας δεδομένου του μοντέλου, υπολογίζεται με πολλαπλασιασμό των πιθανοτήτων εκπομπής και μετάβασης κατά μήκος της διαδρομής.

1.4 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (NN)

Από την άλλη τα νευρωνικά δίκτυα είναι η εξέλιξη και η επόμενη γενιά στην αναγνώριση χειρόγραφων κειμένων καθώς αντικατέστησαν σταδιακά τα HMM. Χρησιμοποιούνται ως μέθοδος βαθιάς μάθησης *deep learning* η οποία είναι ένα από τα πεδία της τεχνητής νοημοσύνης *AI Artificial Intelligence*. Τα τελευταία χρόνια χάρη στην εξέλιξη της τεχνολογίας τα νευρωνικά δίκτυα πολλαπλασιάστηκαν όσον αφορά τις εφαρμογές τις οποίες μπορούσαν να προσομοιώσουν επομένως και την πολυπλοκότητα της μάθησης που μπορούν πλέον να επιτύχουν. Έχουν δημιουργηθεί πολύ βαθιά δίκτυα τα οποία έχουν εκπαιδευτεί και εξειδικευτεί σε τεράστια σύνολα δεδομένων για να επιτευχθούν καινοτομές στη νοημοσύνη των μηχανών.

Οι καινοτομίες αυτές επέτρεψαν στις μηχανές να ταιριάζουν να συγκρίνουν και να υπερβαίνουν τις ικανότητες των ανθρώπων στην εκτέλεση των αναλαμβανόμενων εργασιών. Οι τελευταίες εξελίξεις στο *deep learning* έχουν καταστήσει δυνατή τη δημιουργία νευρωνικών δικτύων που μεταξύ των άλλων μπορούν να αναγνωρίσουν κείμενα. Ένα παράδειγμα που χρησιμοποιείται ευρέως για να γίνει κατανοητή η φύση της μεθόδου, είναι μέσω της εφαρμογής *TensorFlow*. Αφορά τη δημιουργία βιβλιοθήκης *python* ανοιχτού κώδικα η οποία αναπτύχθηκε από τα εργαστήρια του *Google Brain* για έρευνα βαθιάς μάθησης. Μέσω της εφαρμογής μπορούν να εισαχθούν χειρόγραφες εικόνες των αριθμών από το 0-9 για δημιουργία και να χρήση τους προς εκπαίδευση ενός νευρωνικού δικτύου προκειμένου αυτό να αναγνωρίσει αλλά και να μπορεί να προβλέψει το σωστό ψηφίο το οποίο βρίσκεται στην εικόνα. Τα νευρωνικά δίκτυα και το *deep learning* συνδυάζονται και με μεθόδους μηχανικής μάθησης και τεχνητής νοημοσύνης.

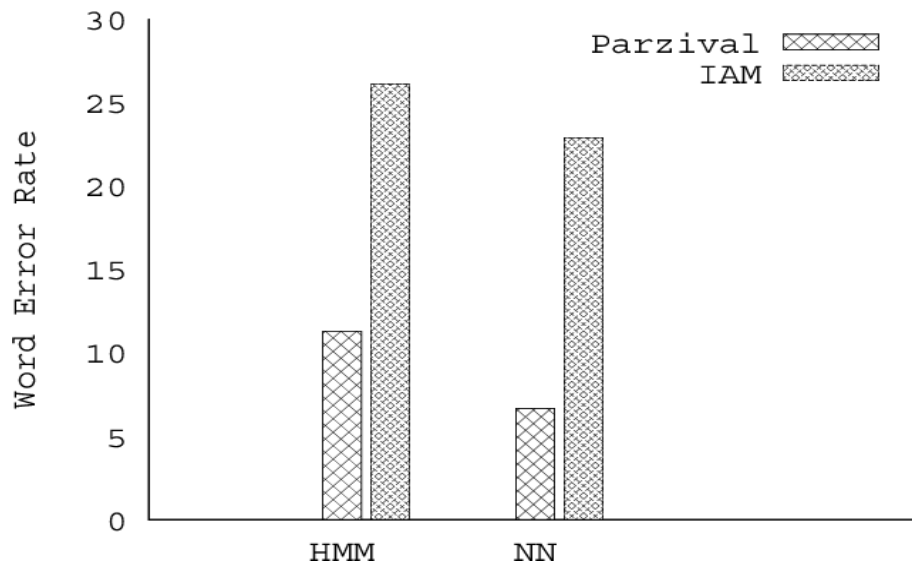
Η αρχιτεκτονική του νευρωνικού δικτύου αναφέρεται σε στοιχεία όπως ο αριθμός των διαφόρων διαστρωματώσεων στο δίκτυο, ο αριθμός των μονάδων σε αυτό καθώς και ο τρόπος με τον οποίο οι μονάδες συνδέονται μεταξύ των διαφόρων επιπέδων. Η λειτουργία είναι ανάλογη με αυτή της λειτουργίας του ανθρώπινου εγκεφάλου. Όπως οι νευρώνες μεταβιβάζουν σήματα στην περιφέρεια του εγκεφάλου, οι μονάδες (νευρώνες) λαμβάνουν ορισμένες τιμές από προηγούμενες μονάδες ως είσοδο, εκτελούν υπολογισμό και εν συνεχεία μεταβιβάζουν τη νέα τιμή ως έξοδο στις άλλες μονάδες.



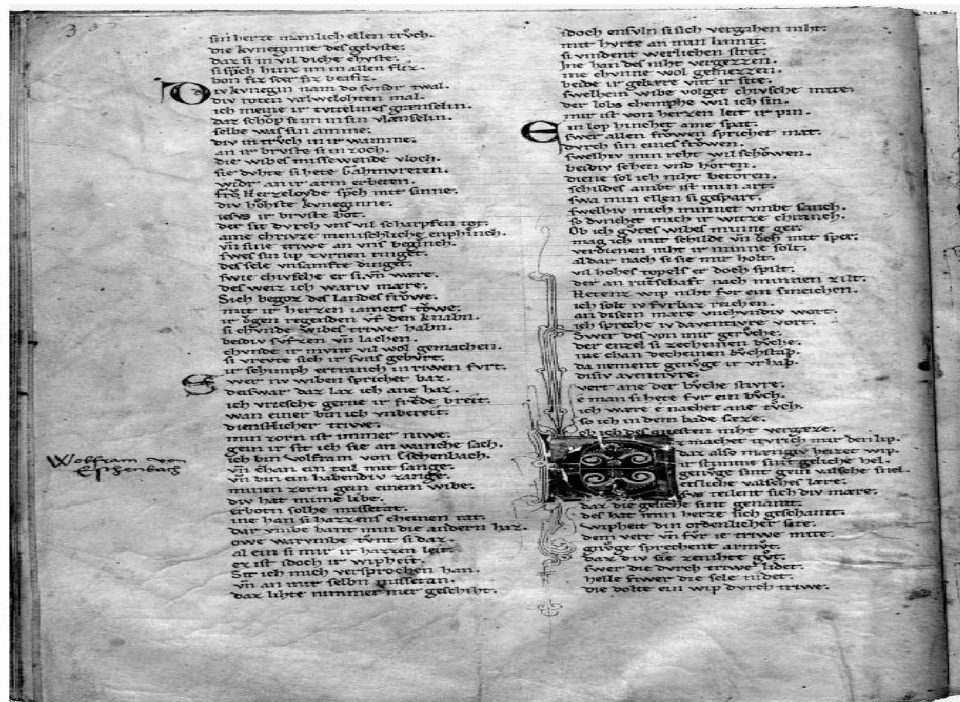
Εικ. 6. Αρχιτεκτονική νευρωνικού δικτύου.

Αυτές οι μονάδες έχουν στρώσεις για να σχηματίσουν ένα δίκτυο, ξεκινώντας με ένα τουλάχιστον επίπεδο για τιμές εξόδου. Ο όρος κρυφό στρώμα, χρησιμοποιείται για όλα τα επίπεδα μεταξύ των επιπέδων εισόδου και εξόδου.

Με βάση τα παραπάνω επιτυγχάνεται υψηλός βαθμός ακρίβειας καθιστώντας τα NN τεχνολογία αιχμής. Ειδικότερα δε τα δύο συστήματα αναγνώρισης κειμένου τα HMM και τα NN δοκιμάστηκαν σε 2 datasets όπως οι βάσεις Parzival όσο και η IAM Handwriting. Η βάση Parzival περιέχει γερμανικά κείμενα χειρόγραφα του 13^{ου} αι. σε middle high German και η IAM Handwriting περιλαμβάνει μια τεραστιά επίσης συλλογή αγγλικών χειρόγραφων κειμένων διαφόρων συγγραφέων. Το αποτέλεσμα της εργασίας των βάσεων με βάση την HMM και την NN μέθοδο ανέδειξε την ανωτερότητα της χρήσης των νευρωνικών δικτύων καθώς σε θέματα ακρίβειας και πιστότητας στη μεταγραφή και αναγνώριση κειμένου έδωσε σημαντικά καλύτερα αποτελέσματα. Στη βάση Parzival το δε ποσοστό ακρίβειας ήταν 93,32% για τη μέθοδο νευρωνικών δικτύων έναντι 88,69% των HMM και στην IAM Handwriting 77.12% και 73.91% αντίστοιχα. [FISCHER, 2009]



Εικ. 7. Αποτελέσματα ως προς το ποσοστό σφάλματος λέξεων των βάσεων Parzival και IAM.



Εικ. 8. Σελίδα από τη βάση Parzival.

1.5 ΕΥΡΕΣΗ ΛΕΞΕΩΝ ΚΛΕΙΔΙΩΝ (KWS)

Το KWS Handwriting (εύρεση λέξεων κλειδιών) αποτέλεσε μια πολλά υποσχόμενη εναλλακτική λύση για τον εντοπισμό συγκεκριμένων όρων αναζήτησης σε σαρωμένες εικόνες εγγράφων σε σχέση με τα υψηλά ποσοστά σφαλμάτων και την αργή ταχύτητα αναγνώρισης για μεταγραφή πλήρους κειμένου μη περιορισμένων εικόνων χειρογράφου.

Προηγουμένως έγινε αναφορά στην παλαιότερη μέθοδο HMM βασισμένη στη μάθηση για εντοπισμό λέξεων κλειδιών χρησιμοποιώντας μοντέλα Markov κρυμμένα με χαρακτήρες που έδειξαν υψηλή απόδοση σε σύγκριση με την παραδοσιακή αντιστοίχιση εικόνας προτύπου.

Η μέθοδος είναι αποτελεσματική σε περιπτώσεις όπου δεν απαιτείται η πλήρης μεταγραφή του κειμένου και η αναζήτηση λέξεων-κλειδιών έχει προταθεί ως εναλλακτική λύση για τον εντοπισμό όρων αναζήτησης σε σαρωμένα έγγραφα. Οι εφαρμογές μεταξύ των άλλων περιλαμβάνουν δυνατότητα επεξεργασίας μεγάλων συλλογών ιστορικών χειρογράφων προκειμένου να καταστεί δυνατή η αναζήτηση του περιεχομένου τους σε ψηφιακές βιβλιοθήκες.

Ο εντοπισμός λέξεων κλειδιών συνήθως προσεγγίζεται με μεθόδους αντιστοίχισης που συγκρίνουν εικόνες προτύπων με έγγραφα στόχου. Η προσέγγιση αυτή εμφανίζει μειονεκτήματα όπου περιλαμβάνει την ανάγκη συλλογής προτύπων για κάθε λέξη κλειδί και τη χαμηλή ικανότητα γενίκευσης για νέα στυλ γραφής που δεν είναι μέσα στις βάσεις των δεδομένων. Οι μέθοδοι βασισμένοι στη μάθηση, έχουν τη δυνατότητα ενσωμάτωσης μιας σχετικής παραλλαγής στο στυλ της γραφής. Δηλαδή τα μοντέλα μάθησης χαρακτήρων σε σύγκριση με τα μοντέλα λέξεων προσφέρουν τη δυνατότητα αναζήτησης αυθαίρετων λέξεων-κλειδιών ακόμα και στην περίπτωση που αυτά δεν εμφανίζονται στα δεδομένα εκπαίδευσης.

Ακόμα δε και στην περίπτωση μη ύπαρξης λεξικού λεξικού η μέθοδος είναι εφαρμόσιμη. Μειονέκτημα της προσέγγισης χωρίς λεξικά αποτελεί το γεγονός ότι το μοντέλο των γλωσσών χρησιμοποιήθηκε για την υποστήριξη συστήματος εντοπισμού το οποίο λαμβάνει υπόψη μόνο την εμφάνιση του κειμένου. Από παλαιότερες έρευνες τα μοντέλα γλώσσας είχαν διαφανεί πως

είναι εξαιρετικά αποτελεσματικά στην αναγνώριση χειρόγραφου κειμένου όπου αποτέλεσαν τυπικό αναπόσπαστο μέρος του συστήματος με τη μορφή LM wordn-gram. [FISCHER, 2009]

Τα βασικά στάδια αποτελούν οι προ επεξεργασία και εξαγωγή των λειτουργιών. Το σύστημα εντοπισμού λειτουργεί σε επίπεδο γραμμής κειμένου στο οποίο δεν απαιτείται η τμηματοποίηση των εικόνων της γραμμής κειμένου σε λέξεις ή χαρακτήρες. Γίνεται παράβλεψη σφαλμάτων από ανάλυση διάταξης τμηματοποίηση γραμμής κειμένου. Στη συνέχεια η προεργασία της εικόνας περιλαμβάνει binarization (δυναδικοποίηση), διόρθωση της γραμμής βάσης στο θέμα της κλίσης. [FISCHER, 2009]

Ο στόχος του εντοπισμού λέξεων-κλειδιών είναι να προσδιοριστεί ο όρος αναζήτησης εντός των εικόνων της γραμμής κειμένου. Με τον τρόπο αυτό επιτυγχάνεται με τον υπολογισμό της αναλογίας μεταξύ δύο μοντέλων γραμμής κειμένου.

ΚΕΦΑΛΑΙΟ 2° : ΣΥΓΧΡΟΝΕΣ ΕΞΕΛΙΞΕΙΣ-ΤΑΣΕΙΣ-ΜΕΘΟΔΟΙ. ΣΤΑΔΙΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΙΚΟΝΑΣ ΣΤΗ ΜΕΤΕΓΓΡΑΦΗ ΚΑΙ ΑΝΑΚΤΗΣΗ ΚΕΙΜΕΝΟΥ

2. ΣΥΓΧΡΟΝΕΣ ΕΞΕΛΙΞΕΙΣ-ΤΑΣΕΙΣ-ΜΕΘΟΔΟΙ. ΣΤΑΔΙΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΙΚΟΝΑΣ ΣΤΗ ΜΕΤΕΓΓΡΑΦΗ ΚΑΙ ΑΝΑΚΤΗΣΗ ΚΕΙΜΕΝΟΥ

Η επεξεργασία εγγράφων ιστορικής αξίας και ειδικότερα εκείνων που ανήκουν στην περίοδο του Μεσαίωνα είναι η διαδικασία ψηφιοποίησης του γραπτού λόγου και έντυπου υλικού από το παρελθόν για μελλοντική χρήση από τους ιστορικούς. Η ψηφιοποίηση των εγγράφων τα διατηρεί με ασφάλεια στην αιωνιότητα διασφαλίζοντας ότι η ψηφιακή έκδοσή του θα διασωθεί στην περίπτωση που το πρωτότυπο έγγραφο καταστραφεί.

Δεδομένου ότι μεγάλος αριθμός ιστορικών εγγράφων βρίσκονται σε Βιβλιοθήκες και αρχεία, η πρόσβαση σε αυτά παρεμποδίζεται συχνά. Ιστορικοί και μελετητές που θέλουν να μελετήσουν στα πρωτότυπα έγγραφα θα πρέπει να μετακινηθούν στο χώρο όπου βρίσκεται το πρωτότυπο έγγραφο αποθηκευμένο, ή να χρησιμοποιήσουν έκδοση μικροφίλμ ή να βρουν έντυπη μεταγραφή ή φωτογραφική τηλεμοιοτυπία.

Με την ψηφιοποίηση αυτών των ιστορικών εγγράφων, επεκτείνονται οι δυνατότητες πρόσβασης των μελετητών στις αρχειακές συλλογές ως εικόνες που ανεβαίνουν και δημοσιεύονται στο διαδίκτυο και τους επιτρέπει ακόμη και να εμπλέκουν αυτά τα κείμενα με νέους τρόπους μέσω ψηφιακών εφαρμογών. [CHANDNA 2016]

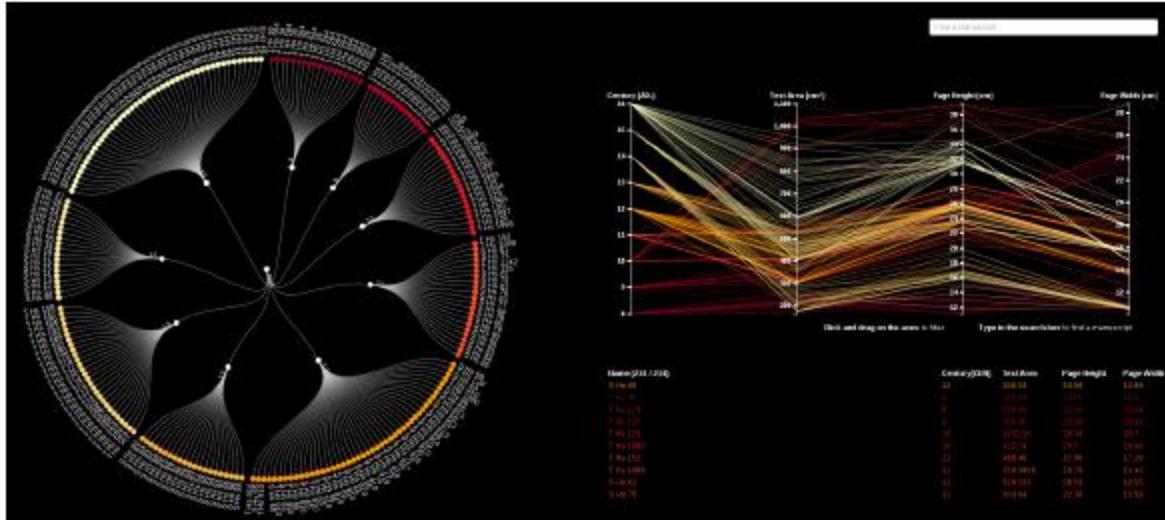
Ένα τέτοιο παράδειγμα είναι ο σχεδιασμός ενός πλαισίου οπτικοποίησης και μεγάλων συνόλων μεσαιωνικών χειρογράφων το οποίο περιλαμβάνει τρία στάδια όπως την διαχείριση δεδομένων, την εξαγωγή τους καθώς και την οπτικοποίησή τους όπως φαίνεται στα αμέσως επόμενα σχήματα. Σε αυτά τα πλαίσια λειτουργεί το λεγόμενο Codi Vis.



Εικ. 9. Περιγραφή σταδίων οπτικοποίησης Codi Vis.



Εικ. 10. Διάφορα χαρακτηριστικά διάταξης, μέγεθος εικόνας, χώρος κειμένου, χώρος εικονογράφησης



Εικ. 11. Οπτικοποίηση δεδομένων σε προβολική εικόνα δέντρου στα αριστερά περιλαμβάνοντας τον αιώνα των χειρογράφων και στη δεξιά πλευρά τα χαρακτηριστικά διάταξης.

Μέσα στις τελευταίες δύο δεκαετίες έχει γίνει τεράστια προσπάθεια για την διάσωση διατήρηση ιστορικών εγγράφων μέσω της ψηφιοποίησης. Η ανάγκη για μεταγραφή του πλήρους των κειμένων αρχειακών εγγράφων έχει αυξηθεί. Η επεξεργασία ιστορικών εγγράφων ενσωματώνει αλγόριθμους και λογισμικό και εργαλεία από διάφορα πεδία της επιστήμης των υπολογιστών για την μετατροπή εικόνων ιστορικών αρχαίων χειρογράφων σε ψηφιακή μορφή η οποία μπορεί να χρησιμοποιηθεί σε συστήματα εξόρυξης δεδομένων και ανάκτησης πληροφοριών.

Αξιοποιώντας τεχνικές εργαλεία από περιοχές συμπεριλαμβανομένης της οπτικής του υπολογιστή, της ανάλυσης και αναγνώρισης εγγράφων, της επεξεργασίας της φυσικής γλώσσας καθώς και της μηχανικής μάθησης, η επεξεργασία ιστορικών εγγράφων αποτελεί ένα νέο υβριδικό πεδίο. Με δεδομένο ότι ενσωματώνει πολλαπλούς υποτομείς της επιστήμης των υπολογιστών, η γνώση η οποία συσχετίζεται με την στόχευσή της είναι διασκορπισμένη σε δημοσιευθείσες έρευνες, άρθρα, πρακτικά συνεδρίων. Πάντως παρόλες τις έρευνες που υπάρχουν σε κάθε υποπεδίο, δεν υπάρχει ακόμα πλήρης χαρτογράφηση της εικόνας για την έρευνα στο εν λόγω πεδίο.

2.1 ΠΡΟΚΛΗΣΕΙΣ-ΤΕΧΝΙΚΕΣ

Οι υπάρχουσες μελέτες δεν αντικατοπτρίζουν πάντα τις προκλήσεις τις οποίες αντιμετωπίζουν επί του πρακτέου οι αρχειονόμοι ερευνητές και προγραμματιστές οι οποίοι εργάζονται πάνω στα project της ψηφιοποίησης και μεταγραφής ιστορικών κειμένων. Σημαντικό μέρος των προκλήσεων αυτών συνίσταται στην αντιμετώπιση προβλημάτων στην ποιότητα των πρωτότυπων εγγράφων, στην ποιότητα των ψηφιακών εικόνων και την σχετική έλλειψη σχολιασμένων δεδομένων εκπαίδευσης και δεδομένων για δοκιμές για εργαλεία μηχανικής μαθησης σε σύγκριση με την τεράστια ποσότητα δεδομένων χωρίς ετικέτα.

Τεράστια συμβολή στην αξιοποίηση και εξέλιξη της τεχνολογίας της επεξεργασίας κειμένων έδωσαν τα ιστορικά κείμενα εν συγκρίσει με τα σύγχρονα, καθώς τα πρώτα εμπεριέχουν περίπλοκες διατάξεις εγγράφων, σπάνιο λεξιλόγιο, αρχαϊκά κείμενα σε χειρόγραφα έγγραφα καθώς και τυπικές γραμματοσειρές, έντυποι σύνδεσμοι και αχνές γραμματοσειρές στην έντυπη μορφή.[SPRINGMANN 2014].



Εικ. 12. Εικόνες πρώιμων τυπωμένων βιβλίων (αριστερά: Libanius 1603, δεξιά: Curioser Botanicus 1675), κείμενα αναμειγμένα με εικόνες, ιστορικές γραμματοσειρές, κηλίδες αριστερά, κακό φόντο δεξιά.

Η ιστορική επεξεργασία εγγράφων απαιτείται ως βοήθημα για ερευνητές παρέχοντάς τους αναβαθμισμένες τεχνικές και βελτιωμένα εργαλεία απαραίτητα για επαγγελματίες επιστήμονες υπολογιστών ερευνητές και αρχειοθέτες, με σχεδιασμό και εφαρμογή συστημάτων επεξεργασίας ιστορικών κειμένων.

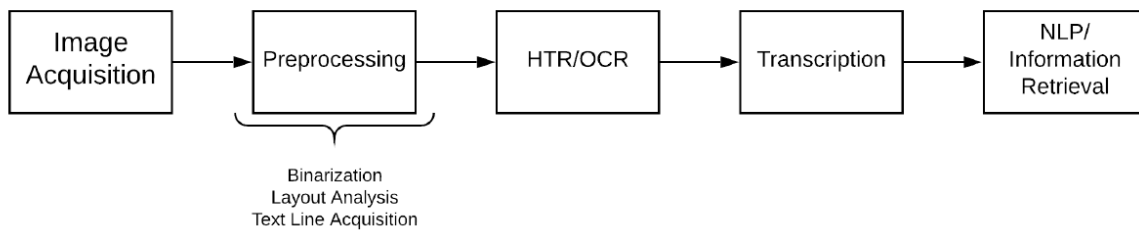
Προκειμένου να αντιμετωπιστεί η παραπάνω ανάγκη γίνεται έρευνα στις κύριες φάσεις της Επεξεργασίας ιστορικών εγγράφων σχετικά με τις τεχνικές τα εργαλεία και τις τάσεις που επικρατούν. Πολλές έρευνες αναφορές και άρθρα έχουν επικεντρωθεί σε μια διάσταση του προβλήματος, όπως την ανάλυση διάταξης, δυαδικοποίηση εικόνας, ή πραγματική μεταγραφή

Σπάνια έχει γίνει αναφορά σε μια πλήρη ροή επεξεργασίας ιστορικών εγγραφων. Επιπλέον μεγάλο μέρος της έρευνας είναι επικεντρωμένο σε διαφορετικούς αλγόριθμους και εργαλεία τα οποία είναι διασκορπισμένα σε πληθώρα δημοσιεύσεων σε διαφορετικά υποπεδία της επιστήμης των υπολογιστών.

Νέοι ερευνητές και αρχειονόμοι της πολιτιστικής κληρονομιάς θα είναι απαραίτητοι για την ανασύνθεση της υπάρχουσας εργασίας με τις νέες μορφές που επέφερε η τεχνολογική εξέλιξη. Ανασύνθεση του ευρύτερου πεδίου μέσα στο οποίο εργάζονται οι ερευνητές όπου η προσπάθειά τους θα πρέπει να συσχετισθεί και να επικεντρωθεί σε έργα ψηφιοποίησης με επιλογή των κατάλληλων εργαλείων με πλήρη κατανόηση του πεδίου στο οποίου λειτουργούν.
[PHILIPS, 2020]

2.2 ΦΑΣΕΙΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΚΕΙΜΕΝΟΥ ΚΑΙ ΕΙΚΟΝΑΣ

Η επεξεργασία ιστορικών εγγράφων ενσωματώνει στάδια σε διάφορες φάσεις. Το χειρόγραφο ή το αρχικό έντυπο βιβλίο ψηφιοποιούνται . Εν συνεχεία, μόλις απεικονιστεί το έγγραφο, οι σαρώσεις σελίδων συνήθως ακολουθούν την εξής προ επεξεργασία.



Εικ. 13. Διάγραμμα επεξήγησης επεξεργασίας εικόνας.

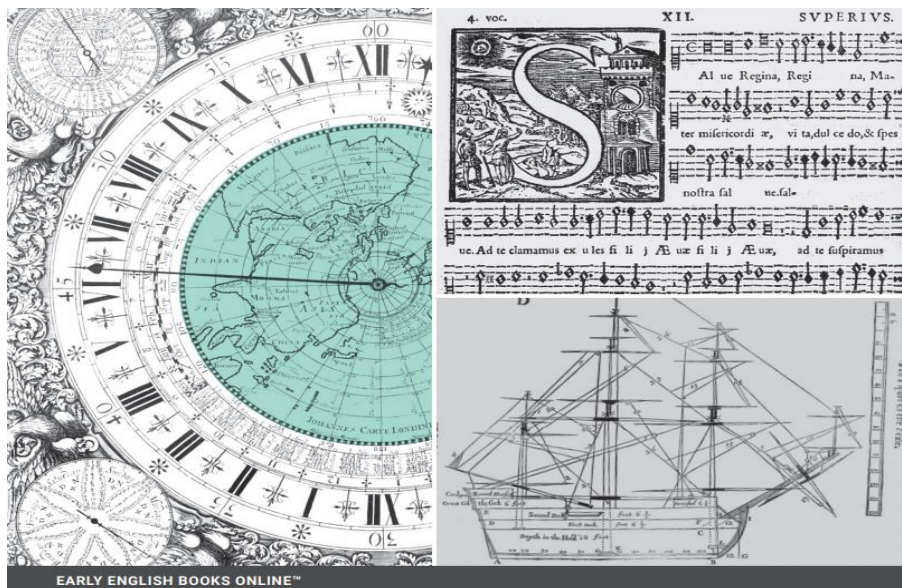
Έχοντας τελειώσει με το στάδιο προ επεξεργασίας, είτε μέσω οπτικής αναγνώρισης χαρακτήρων OCR, είτε χειρόγραφης αναγνώρισης κειμένου HTR, ανάλογα με τον τύπο του αρχείου, χρησιμοποιείται μηχανική μάθηση και χρήση λογισμικού αναγνώρισης. Αυτό το στάδιο μεταγραφής, ανεξάρτητα από το είτε προσπαθεί να διεξαγάγει μεταγραφή μία προς μία λέξη με εντοπισμό λέξεων κλειδιών, είτε ολόκληρου του εγγράφου, αντλεί την ποιότητα των αποτελεσμάτων του από τον συνδυασμό της ποιότητας που έχουν εξαχθεί από τα δεδομένα εκπαίδευσης και από τα προερχόμενα από αυτά μοντέλα.

Πρώιμες προσπάθειες διατήρησης και διάσωσης ιστορικών εγγράφων αρχικά έγιναν με τη μέθοδο του μικροφίλμ και μετά μετατράπηκαν σε ψηφιακά αρχεία με στόχο να επιτευχθεί η καλύτερη δυνατή ανάλυση εικόνας του εγγράφου με χρήση μηχανημάτων και εργαλείων ακριβείας με δυνατότητες αυτόματης ανάλυσης διάταξης και μεταγραφής όπως φαίνεται πιο κάτω στην εικόνα 15. [BAECHLER 2010]



Εικ. 14. Σχολιασμένα μεσαιωνικά χειρόγραφα όπου το επίπεδο κειμένου τονίζεται με μπλε, των σχολίων με πράσινο και το διακοσμητικό με πορτοκαλί.

Ως μία από τις πλέον χαρακτηριστικές προσπάθειες διάσωσης πρώιμων ιστορικών εγγράφων το Early English Books Online Project το οποίο περιλαμβάνει υλικό 150 βιβλιοθηκών που αφορούν την ιστορική περίοδο της Αγγλίας μεταξύ του 1475-1700. Σε συνεργασία με την Text Creation Partnership δημιουργήθηκαν χιλιάδες, ακριβείς, με δυνατότητα πλήρους αναζήτησης μεταγραφές πλήρους κειμένου πρώιμων έντυπων βιβλίων διαθέσιμες σε όλους στο διαδίκτυο.



Εικ. 15. Early English Books Online

2.2.1 ΕΙΔΗ ΑΡΧΕΙΑΚΩΝ ΕΓΓΡΑΦΩΝ-ΠΡΟΒΛΗΜΑΤΑ-ΠΡΟΚΛΗΣΕΙΣ

Η πλειονότητα των ιστορικών εγγράφων αποτελούνταν από χειρόγραφα. Αυτό άλλαξε και διαφοροποιήθηκε με την έλευση της τεχνολογίας της τυπογραφίας ήδη από τον 15^ο αι. χάρη στον Γουτεμβέργιο. Έτσι προέκυψαν διαφορετικοί τύποι καθώς είχαμε πλέον πέρα από χειρόγραφα και έργα που παράγονταν στα τυπογραφικά πιεστήρια και μέρος της ιδιωτικής αλληλογραφίας προκειμένου να διατηρηθεί πέρασε σε αρχεία ήχου. Ως αποτέλεσμα αυτής της διαφοροποίησης των τύπων των ντοκουμέντων ως προς την δημιουργία και τη διατήρησή τους έχει οδηγήσει σε επίσης διαφορετικούς τρόπους αντιμετώπισης κατά τη διαδικασία της ψηφιοποίησης.

Ειδικότερα, ένα μεσαιωνικό χειρόγραφο πιθανότατα έχει πιο περίπλοκη διάταξη από ένα άλλο το οποίο συντάχθηκε τον 17^ο αιώνα αλλά επίσης η μεταγραφή του δεύτερου να είναι πιο δύσκολη σε θέματα ακρίβειας από το πρώτο για μια σειρά από παράγοντες.

Τα μεσαιωνικά ιστορικά έγγραφα παράγονταν συνήθως σε μοναστήρια τα οποία ονομάζονταν και scriptoria. Επίσης σημαντικός παράγοντας είναι και το υλικό πάνω στο οποίο δημιουργήθηκαν τα χειρόγραφα. Ιδίως πριν από την εποχή της τυπογραφίας αυτά γράφονταν

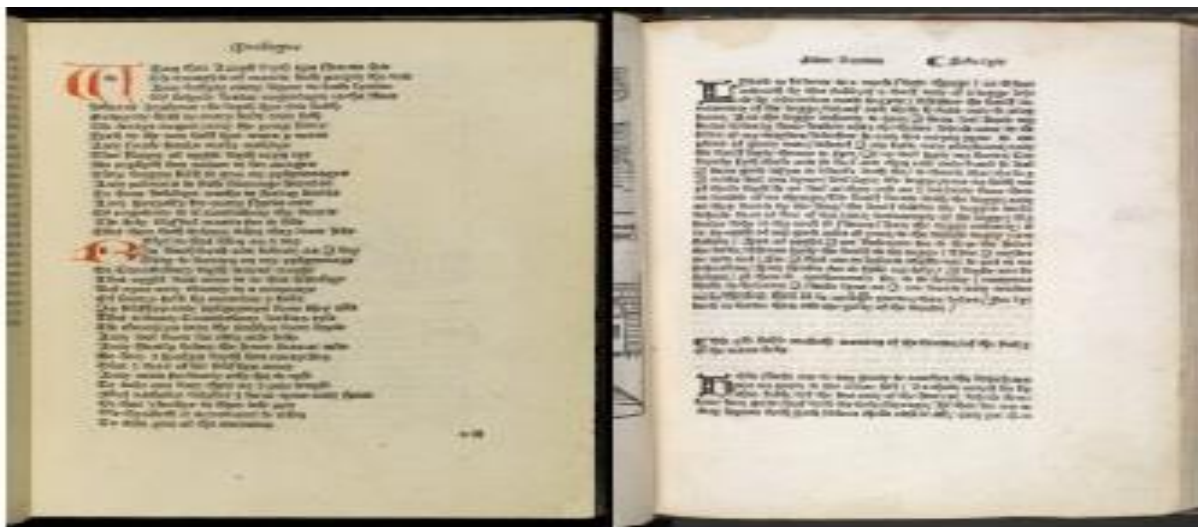
πάνω σε πάπυρο ή περγαμηνή, υλικό από δέρματα ζώων. Παρότι πληθώρα αποσπασματικών παπύρων πολύ σημαντικών ιστορικών ντοκουμέντων επιβιώνει ως τις μέρες μας, ένα άλλο είδος υλικού το πέπλο αποτέλεσε το κύριο μέσο επιλογής για χειρόγραφα έγγραφα κατά τη μεσαιωνική περίοδο. Στα πλαίσια αυτά, έχει γίνει σημαντική προσπάθεια στην συγκέντρωση και παραγωγή αλγόριθμων και βάσεων δεδομένων για αποτελεσματική αντιμετώπιση των προκλήσεων χειρογράφων από διαφορετικό υλικό σε λογισμικά αυτόματα εργαλεία μεταγραφής. Για παράδειγμα η παραγωγή υλικού που είναι σε πάπυρους που το υλικό έχει φθαρεί από το χρόνο δεν αποδίδει ιδιαίτερα καλά αποτελέσματα στον τομέα της αυτόματης μεταγραφής. [PHILIPS, 2020]



Εικ. 16. Αποσπάσματα στοιχείων ελληνικών βιβλικών παπύρων (αριστερά) και μεσαιωνικά λατινικά έγγραφα από το e-codices με απεικόνιση σύνθετης διάταξης

Σε αντίθεση με το παραπάνω, ιστορικά έγγραφα και κώδικες από την ύστερη αρχαιότητα και τον Μεσαίωνα τα οποία γραφτήκαν σε πέπλο τείνουν στο να είναι καλύτερα διατηρημένα από το υλικό του πάπυρου και ως αποτέλεσμα οι αυτόματες μεταγραφές να είναι πολύ υψηλότερης αξιοπιστίας. Επίσης τα γραφέντα είναι πιο ευανάγνωστα και η διάταξη μεταξύ των χαρακτήρων των γραμμάτων επίσης διευκολύνει την μάθηση και εκπαίδευση μέσω της μηχανικής μάθησης βασιζόμενα σε ταξινομητές σε σύγκριση με τα αρχικά σύγχρονα τυπωμένα και επίσης σύγχρονα χειρόγραφα κείμενα. Μια από τις κυριότερες προκλήσεις των μεσαιωνικών εγγράφων αποτελεί

η πολύπλοκη διάταξή τους που δεν επιτρέπει την διάκριση τόσο των γραμμών όσο και των χαρακτήρων για να γίνει ευχερέστερη η διαδικασία της αναγνώρισης κειμένου. [SIMISTIRA F., 2016]



Εικ. 17. Το *incunabula* στα τέλη του 15^{ου} αιώνα όπως έχει σχεδιαστεί για να αναπαράγει τα μεσαιωνικά χειρόγραφα σε διατάξεις και τα διακοσμητικά σε έντυπη μορφή.

Σύνηθες γεγονός αποτελούσε και επιπλέον δυσκολία ότι πέρα από το πρωτεύον κείμενο αυτού του είδους τα χειρόγραφα συχνά διέθεταν και εκτεταμένο περιθωριακό δευτερεύον συμπληρωματικό κείμενο στα περιθώρια ή σημειώσεις τροποποιήσεις του πρωτεύοντος κειμένου που για να μπορέσει να ταυτοποιηθεί και να μεταγραφεί έπρεπε να προηγηθεί μια σωστή κατάτμηση του κειμένου.

Μια άλλη παράμετρος σημαντική είναι ότι αυτού του τύπου τα αρχεία συχνά εμπεριείχαν και εικονογραφήσεις και στολίδια συχνά απαντούμενα σε θρησκευτικά έργα τέχνης και όχι μόνο. Προκειμένου λοιπόν αυτού του τύπου τα αρχεία να τύχουν σωστής επεξεργασίας και να μεταγραφούν κατά το δυνατόν σωστά, θα πρέπει να αντιμετωπιστούν κατά τη φάση της προεπεξεργασίας σύμφωνα με το HDP cycle.

Η εξέλιξη χειρόγραφων ιστορικών ντοκουμέντων που κατά την πρώιμη φάση του Μεσαίωνα κατά κύριο λόγο ήταν για ιδιωτική χρήση και σε περιορισμένη κυκλοφορία όπως π.χ προσωπική αλληλογραφία, βιβλία, ημερολογια. Η εξέλιξη των χειρόγραφων και η ιδιωτικότητα ακόμα και

για άτομα που έζησαν σε επικαλυπτόμενες χρονικές περιόδους αποδεικνύουν την πολυπλοκότητα του HDP για αυτά τα μετα μεσαιωνικά ντοκουμέντα. Λόγω της συνεχούς καμπυλότητας και καλλιγραφίας των χαρακτήρων των γραμμάτων η αντιμετώπισή τους αποτελεί πρόκληση κατά τη φάση HTR, ενώ υπάρχουν και μεσαιωνικά έγγραφα με ακόμα μεγαλύτερη πρόκληση την κατάτμηση του κειμένου και την ανάλυση της διάταξης.

Συνήθως τα έγγραφα της πρώιμης εποχής του μεσαίωνα πρέπει να είναι απομονωμένα από οριακούς ή διαγραμμικά περιθωριακά κείμενα ή δυσανάγνωστες ενότητες προκειμένου να έχουμε μια όσο το δυνατόν πιο ακριβή μεταγραφή.

Ανεξάρτητα από την εποχή στην οποία αυτά δημιουργήθηκαν, παρουσιάζονται και άλλες προκλήσεις για τη διαδικασία HDP, όπως του λεγόμενου «ποτίσματος» του υλικού με τη μελάνι και στην πίσω πλευρά της σελίδας όπου υπάρχει επίσης χειρόγραφο υλικό, διφορούμενη ή δυσανάγνωστη γραφή λόγω πιθανής κακής κατάστασης της ποιότητας της εικόνας κατά τη διαδικασία ίσως της απεικόνισής του, με χρήση προστατευτικών πλαστικών φύλλων τα οποία λόγω ανακλαστικότητας του φωτός να προκαλέσουν θαμπάδα ή πολύ έντονο φωτισμό στην εικόνα του εγγράφου. Τα παραπάνω αποτελούν παράγοντες που μπορούν να αποτελέσουν σημαντικό πρόβλημα στη διαδικασία της ψηφιοποίησης.

Επιπλέον έγγραφα τα οποία ήδη έχουν υποστεί μια αρχική επεξεργασία στη μορφή του μικροφίλμ στη συνέχεια απεικονίστηκαν και ψηφιακά. Η διαδικασία της νέας επεξεργασίας υλικού ήδη επεξεργασμένου από το πρωτότυπο σε ψηφιακή εικόνα μειώνει την ποιότητα που αποδίδεται. [PHILIPS, 2020] Ιδίως τα χειρόγραφα εκείνα της περιόδου που τυπώθηκαν πριν το 1500 στα αρχικά στάδια της εφαρμογής της τυπογραφίας έχουν δημιουργήσει τις περισσότερες δυσκολίες για ακριβή ψηφιακή μεταγραφή.

Με δεδομένο ότι τα βιβλία αυτά είναι στοιχειοθετημένα σε γραμματοσειρές οι οποίες διαφέρουν πολύ σε σχέση με τις σύγχρονες, το σύγχρονο λογισμικό οπτικής αναγνώρισης κειμένου OCR δεν παράγει ευανάγνωστα αποτελέσματα αναγνώρισης. Η εκτεταμένη χρήση συνδέσμων στο κείμενο επίσης αποτελεί έναν ακόμα παράγοντα συσκόλισης καθώς αυτοί μειώθηκαν στη χρήση καθώς η διαδικασία της εκτύπωσης γινόταν πιο πρότυπη. Μετά το 1500

επήλθε μεγαλύτερη ομοιομορφία σε εκτυπωμένα βιβλία και κατά τις αρχές του 19^{ου} αι. η μαζική παραγωγή έντυπου κειμένου οδήγησε σε βιβλία με πιο μοντέρνη διάταξη κειμένου γραμμάτων και σειρών και ως αποτέλεσμα τα εργαλεία οπτικής αναγνώρισης OCR μπορούσαν με μεγαλύτερη ευχέρεια και αξιοπιστία να ψηφιοποιήσουν σε μεγάλη κλίμακα υλικό όπως έχει αποδοθεί στα Internet Archive και στα Google Books με τη συνέργεια μεγάλων Βιβλιοθηκών. [PHILIPS, 2020]

2.3 ΑΝΑΓΝΩΡΙΣΗ ΧΕΙΡΟΓΡΑΦΩΝ-ΕΠΕΞΕΡΓΑΣΙΑ

Παρά τις σημαντικές και διαφαινόμενες διαφορές μεταξύ ιστορικών χειρογράφων και έντυπων, η μέθοδος του οπτικού χαρακτήρα αναγνώρισης αλλά και αναγνώρισης της γραφής είναι οι ουσιαστικές λύσεις στο ίδιο πρόβλημα δηλαδή το πώς θα δοθεί λύση στην αξιόπιστη εξαγωγή του κειμένου.

Μετά τη φάση της προ-επεξεργασίας, η αναγνώριση κειμένου προσδιορίζει είτε λέξεις κλειδιά, είτε δημιουργεί μια λέξη προς λέξη μεταγραφή από μια γραμμή του κειμένου. Ανεξάρτητα λοιπόν από το εφόσον το κείμενο είναι χειρόγραφο ή έντυπο, βασικός στόχος είναι η μετατροπή των λέξεων στην εικόνα που είναι απεικονισμένο το έγγραφο σε ψηφιακή μορφή με ακρίβεια.

Η οπτική αναγνώριση κειμένου βασίζεται πάνω στην προβλέψιμη ακολουθία του διαστήματος μεταξύ χαρακτήρων και λέξεων ως βάση για την ταξινόμησή του. Ο χαρακτήρας του γράμματος είναι το πρώτο και βασικό θεμελιώδες στοιχείο αναγνώρισης. Με δεδομένο ότι οι λέξεις και οι χαρακτήρες οι οποίοι την απαρτίζουν μπορεί να διέπονται από συγκεκριμένες προβλέψιμες και με ακρίβεια τμηματοποιημένους και επαναλαμβανόμενους, οι ταξινομητές αναγνώρισης οπτικών χαρακτήρων μπορούν αφενός να αναγνωρίσουν και αφετέρου να δημιουργήσουν μια μεταγραφή χρησιμοποιώντας το ιδιαίτερο γραφιστικό χαρακτήρα του κάθε γράμματος.

Από την άλλη η αναγνώριση χειρογράφου δεν μπορεί να βασιστεί μονάχα στην τακτική απόσταση χαρακτήρων και λέξεων λόγω ιδίως των ιδιοτεροτήτων και ιδιοσυγκρασιών της ανθρώπινης γραφής. Αν πάρουμε ως βάση το παράδοξο του Sayre, μεμονωμένα γράμματα δεν είναι δυνατό να αναγνωριστούν δίχως τη διαδικασία της τμηματοποίησης και την ίδια στιγμή η πράξη της τμηματοποίησης δεν μπορεί να προχωρήσει δίχως προηγούμενη αναγνώριση. Σε

κάθε περίπτωση η αναγνώριση χειρογράφου βασίζεται συνήθως στη μεθοδολογία αναγνώρισης χωρίς τμηματοποίηση σε επίπεδο χαρακτήρων. Προκειμένου να βελτιωθεί η ακρίβεια τόσο της οπτικής αναγνώρισης χαρακτήρων όσο και της χειρόγραφης αναγνώρισης κειμένου, συχνά τα συστήματα ενσωματώνουν ένα λεγόμενο μοντέλο στατιστικής γλώσσας ιδίως δε στην περίπτωση που αυτή είναι γνωστή εκ των προτέρων. Αυτό το γλωσσικό μοντέλο είναι απαραίτητο για την διασφάλιση ότι οι λέξεις οι οποίες αναγνωρίζονται από το λογισμικό, αντιστοιχούν σε γνωστό λεξιλόγιο και αντίστοιχη γραμματική στην καθομιλουμένη. Οι τεχνικές για αναγνώριση χειρογράφου και οπτική αναγνώριση χαρακτήρων μπορούν να συνδυαστούν και να χρησιμεύσουν σε προσεγγίσεις μηχανικής μάθησης που βασίζονται στη χρήση νευρωνικών δικτύων.

2.3.1 Στάδιο προεπεξεργασίας

Κατά τη διάρκεια μιας τυπικής ροής επεξεργασίας ιστορικού εγγράφου, την φάση προεπεξεργασίας ακολουθεί η απόκτηση της εικόνας και προηγείται κάθε απόπειρας για μεταγραφής του περιεχομένου του κειμένου του εγγράφου. Σε αυτό το στάδιο είναι διακριτή η δυαδικοποίηση όπου στο κατώτατο όριο εφαρμόζεται στην εικόνα του εγγράφου η όποια δυνατότητα για προσαρμογή για λοξή, ανάλυση διάταξης και τμηματοποίηση της γραμμής κειμένου. Από μελέτες που έχουν γίνει έχουν προταθεί διάφορες μέθοδες δυαδικοποίησης όπως οι Dewarping και skew reduction [BUKHARI 2011]. Μια από τις πιο απαιτητικές πτυχές επεξεργασίας ιστορικών εγγράφων είναι η ανάλυση διάταξης κειμένου. Πρόκειται για μια ιδιαίτερος δύσκολη πρόκληση που αφορά μεσαιωνικά έγγραφα, το οποίο δημιουργείται λόγω της πολυπλοκότητας της διάταξης των σελίδων τους.

Για την αντιμετώπιση της ανάλυσης διάταξης έγιναν διάφορες μελέτες και κατά το παρελθόν στις αρχές της πρώτης δεκαετίας του 21^{ου} αι. συζητήθηκαν προτάθηκαν ημιαυτόματα εργαλεία και μεθοδολογίες για ανάλυση διάταξης. Αυτά τα εργαλεία απαιτούσαν την αλληλεπίδραση των χρηστών για την αναγνώριση ή σχολία περιοχής εγγράφων. Βασισμένα σε pixel ή διασυνδεδεμένα στοιχεία εικόνας που χρησιμοποιήθηκαν από τα πιο πάνω εργαλεία για τον προσδιορισμό των περιοχών των εγγράφων.

Στη μελέτη των Baechler Ingold είχε προταθεί μοντέλο διάταξης για μεσαιωνικά έγγραφα το οποίο θα μπορούσε να υποστηρίξει απόλυτα αυτόματα εργαλεία σχολιασμού και μεταγραφής μεσαιωνικών κειμένων. Χρησιμοποιώντας εικόνες χειρογράφων από το πρόγραμμα e-codices του Πανεπιστημίου του Freiburg μοντελοποίησαν μια σελίδα σε ένα μεσαιωνικό χειρόγραφο βασισμένα σε διάφορα επίπεδα διαστρωμάτωσης ανάλυσης διάταξης, περιλαμβάνοντας το κείμενο του εγγράφου, τα σχόλια στο περιθώριο, την διακόσμηση του αρχείου αλλά και την φθορά και υποβάθμιση της ποιότητας από το υλικό με την όποια επίπτωση στην αναγνωσιμότητα του χειρογράφου. Αλληλεπικαλυπτόμενα πολυγωνικά κουτιά χρησιμοποιούνται για να τύχουν αναγνώρισης τα συστατικά της ανάλυσης διάταξης τα οποία και εμφανίζονται σε λογισμικό σε format XHTML.

Σε μελέτη του Γάτου κλπ κατά το 2014 αναπτύχθηκε λογισμικό διάταξης ανάλυσης και τμηματοποίησης γραμμών σειράς το οποίο σχεδιάστηκε για να δημιουργεί είσοδο δεδομένων σε εργαλεία χειρόγραφης αναγνώρισης κειμένου (HTR). Εν συνεχεία τα αποτελέσματα της δουλειάς τους ενσωματώθηκαν στην πλατφόρμα Transcribus. [PHILIPS, 2020]

Η λειτουργία του εργαλείου τους πρώτα προσδιορίζει τις ζώνες του κειμένου χρησιμοποιώντας έναν συνδυασμό οποιονδήποτε κατακόρυφων γραμμών στο έγγραφο, κατακόρυφα τμήματα λευκού χώρου και στο τέλος οριζόντια βελτίωση. Αυτός ο συνδυασμός δίνει τη δυνατότητα ανίχνευσης/ενοτήτων εγγράφων διαχωρισμένες με γραμμές, διάταξη πολλών στηλών εγγράφων διαχωρισμένων από διαδρομές λευκού χώρου και τη χρήση οριζοντίων γραμμών για ακριβέστερη αναγνώριση των ζωνών του κυρίως κειμένου. Οι γραμμές κειμένου προσδιορίζονται σε κάθε περιοχή του αναλογικά με τον υπολογισμό του μέσου ύψους των χαρακτήρων. Με βάση αυτή τη μεθοδολογία και χρήση δείγματος εικόνων ιστορικών εγγράφων από το Transcriptorium, δημιουργήθηκε μια βάση δεδομένων αναφοράς για να αξιολογηθεί η απόδοση και η ποιότητα του εργαλείου ανάλυσης διάταξης με ποσοστό ακρίβειας 84,7% στην αναγνώριση ζωνών κειμένου σε πρωτογενές υλικό. Έγινε αξιολόγηση της κατ' ακρίβεια προσέγγισης τμηματοποίησης της γραμμής κειμένου στο Transcriptorium με ποσοστά ανίχνευσης 83,08%, ακρίβεια αναγνώρισης 86,35% και με βάση το μέτρο F 84,68%,

επιτυγχάνοντας εξαιρετικά ποσοστά και ξεπερνώντας κατά πολύ προηγούμενες προτεινόμενες μεθοδολογίες.

Σε άλλες έρευνες των Pintus Rushmeier και Yang γίνεται επίσης διερεύνηση της ανάλυσης διάταξης και εξαγωγής γραμμής κειμένου δίνοντας έμφαση σε μεσαιωνικά χειρόγραφα. Γίνεται ιδιαίτερη μνεία στην ανάγκη για δημιουργία συστήματος αυτόματης ανάληψης διάταξης.

Αρχικά το πρόβλημα εντοπίζεται στον αρχικό υπολογισμό του ύψους της γραμμής του κειμένου. Υπογραμμίζεται η διαδικασία εκτίμησης υπολογισμού της αναγκαίας βασικής μέτρησης η οποία επιδεινώνεται από την επικάλυψη κατερχόμενων και ανερχόμενων από διπλανές γραμμές κειμένου με στενή απόσταση μεταξύ των γραμμών πράγμα σύνηθες σε μεσαιωνικά χειρόγραφα.

Πρόταση των Pintus Rushmeier και Yang είναι η δημιουργία αλγόριθμου αυτόματης εξαγωγής γραμμής κλειμένου το οποίο υπολογίζει το ύψος γραμμής ανά σελίδα χειρογράφου. Με την συγκεκριμένη μέθοδο υπολογίζεται το ύψος του κειμένου δημιουργώντας μια πολυδιάστατη κλίμακα παρουσίας της εικόνας του κειμένου όπου παράγεται μια συλλογή από υποσύνολο εικόνων. Κάθε ένα από αυτά τίθεται σε επεξεργασία με τη χρήση του λεγόμενου NACF, προφίλ προβολής, διακριτικό μετασχηματισμό Fourier με τα σχετικά αποτελέσματα αυτού του PMF να συνοζίζονται στον υπολογισμό της εκτίμησης της γραμμής.

Με τη χρήση αυτής της τιμής στη συνέχεια τμηματοποιούνται οι περιοχές του κειμένου χοντρικώς και εφαρμόζοντας ένα SVM έναν ταξινομητή για την παραγωγή μιας πιο εκλεπτυσμένης γραμμής αναγνώρισης κειμένου. Σύμφωνα με την μέθοδο δεν επηρεάζεται η αξιοπιστία της από χειρόγραφα που είναι σε λοξή γραφή και συνήθως δεν απαιτούν διόρθωση ευθυγράμμισης. Σε δοκιμές που έγιναν με την παραπάνω προτεινόμενη μέθοδο σε σύνολα δεδομένων άνω των 15.552 σελίδων και 80.000 γραμμών κειμένου τα αποτελέσματα ήταν εντυπωσιακά της τάξης του 98,55% σε μέση ακρίβεια και 96,31% για μέση ανάκληση τμηματοποίησης γραμμής κειμένου. Επίσης δώθηκαν και μετρήσεις χρόνου για το απαιτούμενο χρονικό διάστημα εκτέλεσης των παραπάνω εργασιών των αλγορίθμων. Ενδεικτικά μέσος όρος

επεξεργασίας για την εκτέλεση πολλαπλών νημάτων για κάθε σελίδα ανέρχεται σε 3 λεπτά. [PHILIPS, 2020]

Σε άλλη πρόταση των Yang et al γίνεται επέκταση στο έργο τους σχετικά με την εκτίμηση ύψους κειμένου και ανάλυση διάταξης σε ένα αυτοματοποιημένο σύστημα το οποίο μπορεί να λειτουργήσει ανά σελίδα και όχι ανά χειρόγραφο. Προτείνονται τρεις αλγόριθμοι, ο ένας για την εξαγωγή γραμμής κειμένου, ένας για εξαγωγή μπλοκ κειμένου και ένας για εξακρίβωση ειδικών στοιχείων.

Η χρήση τους γίνεται με ημι-εποπτευόμενη τεχνική μηχανικής μάθησης και εστιάζουν σε μεσαιωνικά χειρόγραφα τα οποία δημιουργήθηκαν από επαγγελματίες γραφείς. Ειδικότερα σημειώνεται πως οι αλγόριθμοί τους είναι ειδικά σχεδιασμένοι για να αντιμετωπίσουν παραμορφωμένες σελίδες και σοβαρά κατεστραμμένα έγγραφα.

Ενδεικτικά με την μέθοδο των Yang et al [PHILIPS, 2020] στον υπολογισμό μπλοκ κειμένου επιτυγχάνεται ακρίβεια 97%, ανάκληση στο 96% από σύνολο 35 εικόνων που περιέχονται σε 42 μπλοκ κειμένου. Αναφερόμενο μειονέκτημα της παραπάνω μεθόδου είναι ότι δεν αξιολογούν τους αλγόριθμους τους έναντι του συνόλου δεδομένων Simistira et al ή των αποτελεσμάτων του Gatos et al. [PHILIPS, 2020]. Παρόλα ταύτα πάντως τα αποτελέσματά τους κρίνονται ιδιαίτερα αποτελεσματικά και υποδηλώνουν ότι η αυτόματη ανάλυση αλγοριθμικής διάταξης παρουσιάζει υψηλή ακρίβεια, και η ανάκληση και ακρίβεια προσεγγίζουν εγγύτερα στην πραγματικότητα του κειμένου.

2.4 ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΧΕΙΡΟΓΡΑΦΟΥ ΚΕΙΜΕΝΟΥ

Οι κυρίαρχες σύγχρονες τάσεις και τεχνικές στην αναγνώριση κειμένου είναι εκείνες που χρησιμοποιούν την παραδοσιακή μηχανική μάθηση machine learning και τη βαθιά μάθηση deep learning και τα οποία είναι αντικείμενο της σχετικά πρόσφατης βιβλιογραφίας και επιστημονικής έρευνας. Με δεδομένες της προκλήσεις στη χειρόγραφη αναγνώριση ιστορικών κειμένων έχουν υπάρξει μελέτες οι οποίες διερεύνησαν τεχνικές εντοπισμού λέξεων κλειδιών

κως ως μια εναλλακτική λύση για παραγωγή μιας πλήρους μεταγραφής. Οι αρχικές τεχνικές αναγνώρισης κειμένου εφαρμόστηκαν σε πρώιμα ιστορικά ντοκουμένα ως πρόβλημα ομοιότητας της εικόνας. Σε αυτή την μέθοδο, συστάδες εικόνων από λέξεις δημιουργούνται οι οποίες έχουν συγκριθεί για ομοιότητα με χρήση ζεύγους λέξεων από απόσταση. Αυτές οι συστάδες οι οποίες αντιστοιχούν σε σημαντικές λέξεις εντός του εγγράφου εν συνεχεία επισημαίνονται χειροκίνητα με τη λέξη τους. Οι ετικέτες επί των λέξεων μπορούν να καταχωρηθούν να ευρετηριαστούν και να υπάρχει δυνατότητα εντοπισμού σε ένα σύστημα ανάκτησης πληροφοριών. Πρόκειται για την προσέγγιση βάση προτύπου για τον εντοπισμό λέξεων κλειδιών .

Μέσω μιας μελέτης των Rath και Manmatha έγινε χρήση ιδιαίτερου αλγόριθμου στρέβλωσης του χρόνου προκειμένου να γίνει υπολογισμός της ομοιότητας της εικόνας και σύγκρισή της με την απόδοση διαφόρων αλγορίθμων ομαδοποίησης περιλαμβάνοντας Ward linkage και K-means σε μια αρχική έκδοση της βάσης του G. Washington.

Ως κομμάτι άλλου πρόγραμματος στο HisDoc του Fischer κα έγινε διερεύνηση διάφορων τεχνικών και για εντοπισμό λέξεων κλειδιών και για ολοκληρωμένη μεταγραφή. Σημαντική παράμετρος προβληματισμού είναι ότι στη βασιζόμενη αντιστοίχιση βάσει προτύπου λέξεων, το σύστημα έχει τη δυνατότητα αναγνώρισης μόνο για λέξεις για τις οποίες έχει εικόνες αναφοράς. Όσες λέξεις είναι εκτός της βάσης δεν μπορούν να αναγνωριστούν. Για να επιλυθεί αυτός ο περιορισμός οι ερευνητές δοκίμασαν την αναγνώριση βάση των μοντέλων HMM για εντοπισμό λέξεων κλειδιών. Σε αυτή τη μέθοδο γίνεται προσέγγιση η οποία βασίζεται στη μάθηση και η οποία εφαρμόζεται στο επίπεδο χαρακτήρων, ένα σύστημα εντοπισμού λέξεων έχει την ικανότητα εντοπισμού αυθαίρετων λέξεων κλειδιών και συνδυάζουν κατάλληλα τα μοντέλα των γραφικών χαρακτήρων. Η ανάλυση εντοπισμού λέξεων κλειδιών συγκρίνεται με το σύστημα που βασίζεται σε χαρακτήρες με βασικό δυναμικό σύστημα αλγόριθμου στρέβλωσης του χρόνου.

Με βάση τη χρήση της μέσης ακρίβειας ως μέτρηση αξιολόγησης έγινε διαπίστωση πως τα κρυφά μοντέλα Markov ξεπέρασαν σημαντικά τον δυναμικό χρόνο διαστρέβλωσης του

συστήματος τόσο σε τοπικά όσο και παγκόσμια όρια για τις βάσεις δεδομένων του G. Washington και Parzival.

Κάνοντας σύγκριση στο project HisDoc με τα μοντέλα HMM και την απόδοση του νευρωνικού δικτύου σε τρεις βάσεις δεδομένων που αποτελούν το IAM-HisDB όπως τα St Gall, G. Washington, Parzival προκειμένου να παράγουν ολοκληρωμένες μεταγραφές. Τα HMM ήταν ανάλογο με αυτό που επεξηγήθηκε πιο πριν στην μελέτη εντοπισμού λέξεων κλειδιών, για το σύστημα βασιζόμενο σε νευρωνικά δίκτυα, χρησιμοποιείται μια αμφίδρομη αρχιτεκτονική βραχυπρόθεσμης μνήμης η οποία προσπαθεί να μετριάσει το πρόβλημα διαγραφής σχεδίων άλλων νευρωνικών δικτύων. Κάθε ένα από τα γεωμετρικά σχήματα που χρησιμοποιούνται για εκπαίδευση αντιστοιχούν σε μεμονωμένο κόμβο στο επίπεδο εισόδου του δικτύου. Οι κόμβοι εξόδου στο δίκτυο αντιστοιχούν στους μεμονωμένους χαρακτήρες στο σύνολο των χαρακτήρων. Η πιθανότητα μιας λέξης υπολογίζεται με βάση τις πιθανότητες των χαρακτήρων.

Αποτελέσματα μελετών έδωσαν ποσοστά καλύτερης απόδοσης σε ποσοστά σφάλματος για τεχνικές νευρωνικών αρχιτεκτονικών δικτύων σε όλες τις περιπτώσεις απέναντι σε σύστημα που βασίζεται στα κρυφά μοντέλα Markov. Ενδεικτικά στις προαναφερόμενες βάσεις δεδομένων στα νευρωνικά δίκτυα στις St Gall, Parzival και G. Washington τα ποσοστά ήταν 6,2%, 6,7%, και 18,1% έναντι 10,6%, 15,5% και 24,1% αντίστοιχα εκείνων σε HMM.

Οι εμφανώς ανώτερες επιδόσεις των τεχνικών που βασίζονται σε αρχιτεκτονικές νευρωνικών δικτύων αποτελούν την αιτία για ολοένα και μεγαλύτερη ενασχόληση των ερευνητών για βελτιστοποίηση τεχνικών που βασίζονται στην εν λόγω τεχνική όσον αφορά το πεδίο της αναγνώρισης χειρόγραφου κειμένου.

Σε άλλες μελέτες έχουν γίνει αναφορές σχετικά με τη χρήση μιας άλλης σύγχρονης τεχνικής ονομαζόμενης ως συνελλικτικό επαναλαμβανόμενο νευρωνικό δίκτυο για μεσαιωνικά έγγραφα. Πρόκειται για κατηγορία βαθιών νευρωνικών δικτύων τα οποία επίσης χρησιμοποιούνται συνήθως για την ανάλυση οπτικών εικόνων. Οι συνελλικτικές διαστρωματώσεις πραγματοποιούν αυτόματη εξαγωγή χαρακτηριστικών πράγμα που αποκλείει την ανάγκη για ύπαρξη χειρόγραφων γεωμετρικών σχημάτων ή βασισμένων σε γραφήματα

χαρακτηριστικών όπως εκείνα της περίπτωσης του HisDoc. Η βασική διαφορά είναι ότι οι τεχνικές αρχιτεκτονικής βαθέως νευρωνικών δικτύων είναι περισσότερο ανταγωνιστικές σε σχέση με άλλες τεχνικές ιδίως για την αποδοτικότητα που επιτυγχάνουν από πλευρά αποδοτικότητας του χρόνου αλλά και το γεγονός ότι απαιτούν σημαντικότερη υπολογιστική ισχύ. Αυτή επιτυγχάνεται μέσω της χρήσης γραφικής μονάδας επεξεργασίας και όχι μέσω της κεντρικής μονάδας επεξεργασίας του υπολογιστή.

Στη βάση Rodrigo σε έρευνα που έγινε επετεύχθη το καλύτερο δυνατό αποτέλεσμα τους όπου το ποσοστό σφάλματος λέξεων ήταν μόλις 14%. Με την ίδια τεχνική με νευρωνικό δίκτυο έγινε απόπειρα μεταγραφής της παπικής αλληλογραφίας στα αρχεία του Βατικανού στο project Codice Ratio.

Η χρήση εργαλείου που εφαρμόζει επαναλαμβανόμενες τεχνικές που βασίζονται σε αρχιτεκτονική νευρωνικών δικτύων αποδίδει αποτελέσματα υψηλής ακρίβειας και υπεροχής και στα δύο είδη επεξεργασίας ιστορικών ντοκουμέντων τόσο στα οπτικής αναγνώρισης χαρακτήρων όσο και στην αναγνώριση χειρογράφων.

2.5 ΧΡΗΣΗ ΕΡΓΑΛΕΙΩΝ ΑΝΑΓΝΩΡΙΣΗΣ ΧΕΙΡΟΓΡΑΦΟΥ ΚΕΙΜΕΝΟΥ

Σε αυτή την ενότητα γίνεται μια αναφορά στα εργαλεία λογισμικού και σύνολα δεδομένων τα οποία χρησιμοποιούνται από ερευνητές και άλλους χρήστες στην προσπάθεια επεξεργασίας ιστορικών ντοκουμέντων.

Ειδικότερα όσον αφορά τα εργαλεία αναγνώρισης χειρογράφων σε σχέση με τα πρώιμα έντυπα που είναι αρκετά και ποικίλα για αναγνώριση οπτικών χαρακτήρων (όπως τα OCR, Abbey FineReader, Tesseract, OCRopus, και AnyOCR εργαλεία και ιδιαίτερα το IMPACT) είναι σχετικά λιγότερα.

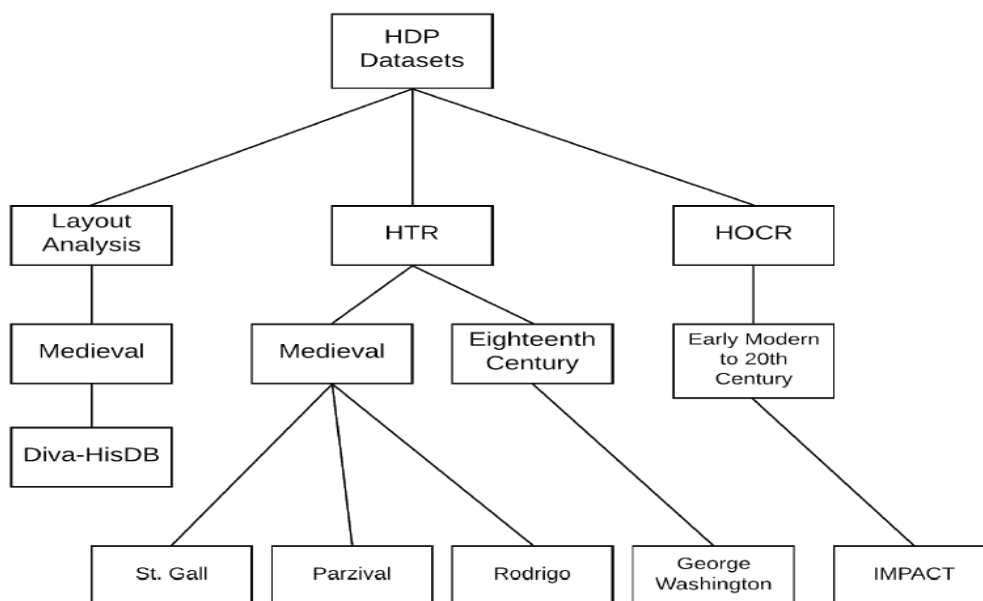
Στα χειρόγραφα οι ερευνητές έχουν πρόσβαση στο IAM-HistDB, Rodrigo και στη βάση δεδομένων Codice Ratio. Σε αυτά περιλαμβάνονται εικόνες από πλήρεις σελίδες χειρογράφων

καθώς και από μεμονωμένες λέξεις αλλά και χαρακτήρες γραμμάτων που αντιστοιχούν σε μεσαιωνικά λατινικά αλλά και σε πρώιμα ισπανικά και γερμανικά χειρόγραφα.

Επίσης το IAM-HistDB περιλαμβάνει την βάση δεδομένων Washington για ιστορική αναγνώριση χειρόγραφου κειμένου. Εκτός της φάση της μεταγραφής και επεξεργασίας ιστορικών εγγράφων από το εργαλείο λογισμικού και βάσεις δεδομένων υπάρχουν και εκείνες οι βάσεις όπως Diva-HistDB που χρησιμοποιούνται για την έρευνα ανάλυσης διάταξης όπως και για ενέργειες άλλων προεπεξεργασιών.

2.6 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Παρότι έχει γίνει σε τεράστια κλίμακα επεξεργασία υλικού ιστορικών εγγράφων, οι ανάλογες βάσεις δεδομένων για εκπαίδευση και δοκιμές και αξιολόγηση παραμένουν σχετικά λιγοστές με εκείνες που χρησιμοποιούνται στο πεδίο της σύγχρονης αναγνώρισης χειρογράφων. Για δυτικής προέλευσης ιστορικά έγγραφα, υπάρχουν διαθέσιμες βάσεις από γλώσσες όπως τα μεσαιωνικά λατινικά, μεσαιωνικά γερμανικά όπως και ισπανικά και μια πληθώρα άλλων σύγχρονων ευρωπαϊκών γλωσσών όπως επίσης και αγγλικά μέχρι το 18^ο αιώνα.



Εικ. 18. Εικόνα ιεραρχίας βάσης συνόλου δεδομένων ανάλογα την περίπτωση χρήσης και το χρονικό διάστημα.

Όσο για τις όποιες επιλογές υπάρχουν διαθέσιμες για τους ερευνητές για να επεξεργαστούν και να δουλέψουν επάνω σε μεταγραφές μεσαιωνικών χειρογράφων είναι μετρημένες. Για παράδειγμα εντός της IAM Historical Database περιλαμβάνονται 2 υποσύνολα δεδομένων όπως εκείνη της St Gall. Η συγκεκριμένη περιλαμβάνει εικόνες με χειρόγραφα στα λατινικά του 9^{ου} αιώνα γραμμένα σε Καρολίνεια γραφή από έναν γραφέα. Πέρα από τις εικόνες της σελίδας και τις μεταγραφές, περιλαμβάνεται εκτεταμένης αλήθειας εδάφους, γραμμές κειμένου και μία προς μία απομονωμένες λέξεις όπου οι εικόνες έχουν μεταφερθεί σε δυαδικές και σημειώθηκαν με κατά λέξη μεταγραφή σε επίπεδο κάθε γραμμής.

Η αρχική σχετική ανάπτυξη προήλθε από το Project HisDoc για την ολοκλήρωση της εργασίας στην ευθυγράμμιση της μεταγραφής αρχειακών εγγράφων σε μορφή εικόνας και την μετέπειτα χρήση της βάσης σε περεταίρω ερευνητικές διαδικασίες. Όσον αφορά το γλωσσικό επίπεδο, παρότι τα Λατινικά ήταν η τότε κυρίαρχη επιστημονική και εκκλησιαστική γλώσσα σε πανευρωπαϊκό επίπεδο κατά τη διάρκεια του μεσαίωνα, μέρος της λογοτεχνικής παραγωγής δημιουργήθηκε σε κοινές γλώσσες. Ειδικότερα για όσους ερευνητές ασχολούνται με αναγνώριση χειρογράφων σε κλασικά χειρόγραφα όπως στα παλαιά γερμανικά και παλαιά ισπανικά ή σε διαλέκτους ισπανικές όπως τα καστιλιάνικα υπάρχουν δύο σύνολα δεδομένων.

Εντός της IAMHistDB υπάρχει και η βάση Parzival η οποία σε αντίθεση με την βάση St Gall χειρόγραφα της οποίας δημιουργήθηκαν από έναν γραφέα, στην πρώτη οι χειρόγραφες εικόνες της πάρθηκαν από τρία διαφορετικά χειρόγραφα και είναι γραμμένα από τρεις διαφορετικούς γραφείς χρησιμοποιώντας γοτθικό γραφικό στυλ με διατάξεις πολλαπλών στηλών. Οι γραμμές κειμένου και οι εικόνες μιας λέξης είναι δυαδικές και κανονικοποιημένες και σχολιασμένες με πλήρη μεταγραφή επιπέδου γραμμής.

Επίσης υπάρχει η βάση Rodrigo Corpus, που περιλαμβάνει ισπανικό σύνολο δεδομένων σε πάνω από 850 σελίδες. Η κατά αρχήν δημιουργία του είναι ανάλογη προηγούμενων βάσεων δεδομένων για αναγνώριση χειρογράφων και γραμμές εξόρυξης έρευνας, οι ερευνητές χρησιμοποίησαν για το έργο ψηφιοποίηση εικόνων από παλαιά ισπανικά χρονικά που αποκαλούνται «Historia de Espana el Archbisipo Don Rodrigo».

Παρά το γεγονός ότι το έργο χρονολογείται στον 14^ο αιώνα, το σύνολο δεδομένων του χειρογράφου είναι περίπου από το 1545 και προφανώς μπορεί να εντοπιστεί στα όρια της πρώιμης εμφάνισης της τυπογραφίας. Παρά το γεγονός ότι οι δημιουργοί του δημοσίευσαν αποτελέσματα υβριδικού ταξινομητή εικόνας με βάση τα γλωσσικά μοντέλα Hidden Markov, κατ'αντιστοιχία οι Granell χρησιμοποίησαν το σύνολο των δεδομένων της βάσης σύμφωνα με το αρχιτεκτονικό πρότυπο των βαθιών νευρωνικών δικτύων.

Μια άλλη βάση δεδομένων που περιλαμβάνεται στην IAM HistDB είναι εκείνη της Washington. Περιλαμβάνει επιλεγόμενη αλληλογραφία γραμμένη από τον G. Washington αλλά καθώς αναφέρεται χρονολογικά σε εποχή πιο σύγχρονη από εκείνη του μεσαίωνα δε θα κάνουμε επεκταθούμε στην ανάλυσή της. Αυτό που χρήζει ιδιαίτερης αναφοράς είναι πως αναλογικά είναι το μοναδικό ιστορικό σύνολο δεδομένων εγγράφων το οποίο έχει σχεδιαστεί για αναγνώριση χειρόγραφης γραφής και επομένως πρόκειται για ένα ιδιαίτερα πολύτιμο εργαλείο.

Επίσης παρότι τα προηγούμενα σύνολα δεδομένων όσον αφορά την μεσαιωνική αναγνώριση γραφής περιορίστηκαν σε ένα μικρό αριθμό χειρογράφων σε μια σειρά από λατινογενείς πρώιμες γλώσσες από λίγες χώρες προέλευσης, οι ερευνητές και αρχειονόμοι διαθέτουν μια πολύ πιο εμπειριστατωμένη βάση εργασίας στη βάση δεδομένων με την ονομασία IMPACT.

Η δημιουργία της έγινε από κοινοπραξία ευρωπαϊκών βιβλιοθηκών και εμπεριέχει μια πανευρωπαϊκή συλλογή κειμένων μεταξύ του 15^{ου} και 20^{ου} αιώνα σε ποικιλία ευρωπαϊκών γλωσσών. Ένα από τα μειονεκτήματα της εν λόγω βάσης ήταν ότι εμφάνιζε σχετική δυσκολία και περιορισμούς απέναντι σε λατινικά κείμενα. Παρά το γεγονός ότι τα λατινικά αποτελούσαν την κοινή γλώσσα των επιστημών εντούτοις ένα μικρό μέρος της επιστημονικής παραγωγής έχει αναπαραχθεί στα λατινικά οπότε δεν συνιστά μείζων μειονέκτημα η όποια εμφανιζόμενη σχετική αδυναμία στην αναγνώριση λατινικών κειμένων.

Στη συλλογή της εν παραπάνω βάσης περιλαμβάνεται τόσο το πλήρες κείμενο του εγγράφου σε Unicode μαζί με διαστρωματική ανάλυση διάταξης σχολιασμού καθώς και εντολή ανάγνωσης τα οποία καθορίζονται με μορφή PAGE βασιζόμενα σε XML. Η μεταγραφή με βάση το Unicode από το κείμενο της σελίδας επιτρέπει την πλέον ακριβέστερη κωδικοποίηση μιας σειράς από

ποικίλα έντυπα εντός του συνόλου δεδομένων καθώς και των συνδέσμων τους όπου χρησιμοποιούνται ειδικοί χαρακτήρες συντομογραφίας σε μεγάλη κλίμακα σε μια σειρά από πρώιμα σύγχρονα κείμενα.

Άξιο αναφοράς επίσης είναι πως σχετικά με την προαναφερθείσα βάση IAM HistDB το έργο HistDoc συνέχισε με τη δημιουργία του Diva-HistDB το οποίο αποτελεί ένα σύνολο δεδομένων από περίπου 150 σελίδες με αποσπάσματα εικόνων από τρία διαφορετικά χειρόγραφα, συνοδευόμενα από ανάλυση διάταξης, τμηματοποίηση γραμμών. Δύο από τα παραπάνω χειρόγραφα προέρχονται από τον 14^ο αιώνα και ένα είναι του 11^{ου} αιώνα στη λεγόμενη Chancery γραφή. Σε όλα τα χειρόγραφα περιλαμβάνεται μια στήλη κειμένου η οποία περιβάλλεται από εκτεταμένο περιθώριο σχολιασμού. Μερικές σελίδες έχουν διακοσμημένους τους αρχικούς χαρακτήρες. Όσο για τις διατάξεις αυτές επομένως είναι αρκετά περίπλοκες και η κωδικοποίηση της γίνεται στη μορφή PAGE XML. Το σύνολο της βάσης διατίθεται σε ελεύθερη πρόσβαση στο διαδίκτυο μέσω της ιστοσελίδας του έργου HistDoc.

Στη βάση In Codice Ratio το περιεχόμενό της είναι δομημένο έτσι ώστε να εκπαιδεύσει τον ταξινομητή τους σε χειρόγραφες μεσαιωνικές γραφές. Σε σχέση με άλλες βάσεις μεσαιωνικών δεδομένων χειρογράφων, η In Codice Ratio στοχεύει στους ιδιαίτερους χαρακτήρες και τον αντίστοιχο σχολιασμό. Από τα χειρόγραφα που ψηφιοποιούν περιέχονται 23 διακριτοί χαρακτήρες, από τους οποίους εξήγαγαν χαρακτήρες από δύο αντιπροσωπευτικές σελίδες των χειρογράφων. Στη συνέχεια μετά από κάποια συνθετική αύξηση προκειμένου να δημιουργηθούν επιπλέον παραδείγματα από χαρακτήρες για εκπαίδευση, το σύνολο δεδομένων περιείχε έναν αριθμό 1000 παραδειγμάτων χαρακτήρα για κάθε μία κατηγορία χαρακτήρα. Αυτά τα δείγματα εικόνων συμπεριλαμβάνονται σε ένα σύνολο δεδομένων αποτελούμενο από εικόνες λέξεων σε PNG και οι συσχετιζόμενες μεταγραφές κειμένου περιλαμβάνονται σε ξεχωριστή βάση. Και αυτή η βάση είναι ελεύθερης πρόσβασης μπορεί να ληφθεί δωρεάν από τον ιστότοπο του έργου.

2.7 ΤΑΣΕΙΣ ΣΕ ΕΡΓΑ ΨΗΦΙΟΠΟΙΗΣΗΣ

Την τελευταία 10ετία έχει υπάρξει μια πληθώρα ερευνητικών προγραμμάτων και έργων στον τομέα της επεξεργασίας ιστορικών εγγράφων μέσω της δημιουργίας συνόλου δεδομένων, της εξερεύνησης για τις βέλτιστες δυνατές τεχνικές και την εφαρμογή των όποιων εργαλείων που ήδη υπάρχουν στις προσπάθειες διατήρησης και αναγνώρισης ψηφιακών αρχείων. Κάθε προσπάθεια κάθε έργο με τη σειρά του συνέβαλε στο να ανευρίσκονται νέοι τρόποι, επαρκείς μέθοδοι και συνέβαλε με τον τρόπο του στο να συμπληρωθεί το παζλ με τα εργαλεία εκείνα που προσέφεραν το καλύτερο δυνατό αποτέλεσμα.

Μέσα από την δημιουργία διαφόρων έργων στο χώρο των ιστορικών ντοκουμέντων υπήρξε σημαντική συμβολή σε αλγόριθμους εργαλεία και σύνολα δεδομένων για μεσαιωνικά χειρόγραφα. Ένα από τα εμβληματικά έργα αυτής της κατηγορίας ήταν το HisDOC αποτέλεσμα της συνέργειας των Πανεπιστημίων της Ελβετίας, του Fribourg της Βέρνης και του Neuchatel. Το εν λόγω έργο ασχολήθηκε ειδικότερα και με τις τρεις φάσεις στην ιστορική επεξεργασία εγγράφων. Στην ανάλυση της διάταξης, στην αναγνώριση γραφής αλλά και στην ευρετηρίαση αλλά και ανάκτηση των εγγράφων.

Παρά το γεγονός πως το μεγαλύτερο μέρος της έρευνας επικεντρώθηκε σε μεσαιωνικά έγγραφα, ο κύριος στόχος ήταν η δημιουργία μεθόδων επεξεργασίας ιστορικών εγγράφων οι οποίες μπορούν να έχουν γενική εφαρμογή σε πολλά σενάρια και γλώσσα. Στο δε Πανεπιστήμιο του Fribourg οι ερευνητές του επικεντρώθηκαν σε ζητήματα ανάλυσης διάταξης. Δημιουργήθηκε εργαλείο που ειδικεύεται στην ταξινόμηση διάφορων στοιχείων από κάθε σελίδα χειρογράφου που περιλαμβάνει τόσο στοιχεία του κειμένου όσο και εικονογράφηση, διακοσμητικά στοιχεία και χαρακτήρες όπως και στοιχεία που βρίσκονται στο προσκήνιο της σελίδας όσο και ως φόντο.

Η συγκεκριμένη μεθοδολογία χαρακτηρίζεται ως «πυραμυδική προσέγγιση» κατά την οποία τμήματα της διάταξης της σελίδας αναγνωρίζονται διαδοχικά σε διαφορετικές εικόνες ανάλυσεων χάρη σε ταξινομητή ο οποίος βασίζεται σε νευρωνικά δίκτυα. Τόσο στο προσκήνιο όσο και στο φόντο τα στοιχεία ήταν οριοθετημένα σε χαμηλή ανάλυση με εικόνα χαμηλής

κλίμακας και στοιχεία κειμένου και εκτός κειμένου διακρίνονται σε σχετικά υψηλότερη ανάλυση. Τα παραπάνω στοιχεία της μεθόδου επέτρεψαν την μετατροπή των στοιχείων κειμένου σε γραμμές και λέξεις και στοιχεία εκτός κειμένου σε εικονογραφήσεις και διακοσμητικά σχέδια.

Παράλληλα η ερευνητική ομάδα του Πανεπιστημίου της Βέρνης εστίασε την προσοχή της κυρίως στη φάση της μεταγραφής. Στο πλαίσιο αυτό έγινε προσπάθεια διερεύνησης σε επίπεδο πολλαπλών προσεγγίσεων για την μετατροπή εικόνων γραμμής κειμένου σε ψηφιακό κείμενο στο οποίο συμπεριλαμβάνονταν ο εντοπισμός λέξεων κλειδιών με πλήρη μεταγραφή, όπως επίσης και τεχνικές στρατηγικής μηχανικής μάθησης όπως κρυφά μοντέλα Markov και νευρωνικά δίκτυα αμφίδρομης μακροπρόθεσμης και βραχυπρόθεσμης μνήμης επαναλαμβανόμενων νευρωνικών δικτύων.

Χρησιμοποιήθηκαν επίσης στατιστικά γλωσσικά μοντέλα για τη βελτίωση της ακρίβειας μεταγραφής. Η ερευνητική ομάδα της Βέρνης εμβάθυνε στην ευθυγράμμιση της μεταγραφής σύζευξης μιας υπάρχουσας μεταγραφής με την αντίστοιχη θέση της σε μια γραμμή κειμένου. Το συγκεκριμένο θεωρήθηκε ως πιθανή λύση στην έλλειψη επισημασμένων δεδομένων.

Με στόχο την ανάπτυξη μιας μηχανής αναζήτησης προορισμένης για ιστορικά χειρόγραφα η τρίτη ομάδα του Πανεπιστημίου του Neuchatel εργάστηκε στο να αντιμετωπίσει τις διάφορες προκλήσεις στον τομέα της ανάκτησης πληροφοριών στο πλαίσιο των ιστορικών εγγράφων. Σε συγκεκριμένα ερωτήματα αναζήτησης το σύστημα που ανέπτυξαν αναπαράγει μια λίστα αντιστοίχισης χειρογράφων ταξινομημένα κατά συνάφεια σε συνάρτηση με το αρχικό ερώτημα. Ιδίως δε, η έρευνα αντιμετώπισε τις προκλήσεις που αφορούν δυσκολίες σε ιστορικά κείμενα όπως στην αναγνώριση λέξεων κατά τη φάση της μεταγραφής όπως επίσης στην καμπτική και ορθογραφική ποικιλία. Ως επίλυση του προβλήματος στη μείωση επιπέδων σφάλματος κατά την μεταγραφή, υιοθετήθηκαν εναλλακτικές υποθέσεις με υψηλές πιθανότητες για να αυξήσουν την κύρια μεταγραφή ως μέρος του κειμένου το οποίο ευρετηριάστηκε από το σύστημα ανάκτησης πληροφοριών.

Η ανάκτηση των πληροφοριών είναι ο τελικός στόχος της επεξεργασίας ιστορικών εγγράφων και σε αυτό το επίπεδο η ερευνητική ομάδα του Neuchatel έκανε σημαντική δουλειά. Μέσω της

ταυτόχρονης διερεύνησης της ανάλυσης διάταξης, της αναγνώρισης και ανακτησης πληροφοριών, το εναρκτήριο έργο HisDoc αποτέλεσε μια εμβληματική προσπάθεια στο χώρο καθώς προχώρησε ταυτόχρονα σημαντικές πτυχές του τομέα επεξεργασίας ιστορικών χειρογράφων.

Σημαντική κρίνεται ιδιαίτερα η αναγνώριση για την ύπαρξη ενός ενοποιημένου συστήματος για την επεξεργασία ιστορικών εγγράφων και τη δημιουργία ενός σημαντικού συνόλου δεδομένων όπως το IAM-HistDB.

Μια από τις συνεισφορές του HisDoc ήταν και η δημιουργία μιας διαδικτυακής υπηρεσίας την Diva Services η οποία παρέχει ιστορικούς αλγόριθμους επεξεργασίας εγγράφων με λεγόμενη αντιπροσωπευτική μεταφορά κατάστασης API για να γίνει παράκαμψη του προβλήματος που αντιμετωπίζουν πολλοί προγραμματιστές και επαγγελματίες ως προς την εγκατάσταση πολύπλοκων εργαλείων λογισμικού, βιβλιοθηκών κλπ.

Πέρα από την Diva Services προέκυψε και άλλη υπηρεσία, ένας ψηφιακός χώρος εργασίας το Diva Desk με λογισμικό το οποίο είναι βασισμένο σε GUI το οποίο μετατρέπει αλγόριθμους επιστήμης υπολογιστών για τη δημιουργία τη διάταξη αλήθειας εδάφους, καθήκοντα προσβάσιμα σε μελετητές και ερευνητές των ανθρωπιστικών επιστημών. Διερευνήθηκε η περιοχή κειμένου αλλά και ανάλυση διάταξης με νευρικό δίκτυο αλλά και ταυτοποίηση των συγγραφέων. Αποτέλεσμα ήταν η δημιουργία του συνόλου δεδομένων Diva HisDB. Στην ίδια λογική πρόσφατα ξεκίνησε μια προσπάθεια του HisDOC προ τριών περίπου ετών, η οποία και επιδιώκει να αναπτύξει μεθόδους βασισμένους στη μάθηση σε βάθος και επίσης να ταξινομήσει ογκώδεις αριθμούς χειρογράφων χωρίς ετικέτα σε αρχεία και δημιουργία ροών εργασίας που θα μπορούσαν να εφαρμοστούν σε μια σειρά από Βιβλιοθήκες και άλλα αρχεία πολιτιστικής κληρονομιάς.

Ένα άλλο project είναι το IMPACT χρηματοδοτούμενο από την Ευρωπαϊκή Ένωση προκειμένου να ανπτυχθεί τεχνογνωσία και τις απαραίτητες τεχνικές μεθόδους γνώσεις και υποδομές για να μπορεί να υποστηριχθεί η ψηφιοποίηση της ευρωπαϊκής πολιτιστικής κληρονομιάς. Πα'ρά την ταχύτητα με την οποία οι ευρωπαϊκές Βιβλιοθήκες προχωρούσαν στην ψηφιοποίηση κειμένων,

η αντίστοιχη διαθεσιμότητα πλήρους κειμένου μεταγραφές δεν συμβάδιζαν. Με την επίλυση παρόμοιων προκλήσεων ψηφιοποίησης από πολλές βιβλιοθήκες, οι διαδικασίες επίλυσης επαναλήφθηκαν με αποτέλεσμα τόσο την σπατάλη χρόνου αλλά και οικονομικών πόρων. Για παράδειγμα το κόστος μιας μη αυτόματης μεταγραφής ενός και μόνο βιβλίου κυμαίνεται μεταξύ 400 και 1000 ευρώ ανάλογα με το μέγεθος του βιβλίου. Μέσω της ευρωπαϊκής κοινοπραξίας βιβλιοθηκών το IMPACT προσέφερε ενοποιημένη τεχνογνωσία στην ψηφιοποίηση και ανέπτυξε εργαλεία πόρους και βέλτιστες πρακτικές για να ξεπεράσει τις προκλήσεις της ψηφιοποίησης σε μια μεγάλη κλίμακα.

Ένα άλλο project είναι το έργο In Codice Ratio το οποίο σε συνεργασία με τα αρχεία του Βατικανού έχει αναπτύξει μια νέα αλγοριθμική τεχνική τμηματοποίησης, εργαλεία λογισμικού και ένα σύνολο δεδομένων δοκιμής για μεταγραφή την ιδιωτική αλληλογραφία των Παπών. Μάλιστα, μέχρι την έναρξη αυτού του έργου ψηφιοποίησης, σχεδόν κανένα από αυτό το πραγματικά τεράστιο αρχείο δεν είχε υποστεί επεξεργασία σάρωσης πόσο μάλλον μεταγραφής. Πολλά από αυτά τα έγγραφα χρονολογούνται από τον 8^ο αι. μ.Χ. και προκειμένου να συμβουλευτεί και να ερευνήσει αυτά τα έγγραφα απαιτούνταν φυσική παρουσία του μελετητή.

Η καινοτομία εδώ έγκειται στο ότι με δεδομένο ότι αυτά τα χειρόγραφα είναι γραμμένο με την Καρολίνειο γραφή, οι ερευνητές του Πανεπιστημίου Roma Tre προσέγγισαν την εργασία μεταγραφής με έναν μοναδικό, υβριδικό τρόπο που επηρέασε τόσο την οπτική αναγνώριση χαρακτήρων όσο και την αναγνώριση χειρογράφων. Μια πρόκληση για εκπαίδευση χωρίς τμηματοποίηση σε συστήματα αναγνώρισης χειρογράφου είναι η ανάγκη για επισημασμένα δεδομένα τα οποία σχολιάζονται από παλαιογράφους με ειδίκευση στην ιστορική γλώσσα της εποχής.

Επίσης οι ποικιλίες και ιδιαιτερότητες των χειρογράφων δυσκολεύουν και αρκετές φορές οδηγούν σε αποτυχία συστήματα αναγνώρισης οπτικών χαρακτήρων. Με βάση την αναγνώριση των πιο πάνω αδυναμιών οι ερευνητές του In Codice Ratio προχώρησαν στην ανάπτυξη συστήματος αναγνώρισης το οποίο χρησιμοποιεί πλέγμα τμηματοποίησης με τα τοπικά ελάχιστα της κατανομής των μαύρων εικονοστοιχείων για την αναγνώριση του υπό-χαρακτήρα «περικοπές». Ένας ταξινομητής χαρακτήρων ο οποίος βασίζεται σε ένα συνελκτικό νευρικό

δίκτυο, προσδιορίζει στη συνέχεια τα υπολείμματα μη προσδιορισμένων γραμμάτων ως χαρακτήρες. Η πρόβλεψη λέξεων επιτυγχάνεται με ένα κρυφό μοντέλο γλώσσας βασισμένο στο μοντέλο Μάρκοβ όπου ο ταξινομητής χαρακτήρων εκπαιδεύεται χρησιμοποιώντας μια προσαρμοσμένη διεπαφή ιστού μέσω της ανάθεσης έργου σε μια ομάδα ανθρώπων με ανοιχτή πρόσκληση στην οποία παρουσιάζονται παραδείγματα μεμονωμένων χαρακτήρων στους συμμετέχοντες όπου στη συνέχεια επισημαίνουν χαρακτήρες δίχως ετικέτα.

Μέσω του παραπάνω συστήματος επετεύχθη ένα μέσο ποσοστό περί το 96% ακριβείας για μεμονωμένους χαρακτήρες και προκαταρκτικά ποσοτά σφάλματος λέξεων σε δείγματα σελίδων από την συλλογή του Βατικανού. Η κυριότερη συνεισφορά των ερευνητών που ασχολήθηκαν στο έργο του In Codice Ratio ήταν ότι οι παραδοσιακές μέθοδοι μηχανικής μάθησης, τα νευρωνικά δίκτυα και ένα γλωσσικό μοντέλο, έχει τη δυνατότητα να ενσωματωθεί με επιτυχία σε ένα σύστημα ποιοτικής παραγωγής στην επεξεργασία ιστορικών εγγράφων.

ΚΕΦΑΛΑΙΟ 3^ο

3. ΚΑΤΕΥΘΥΝΣΕΙΣ ΓΙΑ ΠΕΡΕΤΑΙΡΩ ΕΡΕΥΝΑ

Στην επεξεργασία ιστορικών εγγράφων γίνεται μετατροπή σαρωμένων εικόνων εγγράφων σε ψηφιακές μεταγραφές. Μετά την απαραίτητη προεπεξεργασία εικόνων εγγράφων μέσω δυαδικοποίησης, ανάλυσης διάταξης και τμηματοποίησης γραμμών, οι εικόνες μεμονωμένων γραμμών μετατρέπονται σε ψηφιακό κείμενο μέσω της αναγνώρισης χειρόγραφου κειμένου ή μέσω της οπτικής αναγνώρισης χαρακτήρων.

Στις τελευταίες εξελίξεις, οι αρχικές συμβατικές τεχνικές μηχανικής μάθησης οι οποίες χρησιμοποιούν χειρόγραφους χαρακτήρες και πιο πρόσφατα μεθόδους που βασίζονται σε νευρωνικά δίκτυα, έχουν γίνει πλέον αυτονόητες τεχνικές επίλυσης για την παραγωγή ακριβών μεταγραφών ιστορικών κειμένων από μεσαιωνικά χειρόγραφα και των πρώιμων τυπωμένων εγγράφων του 15^{ου} αι.

Οι μεθοδολογίες βαθιάς μάθησης αποτελούν σημαντικό μέρος της δημιουργίας μιας αυτοματοποιημένης αλυσίδας εργαλείων για επεξεργασία ιστορικών εγγράφων όσο και για σημαντικές εκτενείς ιστορικές συλλογές αρχείων. Έργα όπως το Impact HisDoc το Transcriptorium έχουν δώσει την δική σημαντική συνεισφορά στη δημιουργία σημαντικών συνόλων δεδομένων αλλά και εργαλείων λογισμικού.

Ο συνδυασμός της τεχνογνωσίας των επιστημόνων της πληροφορικής, των ιστορικών, των αρχειονόμων κρίνεται απαραίτητη βάση για την αντιμετώπιση της πρόκλησης με καλύτερους όρους στο μέλλον της επεξεργασίας ιστορικών εγγράφων. Το γεγονός ότι δημιουργείται ένας τεράστιος όγκος ποικιλίας αρχειακών δεδομένων και η ταχύτητα με την οποία δημιουργούνται φαντάζει ως μια πρόκληση από τη σκοπιά αυτού που ονομάζουμε «μεγάλα δεδομένα».

Η δημιουργία υποδομών και αξιόπιστων εργαλείων για αυτή τη νέα εποχή στην επεξεργασία των ιστορικών εγγράφων αποτελεί έναν πραγματικά ιερό σκοπό όλων εκείνων που διακηρυγμένος στόχος τους αποτελεί η διατήρηση της ιστορικής κληρονομιάς κειμένων της ανθρωπότητας στην ψηφιακή εποχή.

3.1 ΣΥΝΕΛΙΚΤΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (Convolutional Neuronic Networks, CNN)

Τα συγκροτήματα νευρωνικών δικτύων πρωτο ξεκίνησαν να έρχονται στο προσκήνιο από την δεκαετία του 1950 με τα πρώτα δειλά βήματα της τεχνητής νοημοσύνης καθώς οι επιστήμονες της πληροφορικής προσπαθούσαν να ενσωματώσουν στο λογισμικό των τότε υπολογιστών δυνατότητες επεξεργασίας οπτικών δεδομένων. Στις επόμενες δεκαετίες η τεχνολογική εξέλιξη αυτής της δυνατότητας απλά προχώρησε σε εκθετικό βαθμό.

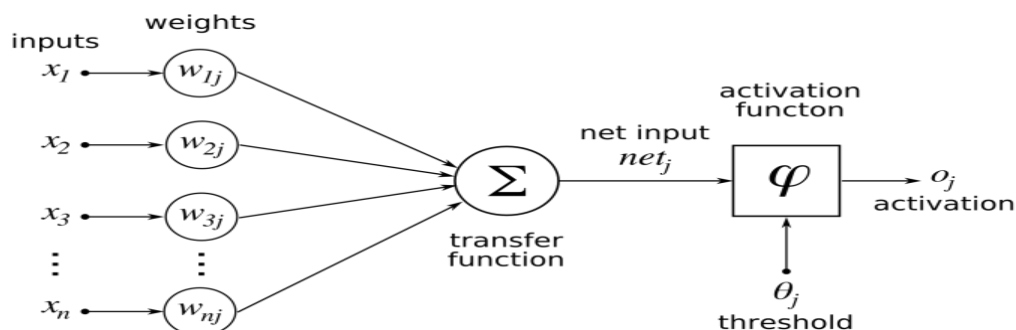
Τα συγκροτήματα νευρωνικών δικτύων έχουν αποτελέσει το θεμέλιο λίθο τόσο στο πεδίο της βαθιάς μάθησης όσο και της τεχνητής νοημοσύνης στην προσπάθεια για όλο και πιο ακριβέστερα αποτελέσματα στην μεταγγραφή χειρόγραφων κειμένων του μεσαίωνα.

Ένα από τα ορόσημα αυτής της τεχνολογικής εξέλιξης ήταν το 2012 καθώς μια ομάδα ερευνητών από το Πανεπιστήμιο του Τορόντο ανέπτυξαν ένα μοντέλο τεχνητής νοημοσύνης το οποίο ξεπέρασε κάθε προσδοκία επιτυγχάνοντας να κερδίσει σε διαγωνισμό ανάλυσης εικόνας με το εκπληκτικό ποσοστό ακρίβειας 85%. [KRIZHEVSKY, 2012]. Το λογισμικό σύστημα τεχνητής νοημοσύνης ονομάστηκε AlexNet χάρη στον δημιουργό του Alex Krizhevsky. Στον τομέα της τεχνητής νοημοσύνης έρχονται να συγκεραστούν τα συγκροτημάτων νευρωνικών δικτύων Convolutional Neural Networks (CNN) τα οποία έρχονται να υποκαταστήσουν μιμούμενα την ανθρώπινη όραση και να αποτελέσουν υπερπολύτιμο εργαλείο στην δημιουργία εφαρμογών αναγνώρισης εικόνων.

Ένα από τα χαρακτηριστικά γνωρίσματα των συγκροτημάτων νευρωνικών δικτύων είναι πως δεν μπορούν να λειτουργήσουν κλιμακωτά. Χρειάζονται πληθώρα δεδομένων προκειμένου να μπορέσουν να επεξεργαστούν μεγάλες σχετικά εικόνες. Αρχικά η τεχνική μπορούσε να εφαρμοστεί σε εικόνες με χαμηλή ανάλυση. Όμως σύμφωνα με τον Krizhevsky δώθηκε η ευκαιρία στην επαναξιολόγηση της βαθιάς μάθησης Deep Learning καθώς πρόκειται για τον κλάδο της τεχνητής νοημοσύνης που χρησιμοποιεί πολυεπίπεδα νευρωνικά δίκτυα. Μάλιστα δε η διαθεσιμότητα μεγάλων συνόλων δεδομένων ImageNet έδωσαν τη δυνατότητα στους ερευνητές να εκτελέσουν μια σειρά από εργασίες αναγνώρισης εικόνων από υπολογιστή που μέχρι τότε δεν ήταν εφικτές.

3.1.1 ΤΡΟΠΟΣ ΛΕΙΤΟΥΡΓΙΑΣ

Τα συγκροτήματα νευρωνικών δικτύων αποτελούνται από σειρές από πολλαπλά στρώματα τεχνητών νευρώνων. Οι τεχνητοί νευρώνες συνιστούν μια πρόχειρη απομίμηση των αντίστοιχων βιολογικών νευρώνων, ως μια σειρά από μαθηματικές συναρτήσεις που υπολογίζουν το σταθμισμένο άθροισμα πολλαπλών εισόδων και εξάγουν μια τιμή ενεργοποίησης.



Εικόνα 19. Δομή ενός τεχνητού νευρώνα, βασικό συστατικό των τεχνητών νευρωνικών δικτύων.

Κάθε νευρώνας, καθορίζεται από τις τιμές. Όταν τροφοδοτείται με τις τιμές των pixel, οι τεχνητοί νευρώνες ενός CNN επιλέγουν μια σειρά από διάφορα οπτικά χαρακτηριστικά. Με την εισαγωγή μιας εικόνας σε ένα συγκρότημα νευρωνικών δικτύων, τα πολυστρωματικά επίπεδα που διαθέτει, δημιουργούν μια σειρά από χάρτες ενεργοποίησης. Οι χάρτες αυτοί, επικεντρώνονται στο να αναγνωρίζουν και να εντοπίζουν τα σχετικά χαρακτηριστικά της όποιας εικόνας. Με τον κάθε νευρώνα να παίρνει ένα κομμάτι pixel ως είσοδο και να πολλαπλασιάζει τις τιμές χρώματος και σε συνδυασμό με το μέγεθος των χαρακτήρων που αναγνωρίζει τα προσλαμβάνει και τα τρέχει προς επεξεργασία μέσω της λειτουργίας ενεργοποίησης.

Συνήθως το αρχικό στρώμα CNN ανιχνεύει και εξετάζει βασικά χαρακτηριστικά μιας εικόνας ή χαρακτήρες συμβόλων που βρίσκονται σε οριζόντια θέση, ή κάθετη, ή σε διαγώνια άκρα. Η έξοδος του πρώτου στρώματος τροφοδοτείται ως είσοδος στο επόμενο στρώμα του οποίου η διεργασία αφορά την εξαγωγή πιο περίπλοκων χαρακτηριστικών όπως γωνίες και συνδυασμούς άκρων. Με την επεξεργασία να προχωρά σε περισσότερα επίπεδα, η ανάλυση

αυξάνεται καθώς αυτά εντοπίζουν αντικείμενα με μεγαλύτερης αξίας ανάλυση και περιγραφή, όπως αντικείμενα πρόσωπα κλπ.

3.2 ΑΝΑΓΝΩΡΙΣΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΣΥΝΕΛΙΚΤΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (CNN Handwriting identification)

Στη συνέχεια θα δούμε πώς η χρήση της μεθόδου συγκροτημάτων νευρωνικών δικτύων διευκολύνει την αναγνώριση κειμένου. Τα χειρόγραφα κείμενα όπως έχουμε δει και προηγουμένως είναι μοναδικά καθώς παράγονται σε διαφορετικές χρονικές περιόδους, τόπους, περιοχές και συγγραφείς που τους προσδίδουν ιδιαίτερα χαρακτηριστικά. Όμως την ίδια ώρα, στοιχεία της μοναδικότητάς τους είναι εκείνα που καθιστούν το χειρόγραφο κείμενο πιο εύκολο να αναγνωριστεί λόγω μιας σειράς παραγόντων όπως η μορφή των γραμμάτων αλλά και τα διαφορετικά στυλ γραφής του κάθε ατόμου/ συγγραφέα.

Η χρήση των CNN έχει ευρεία χρήση από ιστορικές ελληνικές επιγραφές μέχρι λατινικά κείμενα και χειρόγραφα της εποχής μας με αξιόλογα πεδία τόσο σε εκείνα της ιστορικής έρευνας όσο και στην εξιχνίαση εγκλημάτων στην εποχή μας. Οι τελευταίες τεχνολογικές εξελίξεις στην βαθιά μάθηση έχουν επισπεύσει τις διαδικασίες σχετικά με την αναγνώριση χειρόγραφου κειμένου. Πέρα από την τεχνική OCR πλέον υπάρχει και η αναγνώριση κειμένου Intelligent Character Recognition (ICR) κατά την οποία οι αλγόριθμοι που απαιτούνται για την επίλυση αναγνώρισης κειμένου χρειάζονται περισσότερη νοημοσύνη σε σχέση με την παλαιότερη μέθοδο αλγορίθμων εκείνης του OCR.

Τα νευρωνικά δίκτυα αποδεδειγμένα έχουν επιτύχει σημαντικά ποσοστά επιτυχίας αναγνώρισης προτύπων. Τα βαθιά νευρωνικά δίκτυα βρίσκονται παρόντα στα στάδια ταξινόμησης, τμηματοποίησης και ανίχνευσης χαρακτήρων. Η επιτυχία αυτή έχει αναγνωριστεί και στην περίπτωση της μεθόδου αναγνώρισης οπτικών χαρακτήρων OCR, όπου τα νευρωνικά δίκτυα έχουν την ικανότητα να προβλέψουν μια ακολουθία χαρακτήρων από κείμενοι που δημιουργείται από μηχανή με σχεδόν τέλεια ακρίβεια. Η έξυπνη αναγνώριση χαρακτήρων ICR είναι το καθήκον της αποκρυπτογράφησης του ψηφιοποιημένου χειρογράφου του κειμένου. Είναι πολύ πιο απαιτητικό από την οπτική αναγνώριση κειμένου OCR καθώς τα χειρόγραφα

σύμβολα και χαρακτηριστικά είναι μοναδικά. Τα συστήματα ICR μπορεί να είναι διαδικτυακά ή εκτός σύνδεσης. [SUCH, 2018]

3.3 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗ ΚΕΙΜΕΝΟΥ.

Η μηχανική μάθηση είναι ένα πεδίο ολοένα και επεκτεινόμενο και με ιδιαίτερες προοπτικές καθώς αφορά τον πολλά υποσχόμενο τομέα της τεχνητής ευφυΐας artificial intelligence (AI). Η εξέλιξη στον τομέα αφορά ένα αναπτυσσόμενο πεδίο αναγνώρισης χειρόγραφου κειμένου για μεσαιωνικά χειρόγραφα.

Οι μηχανές με βάση την τεχνητή νοημοσύνη είναι σε θέση να εκπαιδευτούν για να μεταγραφούν κείμενα σε σχεδόν οποιοδήποτε σύστημα γραφής και μάλιστα άμεσα μέσω ψηφιακών εικόνων και με επίπεδα ακρίβειας που πλησιάζουν τα ανθρώπινα. Ως πρώτο βήμα της όλης διαδικασίας αποτελεί η δημιουργία ενός ικανού μοντέλου, πράγμα το οποίο απαιτεί χρόνο, δοκιμή και σφάλμα και επαρκή παροχή αλήθειας εδάφους και ακριβείς μεταγραφές οι οποίες έχουν απογυμνωθεί από μη χειρόγραφο υλικό. [Handwritten text recognition, 2021]

Η αναγνώριση χειρογράφων γίνεται μέσω χρήσης του τομέα της μηχανικής μάθησης μέσω της χρήσης μιας συλλογής δεδομένων τα οποία στη συνέχεια δέχονται επεξεργασία για να μελετηθούν τα μοτίβα ώστε να αποκτηθεί η συγκεκριμένη γνώση. Στη συνέχεια, όταν έχουμε εμβάθυνση της διαδικασίας μάθησης στα υπάρχοντα δεδομένα και πρότυπα θα ακολουθήσει και η βαθιά μάθηση. [SUDANA, 2020]

Ως ένα από τα πεδία εφαρμογής της μηχανικής μάθησης και της τεχνητής ευφυΐας στην αναγνώριση κειμένου, αποτελεί η πλατφόρμα e-scriptorium. Η ύπαρξη ποικιλίας ειδών χειρογράφων, διατάξεων από διάφορους πολιτισμούς μεσαιωνικούς εισαγάγει ένα σημαντικό βαθμό δυσκολίας προκειμένου να προγραμματιστεί το ανάλογο λογισμικό σε έναν υπολογιστή ώστε να καταλάβει και αναγνωρίσει όλα αυτά τα διαφορετικά μοτίβα. Επίσης μπορεί οι χρήστες της πλατφόρμας να μην θέλουν να μεταγράψουν τη σελίδα γράμμα προς γράμμα, ή να μην χρειάζονται τις σημειώσεις εκτός περιθωρίου ή και να τις περιλαμβάνουν. Χάρη στη μηχανική

μάθηση ο χρήστης μπορεί να δημιουργεί παραδείγματα του τι χρειάζεται να κάνει ο υπολογιστής και το μηχάνημα τα επεξεργάζεται για να μάθει ανάλογα με το τι χρειάζεται ο χρήστης. Ανάλογα με το τι θα ζητήσει και θα επεξεργαστεί από τον κάθε χρήστη το μηχάνημα θα μπορέσει αυτόματα να τα εφαρμόσει ακολουθώντας τις ίδιες αρχές για άλλα άλλες εικόνες. [RESILIENCE Tool: eScriptorium, 2020]

3.4 ΒΑΘΙΑ ΜΑΘΗΣΗ ΚΑΙ ΣΥΓΚΡΟΤΗΜΑΤΑ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ.

Βαθιά μάθηση είναι η διαδικασία εκμάθησης των τεχνητών νευρωνικών δικτύων στα οποία τα πρότυπα μελετώνται σε περισσότερο βάθος για να παράγουν καλύτερα και πιο αξιόπιστα αποτελέσματα. Χαρακτηριστικό της βαθιάς μάθησης είναι πως μιμείται την ικανότητα του ανθρώπινου εγκεφάλου να θυμάται πράγματα στάδια διαδικασίες.

Μέσω της εφαρμογής της βαθιάς μάθησης επιτυγχάνεται η μάθηση κάθε υπάρχουσας δυνατότητας όσο πιο βαθύτερο γίνεται το δίκτυο. Η σχεδίασή της είναι τέτοια για να χρησιμοποιεί διαφορετικά επίπεδα τεχνητού νευρωνικού δικτύου προκειμένου να μπορεί να μάθει για κάθε ιδιαίτερο χαρακτηριστικό.

Η αρχιτεκτονική των μοντέλων βαθιάς μάθησης αποτελεί πολύ σημαντικό στοιχείο στην ικανότητα του δικτύου να μελετά τα υπάρχοντα δεδομένα. Είναι γεγονός πως υπάρχουν μια σειρά από μοντέλα στη βαθιά μάθηση όπως τα βαθιά νευρωνικά δίκτυα, τα συγκροτήματα νευρωνικών (CNN) όπως και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN).

Τα CNN είναι ένας αλγόριθμος βαθιάς μάθησης που χρησιμοποιεί στρώμα συγκροτημάτων για να εξάγει χαρακτηριστικά καθώς και ένα πλήρως συνδεδεμένο επίπεδο για την ταξινόμηση των χαρακτηριστικών. Η εκπαιδευτική διαδικασία στο συγκρότημα νευρωνικών δικτύων χρησιμοποιεί έναν αλγόριθμο “backpropagation” για να μπορέσει να κάνει τις ανάλογες ενημερώσεις σε σχέση με τα χαρακτηριστικά της κάθε περιόδου.

Κάθε νευρώνας ευρισκόμενος εντός του συνελικτικού νευρωνικού δικτύου CNN συνδέεται με την εικόνα εισόδου, οπότε οι αρχιτεκτονικές του χρειάζονται αρκετά βαθιά στρώματα. Η διαδικασία της ταξινόμησης χρησιμοποιεί πλήρως συνδεδεμένα επίπεδα για να ταξινομήσει τα υπάρχοντα χαρακτηριστικά με προκαθορισμένες τάξεις. Ένα από τα συνήθη μειονεκτήματα της μεθόδου είναι πως ο απαιτούμενος όγκος δεδομένων είναι αρκετά μεγάλος ώστε να επιτευχθούν τα βέλτιστα δυνατά αποτελέσματα. Επομένως το υλικό που απαιτείται για να επιτευχθεί πρέπει να είναι σημαντικό.

Η αναγνώριση χειρογράφου χρησιμοποιώντας αλγορίθμους βαθιάς μάθησης θα είναι σε θέση να παράγει υψηλότερης ακρίβειας τιμές συγκρινόμενη με άλλες μεθόδους. Χειρόγραφα που συλλέχθηκαν από διαφορετικούς συγγραφείς τα δεδομένα τους θα χρησιμοποιηθούν για έρευνα και μάθηση. Η χρήση αλγορίθμου βαθιάς μάθησης αναμένεται να είναι σε θέση να βελτιώσει τα αποτελέσματα της αναγνώρισης χειρογράφου έτσι ώστε να μπορεί να χρησιμοποιηθεί πιο σωστά. [KESTEMONT, 2017]

3.5 ΕΝΤΟΠΙΣΜΟΣ ΛΕΞΕΩΝ ΚΛΕΙΔΙΩΝ (Word Spotting) ΚΑΙ ΧΡΗΣΗ ΣΥΝΕΛΙΚΤΙΚΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ (CNN).

Εργασία άμεσης συνάφειας με την αναγνώριση κειμένου αποτελεί ο εντοπισμός λέξεων κλειδιών στην οποία περιστατικά αναζήτησης λέξεων εντοπίζονται σε ψηφιοποιημένα κείμενα. Η χρήση εντοπισμού λέξεων κλειδιών είναι κατάλληλη στο πεδίο ιστορικών εγγράφων, δεδομένου ότι στα συστήματα αναγνώρισης υπάρχει η πιθανότητα να αποτύχουν εξαιτίας της κακής ποιότητας του χειρόγραφου κειμένου αλλά και της έλλειψης διαθέσιμων και κατάλληλων λεξικών τα οποία χρησιμοποιούνται για τον έλεγχο των αποτελεσμάτων της αναγνώρισης κειμένου. Και σε αυτό το επίπεδο επεξεργασίας για την ανάλυση κειμένου και τη μεταγραφή του, προτείνεται εφαρμογή αρχιτεκτονικών συγκροτημάτων νευρωνικών δικτύων για τον εντοπισμό λέξεων κλειδιών σε χειρόγραφα έγγραφα. [LOMBARDI, 2020]

Ο εντοπισμός λέξεων έχει αναπτύξει ένα έντονο ερευνητικό ενδιαφέρον για ανάλυση εικόνων. Προ ολίγων ετών προτάθηκε η χρήση attributeSVMs προβλέποντας δυαδική παράσταση

χαρακτηριστικών [ALMAZAN 2014]. Αυτή η μέθοδος καθόρισε την τεχνολογική εξέλιξη στο πεδίο της τμηματοποίησης και εντοπισμού λέξεων. Στο κεφάλαιο αυτό γίνεται αναφορά στην εύρεση λέξεων με την χρήση συνελκτικών νευρωνικών δικτύων (CNN). Προτείνονται αρχιτεκτονικές συγκροτημάτων νευρωνικών δικτύων σχεδιασμένες για εντοπισμό λέξεων. Οι αρχιτεκτονικές αυτές είναι σε θέση να εκπαιδευτούν από άκρο σε άκρο. Αποδίδεται στη χρήση CNN η επίτευξη υψηλής αποτελεσματικότητας και τεχνικής μεθόδων τεχνολογίας αιχμής τόσο για την εύρεση λέξεων όσο και στην τμηματοποίηση κειμένων σε μια μεγάλη ποικιλία βάσεων δεδομένων. [SUDHOLT, 2018]

3.6 ΜΕΤΑΦΟΡΑ ΜΑΘΗΣΗΣ (TRANSFER LEARNING)

Η μεταφορά μάθησης αποτελεί υποπεδίο της μηχανικής μάθησης. Μια επιπλέον μέθοδος για την αναγνώριση χειρογράφων από συγκροτήματα νευρωνικών δικτύων είναι η μεταφορά μάθησης ή αλλιώς transfer learning. Η μεταφορά μάθησης σε συνελκτικό νευρικό δίκτυο σημαίνει την επαναχρησιμοποίηση μοντέλου που έχει εκπαιδευτεί χρησιμοποιώντας τεράστιο όγκο δεδομένων με εξαιρετικά αποτελέσματα. [REHMAN A., 2019]

Το προεκπαιδευμένο ουσιαστικά μοντέλο είναι αποτέλεσμα της προσπάθειας εύρεσης ενός αλγορίθμου για ανίχνευση και ταξινόμηση αντικειμένων. Το αποτέλεσμα της χρήσης αυτού του προεκπαιδευμένου μοντέλου είναι εξαρτημένο από το μοντέλο αρχιτεκτονικής το οποίο θα χρησιμοποιηθεί. Επομένως ανάλογα με το μοντέλο χρήσης της αρχιτεκτονικής, ανάλογο θα είναι και το αποτέλεσμα και διαφορετική η τιμή της ακρίβειας. Η έννοια της μεταφοράς μάθησης είναι η χρήση γνώσεων που αποκτήθηκαν για ένα έργο προς επίλυση όμως αυτή τη φορά μιας άλλης σχετικής εργασίας. Τα οφέλη χρήσης της μεταφοράς μάθησης είναι συνήθως ψηλότερη εκκίνηση και κέρδος χρόνου.

Επίσης μια από τις προκλήσεις κατά τη διάρκεια της οποίας έρευνας γίνεται είναι να αποτελέσει αυτή σημαντικό εφιαλτήριο για να κάνει το τρέχον μοντέλο πιο στοιβαρό. Οπότε η χρήση της βάσης δεδομένων με τις όποιες σημαντικές προκλήσεις αυτή προσφέρει και από άποψη

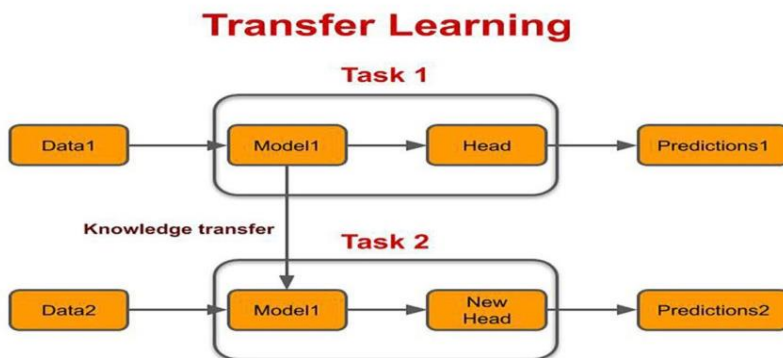
εικόνων, μεταγραφής, θορύβου κλπ βοηθά στην μάθηση ενός ικανού για χρήση μοντέλου για επίλυση πραγματικών αναγκών.

Ένα από τα πεδία χρήσης της μεταφοράς μάθησης αναφέρεται και η υπολογιστική όραση.

Η έλευση της βαθιάς μάθησης προέτρεψε σε νέες προσεγγίσεις μεταφοράς μάθησης αναφορικά με το παρελθόν, με κάποιες να αναφέρονται στη συνέχεια. [PAN, 2010] Σύμφωνα με την έρευνα των Pan & Young (2010) με δυαδική ταξινόμηση εγγράφων, η μεταφορά μάθησης περιλαμβάνει τις έννοιες ενός τομέα και μιας εργασίας. Όπου D ο τομέας και X ο χώρος των χαρακτηριστικών και μια οριακή κατανομή πιθανότητας $P(X)$ πάνω από το χώρο χαρακτηριστικών, όπου $X=x_1, \dots, x_n \in X$.

Ειδικότερα για ταξινόμηση εγγράφων με αναπαράσταση λέξεων bag of words το X είναι ο χώρος των εμφανίσεων των εγγράφων, x_i είναι το i th ο όρος διάνυσμα που αντιστοιχεί σε κάποιο έγγραφο και X είναι το δείγμα εγγράφων τα οποία χρησιμοποιούνται για την εκπαίδευση. Ουσιαστικά μια από τις εφαρμογές της μεταφοράς μάθησης είναι να μαθαίνεις μέσα από προσομοιώσεις, και επιτυγχάνοντας διάδραση μεταξύ των δεδομένων και του μοντέλου που χρησιμοποιείται είναι πολύ πιο οικονομικό και εύχρηστο να δοκιμαστεί ένα πιλοτικό μοντέλο στην αναγνώριση κειμένων πριν αυτό χρησιμοποιηθεί στην πραγματικότητα.

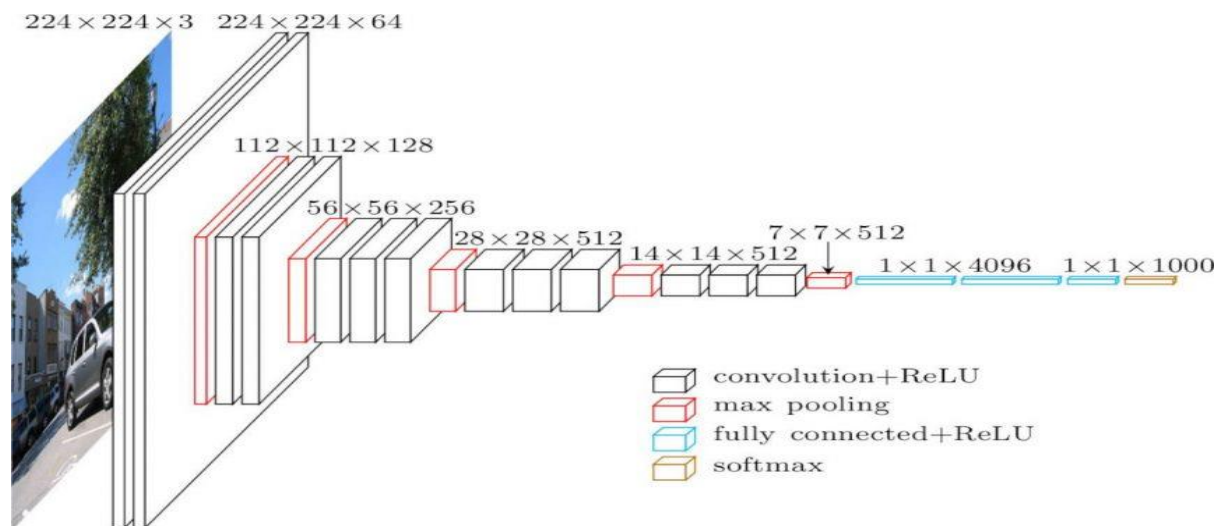
Ο πιο συνηθισμένος τρόπος μεταφοράς μάθησης που έχει εφαρμοστεί, αντιπροσωπεύεται στην εξαιρετική επιτυχία χρήσης των μεγάλων συγκροτημάτων νευρωνικών δικτύων στο ImageNet. [KRIZHEVSKY, 2012]



Εικ. 20. Απεικόνιση μοντέλου μεταφοράς μάθησης

3.7 ΜΟΝΤΕΛΑ ΑΡΧΙΤΕΚΤΟΝΙΚΗΣ ΣΥΝΕΛΚΤΙΚΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ (CNN)

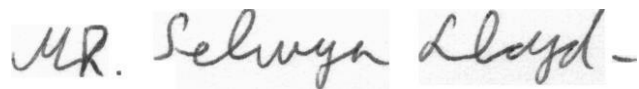
Μια από τις πιο γνωστές αρχιτεκτονικές βαθιάς μάθησης για συνελκτικά νευρωνικά δίκτυα είναι η VGG. Ειδικότερα η VGG16 και η VGG19 ήταν μια από τις πρώτες αρχιτεκτονικές που εμφανίστηκαν και παρουσιάστηκαν από τους Simonyan και Zisserman το 2014 με την εργασία τους [SIMONYAN, 2014]. Πρόκειται για μια σχετικά απλή αρχιτεκτονική η οποία χρησιμοποιεί μόνο μπλοκ αποτελούμενα από έναν αυξανόμενο αριθμό συνελκτικών επιπέδων με φίλτρα μεγέθους 3×3 . Πέρα απ αυτό, για να μπορέσει να μειωθεί το μέγεθος των ληφθέντων χαρτών ενεργοποίησης, τα μπλοκ μέγιστης συγκέντρωσης διασπώνται μεταξύ των συνελκτικών, έτσι μειώνεται το μέγεθος των χαρτών. Οι αριθμοί 16 και 19 στην αρχιτεκτονική VGG αναφέρονται στον αριθμό σταθμισμένων στρωμάτων που έχει κάθε δίκτυο. Πρόκειται για το μοντέλο VGG ή αλλιώς Visual Geometry Group το οποίο ενσωματώνει ανάλογα 16 βαθιά επίπεδα ή 19.



Εικ. 21. Μοντέλο αρχιτεκτονικής δομής επιπέδων VGG16

Η αρχιτεκτονική αυτή δείχνει ότι βαθύ στρωματικό επίπεδο στο CNN αποτελεί σημαντικό παράγοντα για τη δημιουργία ενός συστήματος ταξινόμησης με υψηλή ακρίβεια αποτελεσμάτων. Το δε VGG19 είναι εκπαιδευμένο να χρησιμοποιεί σύνολα δεδομένων ImageNet με 1000 κατηγορίες εικόνων οι οποίες μπορούν να βοηθήσουν να ξεπεραστεί το πρόβλημα του περιορισμένου αριθμού εικόνων στο σύνολο δεδομένων που χρησιμοποιούνται στη διαδικασία της εκπαίδευσης. [Neurohive, 2018]

Στην έρευνα των Sudana et al (2020) χρησιμοποιώντας το μοντέλο VGG19 έχουν επιτευχθεί πολύ υψηλά αποτελέσματα σε εικόνες με κλίμακα του γκρι με πραγματικό ποσοστό αποδοχής 92,3% και ποσοστό σφάλματος ίσο με 7,7% [SUDANA,2020]. Ειδικότερα στην συγκεκριμένη εργασία χρησιμοποιήθηκε η βάση IAM. Οι υπό επεξεργασία εικόνα, χωρίστηκε σε τρεις κατηγορίες, εικόνες κλίμακας του γκρι, δυαδικές και ανεστραμμένες δυαδικές εικόνες. Οι τρεις αυτές κατηγορίες έχουν την ίδια εικόνα με διαφορετικό χρώμα και έχουν εκπαιδευτεί χρησιμοποιώντας την ίδια αρχιτεκτονική CNN



(α)



(β)



(γ)

Εικ. 22. Σύνολο δεδομένων εικόνας με τη μορφή: (α)κλίμακα του γκρι, (β) δυαδικό και (γ) αντίστροφο δυαδικό.

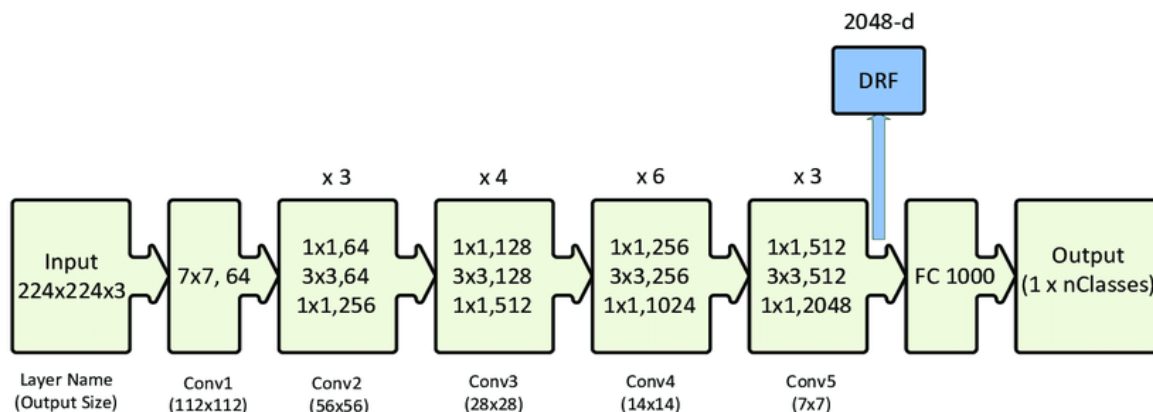
Ως διαδικασία εκπαίδευσης στην πιο πάνω έρευνα, χρησιμοποιήθηκε το Google Colaboratory [CARNEIRO T., 2018] όπου ο κύκλος της πλήρους επεξεργασίας των δεδομένων έφτασε τις 100 εποχές.

Στην πιο πάνω έρευνα χρησιμοποιήθηκε η μέθοδος μεταφοράς μάθησης σε αναγνωρισμένους συγγραφείς με βάση το χειρόγραφό τους και χρησιμοποιήθηκε προ-εκπαιδευόμενο CNN. Εκπαιδεύτηκε μοντέλο χρησιμοποιώντας σύνολο δεδομένων χειρόγραφου IAM αποτελούμενο από 100 κατηγορίες συγγραφέων. Με βάση τα αποτελέσματα της εκπαίδευσης και των δοκιμών, το προεκπαιδευόμενο μοντέλο VGG19 παρουσιάζει καλύτερη απόδοση κατά τη χρήση εικόνων σε κλίμακα του γκρι συγκριμόμενο με τη χρήση δυαδικών ή ανεστραμμένων δυαδικών εικόνων. Επίσης η έρευνα των Sudana et al δείχνει ότι η εικόνα χειρογράφου σε μορφή προτάσεων μπορεί να χρησιμοποιηθεί άμεσα χωρίς να απαιτείται άλλη μέθοδος τμηματοποίησης ή εξαγωγής

χαρακτηριστικών. Μειονέκτημα αποτέλεσε ο χρόνος ολοκλήρωσης της διαδικασίας να τρέξει η έρευνα.

Επίσης υπάρχει και η αρχιτεκτονική ResNet η οποία αναπτύχθηκε το 2015 και η οποία αποτέλεσε ένα πραγματικό ορόσημο στην εισαγωγή ενός τύπου αρχιτεκτονικής που βασίζεται σε modules ή αλλιώς ενότητες, ή σε ότι αναφέρεται ως τα τώρα ως «δίκτυα εντός δικτύων».

Υπάρχουν παραλλαγές του ResNet με διαφορετικό αριθμό διαστρωμάτωσης, αλλά η πιο ευρέως διαδεδομένη και χρησιμοποιούμενη είναι το ResNet50, πράγμα το οποίο σημαίνει πως αποτελείται από 50 επίπεδα διαστρωμάτωσης. Μάλιστα, παρά το γεγονός το ότι περιλαμβάνει πολύ περισσότερα επίπεδα από τα μοντέλα VGG, εντούτοις χρειάζεται πολύ λιγότερη χρήση μνήμης, σχεδόν πέντε φορές λιγότερη. Αυτό οφείλεται στο γεγονός ότι το δίκτυο αντί για πυκνά επίπεδα στο στάδιο της ταξινόμησης, χρησιμοποιεί ένα τύπο επιπέδου το οποίο ονομάζεται GlobalAveragePooling, το οποίο μετατρέπει του 2D χάρτες δραστηριότητας του τελευταίου επιπέδου στο στάδιο εξαγωγής χαρακτηριστικών σε έναν φορέα n-class που χρησιμοποιείται για τον υπολογισμό της πιθανότητας συμμετοχής σε κάθε τάξη. [ROMAN, 2020]



Εικ. 23. Δομή αρχιτεκτονικής ResNet-50 όπως εμφανίζεται με τις υπόλοιπες μονάδες, το μέγεθος των φίλτρων και τις εξόδους κάθε συνελκτικού στρώματος. Εμφανίζεται το εικονιζόμενο DRF το οποίο έχει εξαχθεί από το τελευταίο συνελκτικό επίπεδο αυτού του δικτύου.

3.8 Βάσεις δεδομένων και χρήση τους από CNN

Το ImageNet είναι μια βάση δεδομένων με πάνω από 15 εκατομμύρια εικόνες υψηλής ανάλυσης που ανήκουν σε περίπου 22,000 κατηγορίες. Η συλλογή των εικόνων έγινε από το διαδίκτυο και η επισήμανσή τους με ετικέτες χρησιμοποιώντας τουρκικό εργαλείο μηχανικής Amazon Turk και εύρεσης πλήθους πηγών. Το ImageNet αποτελείται από εικόνες μεταβλητής ανάλυσης. Οι εικόνες είναι σε ανάλυση 256x256. Για παράδειγμα λαμβάνοντας υπόψη μια ορθογώνια εικόνα, αυτή επανασυνδέεται και περικόπτει το κεντρικό εικονίδιο 256x256 από την εικόνα που προκύπτει [“The evolution of ImageNet for Deep Learning in Computer vision, 2020”].

CLaMM16 είναι βάση δεδομένων που χρησιμοποιήθηκε για διαγωνισμό ICFHR2016 σχετικά με την ταξινόμηση μεσαιωνικών χειρογράφων με λατινική γραφή [STUTZMANN D., 2016]. Αποτελεί την πλέον σημαντική βάση για τους διαγωνισμούς για την ταξινόμηση των μεσαιωνικών χειρογράφων σε λατινική γραφή, οι οποίοι διοργανώνονται από κοινού από επιστήμονες πληροφορικής καθώς και από ανθρωπιστές (παλαιογράφους) στα ICFHR2016 και ICDAR2019, με κύριο στόχο την ανάκτηση εικόνας για ιστορικά χειρόγραφα έγγραφα. Παρέχει πρόσβαση σε μια πλούσια βάση δεδομένων ευρωπαϊκών μεσαιωνικών χειρογράφων για ανάλυση και αναγνώριση χειρογράφου. Οι εικόνες του σετ εξάσκησης έχουν επισημανθεί ανάλογα σε 12 ετικέτες. Ο διαχωρισμός των χειρογράφων βασίζεται σε μορφολογικές διαφορές και τύπους γραφής όπως αυτά καθορίζονται σε τυπικά λατινικά γραπτά. Οι τύποι χωρίζονται σε: Caroline, Cursiva, Half-Uncial, Humanistic, Humanistic Cursive, Hybrida, Praegothica, Semihybrida, Rotunda, Textualis, Uncial. Ακολουθούν ενδεικτικές εικόνες από τους παραπάνω τύπους γραφής.

ΑΒCΔΕFCHIKLMNOPQRSTU
ΑΒCΔΕFCHIKLMNOPQRSTUXZ

Εικ. 23 Uncial Script

α β c d e f g h i k l m n o p q r s t u x y
α β c d e f g h i k l m n o p q r r α u x z
α β c d e f g h i k l m n o p q r r γ τ u x y

Εικ. 24. Half-uncial τύπος γραφής

abcdefghijklmnopqrstuvwxyz
 a b c d e f g h i k l m n o p q r s t u x y

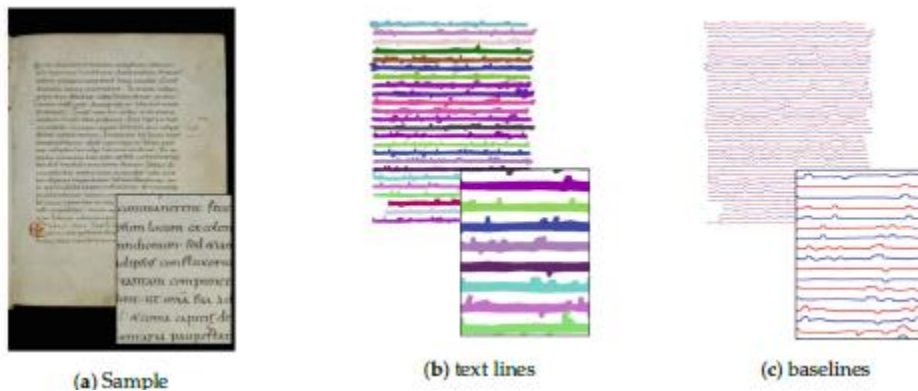
Εικ. 25. Καρολίνειος τύπος γραφής

abcdefghijklmnopqrstuvwxyz
 a b c d e f g h i k l m n o p q r s t u x y z
 a b c d e f g h i k l m n o p q r s t u x y z

Εικ. 26 Cursiva τύπος γραφής

3.9 CNN και μέθοδοι τμηματοποίησης

Στον τομέα της τμηματοποίησης και αναγνώρισης κειμένου η χρήση μεθόδων των νευρωνικών δικτύων είναι πολύ σημαντικής σημασίας καθώς πρόκειται για τεχνολογία αιχμής και διαρκώς εμπλουτίζεται με νέες ερευνητικές προσπάθειες κάποιες από τις οποίες θα καταγραφούν στη συνέχεια. Η τμηματοποίηση γραμμών κειμένου είναι ένα από τα πλέον βασικά απαιτούμενα για την ανάλυση χειρογράφων και γενικότερα ιστορικών εγγράφων. Ενώ η τμηματοποίηση γραμμών κειμένου σε έντυπα είναι ευκολότερη διαδικασία, η τμηματοποίηση ιστορικού χειρογράφου ενέχει δυσκολίες ιδίως λόγω της ακανόνιστης γραμμής του κειμένου αλλά και σε συνδιασμό με την ύπαρξη και άλλων αντικειμένων όπως π.χ σημειώσεων που προστέθηκαν στη σελίδα.



Εικ 27. Προσεγγίσεις στον προσδιορισμό γραμμών κειμένου στο χειρόγραφο. α) γραμμή κειμένου τμηματοποίηση. β) βασική ανίχνευση γ) γραμμές βάσεις

Επιπλέον, γεγονός αποτελεί ότι πολλά από τα χειρόγραφα δημιουργήθηκαν προσπαθώντας να καλύψουν λιγότερες δυνατές σελίδες με αποτέλεσμα όχι μόνο οι λέξεις αλλά και οι γραμμές του κειμένου να έρχονται πλησιέστερα η μία της άλλης. Η τμηματοποίηση γραμμών κειμένου στοχεύει στον

προσδιορισμό της περιοχής που αντιστοιχεί στο κάθε κείμενο και αντιμετωπίζεται με μεθόδους συνήθως της λεγόμενης βαθιάς μάθησης. [ZHAO, 2020]

Σύμφωνα με τη μέλετη του Chen [CHEN K., 2017] προτείνεται ένα απλό CNN το οποίο έχει μόνο ένα στρώμα συνέλιξης. Οι γραμμές κειμένου, η τμηματοποίηση, επιτυγχάνεται μέσω επισήμανσης ενός υπερ-πίξελ, όπου κάθε εικονοστοιχείο ταξινομείται ως φόντο, ή ως κύριο κείμενο, διακόσμηση, είτε σχόλιο. Η είσοδος στην ταξινόμηση συντελείται μέσω μιας ενημερωμένης έκδοσης του κώδικα που επικεντρώνεται στο υπερ-εικονοστοιχείο. Μετά από πειράματα και έρευνες στις βάσεις δεδομένων St Gall και Parzival η υψηλή ακρίβεια ως αποτέλεσμα είναι το κύριο χαρακτηριστικό.



Εικ. 28. Παραδείγματα τριών τύπων χειρογράφων του μεσαίωνα στη βάση DIVA-HisDB

Μια άλλη σύγχρονη μέθοδος τμηματοποίησης έγινε από τους Pastor-Pellicer [PASTOR-PELLICER J., 2016]. Στην μελέτη τους αντιμετωπίζουν την τμηματοποίηση γραμμών κειμένου υπολογίζοντας την λεγόμενη Κύρια Περιοχή Σώματος Main Body Area (MBA) για κάθε γραμμή του κειμένου, όπου η κύρια περιοχή σώματος αποτελεί την περιοχή μεταξύ της γραμμής βάσης και της γραμμής σώματος. Σε αυτήν την περίπτωση, λαμβάνονται δύο επίπεδα ταξινόμησης. Αφενός κάθε pixel εικόνας ταξινομείται σε τρεις κατηγορίες. Στο φόντο, το κυρίως σώμα του κειμένου και τη διακόσμηση. Χρησιμοποιώντας την σήμανση εικονοστοιχείων επιτυγχάνεται ο εντοπισμός περιοχών κειμένου. Στη συνέχεια ένα δεύτερο συγκρότημα νευρωνικών δικτύων ταξινομεί τα εικονοστοιχεία σε σώματα κειμένων και δίνει τη δυνατότητα παρατήρησης ενός συρόμενου παραθύρου στο κέντρο ενός εικονοστοιχείου και απαντά στο εάν αυτό το σύνολο του κειμένου αποτελεί ή δεν αποτελεί Κύρια Περιοχή Σώματος. Η συγκεκριμένη μέθοδος χρησιμοποιήθηκε σε βάσεις δεδομένων που αφορούν μεσαιωνικά χειρόγραφα όπως η Parzival και η St Gall. Ως προς τα αποτελέσματα η μέθοδος Pastor-Pellicer βασίζεται στη βαθιά μάθηση και δείχνει βελτιωμένα αποτελέσματα συγκρινόμενα με άλλες. Άλλες εναλλακτικές προσεγγίσεις που αφορούν αντιμετώπιση που δεν βασίζεται σε λύση με χρήση βαθιάς μάθησης,

που ξεκινούν με ανάλυση ιστορικών σελίδων με διαδικασία εξομάλυνσης κλίμακας και του χώρου του κειμένου, με μεθόδους που βασίζονται σε σχέση με τον προσδιορισμό συνδεδεμένων στοιχείων και φτάνουν μέχρι σε μεθόδους ομαδοποίησης για τμηματοποίηση των γραμμών του κειμένου. Σύγκριση αφορά και την τεχνική εξαγωγής γραμμών κειμένου βασιζόμενη στη δυναμική πολλαπλών διαστρωματώσεων.

Οι μέθοδοι τμηματοποίησης μπορούν να χωριστούν στις παρακάτω κατηγορίες όπως τις μεθόδους κοκκώδεις βάσης (granular), στα μπλοκ κειμένου και της υφής. Στις κοκκώδεις η έρευνα αρχίζει από τα μικρότερα στοιχεία της σελίδας όπου εικονοστοιχεία προσπαθούν να ενσωματωθούν σε μεγαλύτερη ομοιογενή δεδομένα. Σύμφωνα με τους Garz [GARZ R., 2011] χρησιμοποιήθηκε η μέθοδος SIFT η οποία πρότεινε την ανίχνευση οντοτήτων διάταξης τοπικά και εν συνεχεία χρησιμοποίηση ταξινομητή περιοχής SVM.

Με βάση τη μέθοδο των μπλοκ οι εικόνες σελίδας διαχωρίζονται σε μικρότερες περιοχές και στη συνέχεια οι εν λόγω περιοχές τυγχάνουν επεξεργασίας μέχρι να αποδώσουν ολοκληρωμένα εννοιολογικά και σημασιολογικά αλλά και από άποψη δεδομένων χαρακτηριστικά. Μια αντίστοιχη μέθοδος προτάθηκε από τον Chen για τμηματοποίηση κειμένου σε σύνθετες εικόνες εγγράφων [CHEN K., 2017]. Τα έγγραφα κόβονται σε μπλοκ τα οποία αποτελούν πολλαπλά συνιστώμενα όρια για δημιουργία πολλών επιπέδων. Τα συνδεδεμένα στοιχεία του κάθε στρώματος αναγνωρίζονται κατηγοριοποιούνται και ομαδοποιούνται σε μπλοκ με βάση ένα προκαθορισμένο σύνολο χαρακτηριστικών.

Μέθοδοι με βάση την υφή κατά τις οποίες εξάγουν τα χαρακτηριστικά των εικόνων της σελίδας και στη συνέχεια ταξινομούνται τα περιεχόμενα χάρη στη χρήση στατιστικών μοντέλων. Ένα τέτοιο μοντέλο προτάθηκε από τον [ASI A., 2015] όπου περιγράφεται ένα ισχυρό πλαίσιο για την ανάλυση ιστορικών χειρογράφων. Εκεί γίνεται τμηματοποίηση του κειμένου σε χωρικά συνεκτικές περιοχές και γραμμές κειμένου με χρήση φίλτρων με βάση την υφή χαρακτηριστικών και με την βελτιστοποίηση τους χάρη στην αξιοποίηση των Markov Random Fields.

3.10 ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΑ ΣΥΝΕΛΙΚΤΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (CRNN)

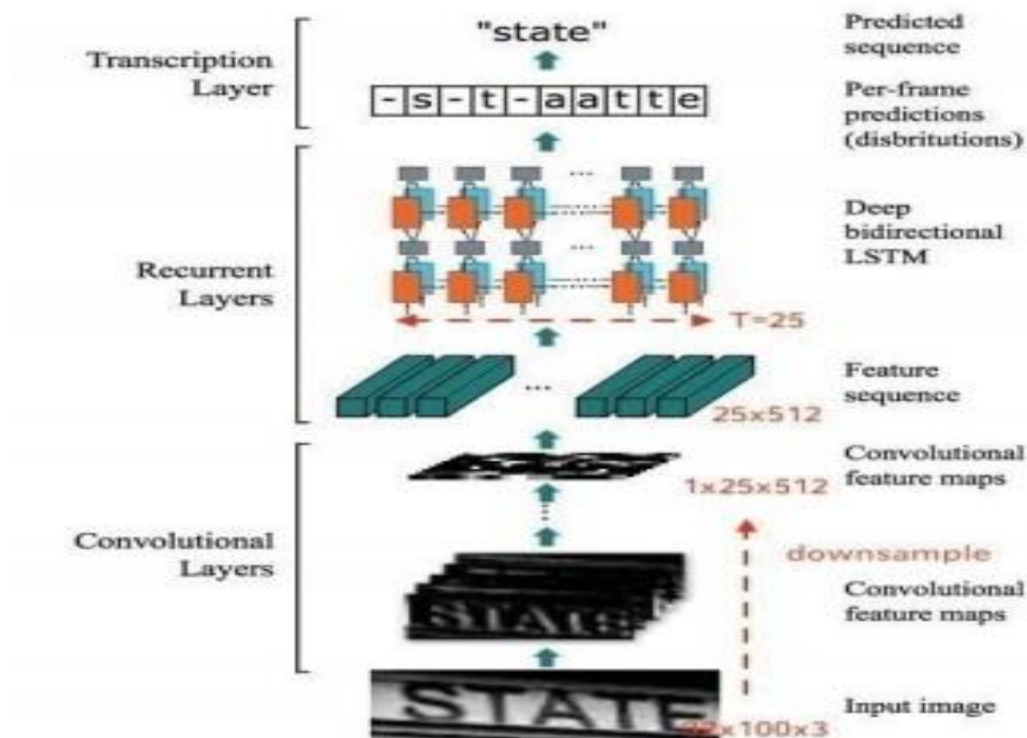
Η έννοια του Convolutional Recurrent Neural Network (CRNN) ή αλλιώς Συγκροτήματα επαναλαμβανόμενων νευρωνικών δικτύων παρουσιάστηκε για πρώτη φορά από τον καθηγητή Shen [SHI 2016] το οποίο κατά κύριο λόγο χρησιμοποιήθηκε για την αναγνώριση από άκρο σε άκρο ακολουθίας κειμένου αορίστου μήκους. Δεν απαιτείται προ-τμηματοποίηση σε συνεχή μεγάλο κείμενο αλλά μετατρέπει την αναγνώριση κειμένου σε πρόβλημα αλληλουχίας σε άμεση συνάρτηση με τον παράγοντα χρόνο, επομένως αναγνώριση ακολουθίας με βάση κατά κύριο λόγο την εικόνα.

Η δομή ενός επαναλαμβανόμενου συνελικτικού νευρωνικού δικτύου αποτελείται από τα παρακάτω τρία μέρη.

1.CNN (συγκρότημα νευρωνικού δικτύου)

2.RNN Το επαναλαμβανόμενο νευρωνικό δίκτυο χρησιμοποιείται για την πρόβλεψη ακολουθίας χαρακτηριστικών όπου ο κάθε φορέας χαρακτηριστικών επισημαίνεται και μαθαίνεται με βάση τη σειρά της ακολουθίας και η παραγματική τιμή αποτελεί την έξοδο.

3.Απώλεια CTC, όπου το επίπεδο μεταγραφής χρησιμοποιεί τη λειτουργία απώλειας για να μετατρέψει μια σειρά διανομών από ετικέτες από το επαναλαμβανόμενο στρώμα στην τελική ακολουθία ετικέτας



Εικ. 29. Δομή επαναλαμβανόμενου συγκροτήματος νευρωνικού δικτύου.

Η δομή CTC αφορά αλγόριθμο ο οποίος δημιουργήθηκε και παρουσιάστηκε το 2006 από τον Graves σαν μια από τις συνηθισμένες επιλογές για να επιλύσουν προβλήματα αναγνώρισης κειμένου και άλλα σχετιζόμενα με ακολουθίες.

3.11 FCNN

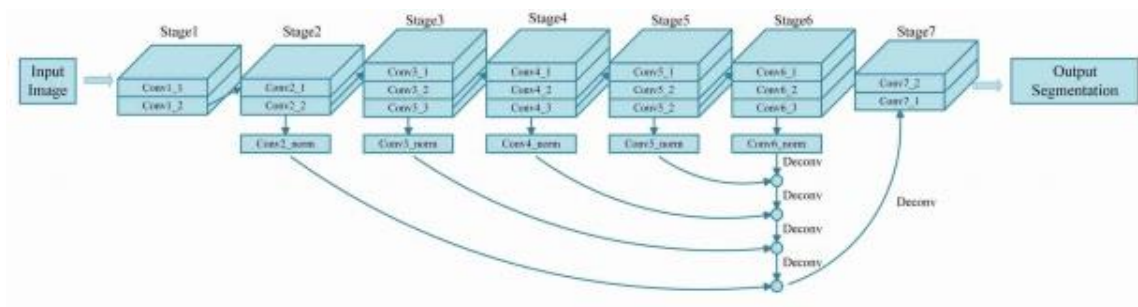
Πρόκειται για πλήρως συνελκτικά νευρωνικά δίκτυα ή αλλιώς FCNN τα οποία αποτελούν μια αρχιτεκτονική η οποία χρησιμοποιείται κυρίως για σημασιολογική τμηματοποίηση. Γίνεται χρήση μόνο τοπικώς διασυνδεδεμένων στρωμάτων όπως εκείνων της συνέλιξης, ομαδοποίησης και δειγματοληψίας. Με την αποφυγή της χρήσης πυκνών στρωμάτων και την υιοθέτηση επεξεργασίας λιγότερων ποσοτικά παραμέτρων, επιτυγχάνεται μεγαλύτερη ευελιξία και αποτελέσματα σε βέλτιστο χρόνο. Το σύνθηδες δίκτυο αποτελείται από μια διαδρομή δειγματοληψίας που χρησιμοποιείται για την εξαγωγή και ερμηνεία του θέματος και μια διαδρομή δειγματοληψίας η οποία επιτρέπει τον εντοπισμό στοιχείων.



Εικ. 30. Η αριστερή εικ. είναι η αρχική. Η δεξιά είναι η βασική αλήθεια της τμηματοποίησης σελίδας. Τα χρώματα αντιπροσωπεύουν το κύριο κείμενο, τα σχόλια και τη διακόσμηση.

Ένα από τα πεδία που χρησιμοποιείται η μέθοδος FCNN είναι στην τμηματοποίηση με βάση τα εικονοστοιχεία σε ιστορικά έγγραφα που χρησιμοποιούν πλήρη συγκροτήματα δικτύων FCN. Η εικόνα του εγγράφου χωρίζεται σε διαφορετικές περιοχές με την ταξινόμηση των εικονοστοιχείων σε διαφορετικές κατηγορίες με βάση: φόντο, κυρίως κείμενο, σχόλια και διακοσμήσεις. Δίνει τη δυνατότητα εξαγωγής διακριτών χαρακτηριστικών και να εκτελέσει με ακρίβεια τμηματοποίηση των εικονοστοιχείων. Μετά την ταξινόμηση βάσει εικονοστοιχείων ακολουθούνται διαδικασίες για τη μείωση του θορύβου, τις διορθώσεις λαθών τμηματοποίησης και εντοπισμό επικαλυπτόμενων περιοχών. [XU Y., 2017]

Σε βάση δεδομένων που περιέχει μεσαιωνικά κείμενα όπως η DIVA-HisDB αναδεικνύεται η αποτελεσματικότητα και ανωτερότητα της χρήσης της μεθόδου FCNN η οποία έχει να δείξει ποσοστό ακρίβειας σε επίπεδο εικονοστοιχείων άνω του 99%. [XU Y., 2017]



Εικ.31. Δομή μοντέλου FCN

Η προτεινόμενη μέθοδος σύμφωνα με την έρευνα του XU [XU Y., 2017] έχει επιφέρει σημαντικές βελτιώσεις και τροποποιήσεις στο πεδίο τμηματοποίησης γενικών αντικειμένων, οπότε μας παρέχεται η προοπτική αντιμετώπισης των προκλήσεων τμηματοποίησης σελίδων με νέους όρους. Το FCN αποτελεί μια δοκιμή εκπαίδευσης από άκρο σε άκρο εντός πλαισίου τμηματοποίησης και ολοκληρώνει την εξαγωγή χαρακτηριστικών με ταυτόχρονη βελτιστοποίηση του ταξινομητή. Μόλις πριν λίγα χρόνια έλαβε χώρα η εφαρμογή της μεθόδου FCN για αντιμετώπιση προκλήσεων τμηματοποίησης σελίδων.

3.12 CNN και αναγνώριση κειμένου-σχολιασμοί

Στο παρόν κεφάλαιο θα γίνουν αναφορές στις προκλήσεις που αναφύονται στην αναγνώριση σχολίων που βρίσκονται σε χειρόγραφα με τη χρήση πλήρως συνελκτικών νευρωνικών δικτύων ή αλλιώς FCNN. Μια τέτοια μέθοδος έχει παρουσιαστεί σε έρευνα του Kolsch [KOLSCH A., 2018] και αφορά την αξιολόγηση της εν λόγω μεθόδου από βάση δεδομένων με ιστορικά γερμανικά χειρόγραφα του μεσαίωνα με ιδιαίτερες δυσκολίες. Ιδιαίτερα αφορά μεθόδους για τον εντοπισμό χειρόγραφων σχολιασμών σε αυτού του τύπου τα έγγραφα. Αυτοί μπορεί να εμφανίζονται σε μορφή υπογραμμίσεων και κειμένου με χρήση εργαλείων γραφής π.χ το μολύβι το οποίο καθιστά πολύ πιο δύσκολη την επεξεργασία και εξαγωγή δεδομένων.

Μια επιπλέον σημαντική παράμετρος σε χειρόγραφα κείμενα και στην προσπάθεια της αναγνώρισης κειμένου και της μεταγραφής τους είναι και τα πεδία σχολιασμού. Η δυνατότητα ανάλυσης χειρόγραφων σχολιασμών σε χειρόγραφα έντυπα συνιστά μια πολύ σημαντική πτυχή καθώς χάρη σε αυτήν δύναται να παρέχονται επιπρόσθετες πληροφορίες όσον αφορά στην ευκρίνεια και ανάλυση του κύριου σώματος του κειμένου αλλά και όσον αφορά τις σκέψεις των αναγνωστών αναφορικά με τα εξεταζόμενα κείμενα.

Προκειμένου να εξασφαλιστεί μια πιο επιτυχημένη αναγνώριση γραφής σε εικόνες εγγράφων, απαραίτητο κρίνεται να υπάρχει αγωγός επεξεργασίας πολλαπλών σταδίων. Μάλιστα, πριν ο συγκεκριμένος αλγόριθμος αναλύσει σε χειρόγραφους σχολιασμούς σε ένα έγγραφο θα πρέπει αρχικά να εντοπιστούν τα ακριβή σημεία που περιέχονται τα εν λόγω χειρόγραφα. Πρωταρχικό στάδιο αποτελεί η δημιουργία ενός επαρκώς ισχυρού αλγόριθμου που μπορεί να εντοπίσει αυτές τις περιοχές. Ως γεγονός, η δημιουργία ενός τέτοιου αλγορίθμου είναι μια ιδιαίτερη

δύσκολη και απαιτητική υπόθεση καθώς η είναι τεράστια η διάταξη και η ποικιλία των εγγράφων που θα πρέπει να διεκπεραιώσει προκειμένου να εντοπίσει αξιόπιστους χειρόγραφους σχολιασμούς. [KOLSCH A., 2018]

Παλαιότερες προσπάθειες στο πεδίο εντοπισμού χειρόγραφων σχολιασμών επικέντρωναν μέσω τμηματοποίησης εγγράφων και περιοχών από πάνω προς τα κάτω και το αντίστροφο, όπως επίσης και μέσω μεθόδων αφής [ASI A., 2015]

Για την επίτευξη της σημασιολογικής τμηματοποίησης υπάρχουν οι μέθοδοι μέσω της χρήσης αρχιτεκτονικών δικτύου, βασιζόμενες σε συγκροτήματα νευρωνικών δικτύων CNN και μέσω μακροπρόθεσμης μνήμης νευρωνικών δικτύων LSTM [Long J., 2015]. Η αποτελεσματικότητα των παραπάνω προσεγγίσεων δεν έχει αμφισβητηθεί, όμως ιδίως η χρήση των συγκροτημάτων νευρωνικών δικτύων είναι μεγαλύτερη και πιο δημοφιλής λόγω αποτελεσματικότητας.

Ειδικότερα δε, μέσω της χρήσης δικτύων όπως τα πλήρη συγκροτήματα νευρωνικών δικτύων αποτέλεσαν μια τομή στο πεδίο της τμηματοποίησης ιστορικών εγγράφων αφήνοντας πίσω τους άλλες μέχρι πρόσφατα μεθόδους. Οι συνήθεις προσεγγίσεις στην σημασιολογική τμηματοποίηση εγγράφων βασίζονταν σε μεθόδους μηχανικής μάθησης εφαρμοσμένες σε χειροποίητο υλικό. [CHEN K., 2017]

Στην έρευνα του Kolsch [KOLSCH A., 2018] η οποία βασίστηκε πάνω στη χρήση FCNN χρησιμοποιήθηκε ως βάση δεδομένων για τα πειράματα, την εκπαίδευση, την συγκριτική αξιολόγηση και χειρόγραφη τμηματοποίηση σχολιασμών, 40 εικόνες για εκπαίδευση και επικύρωση και 10 εικόνες για δοκιμή. Το υλικό σε φορμάτ PAGE αντλήθηκε από βάσεις δεδομένων από Πανεπιστήμια και Βιβλιοθήκες της Γερμανίας όπως:

- Universitätsbibliothek der Humboldt-Universität του Βερολίνου
- Universitätsbibliothek Kassel
- Staatsbibliothek Βερολίνου

- Staatsarchiv Marburg

Οι εικόνες λαμβάνονται από πολλά χειρόγραφα με ποικιλία διατάξεων και διάφορα είδη σχολιασμών. Σε κάθε εικόνα υπάρχει αλήθεια εδάφους με δύο εικονοστοιχεία. Την χειρόγραφη σημείωση και το φόντο.

Υπάρχουν επίσης και άλλες προσεγγίσεις στη σημασιολογική τμηματοποίηση εγγράφων όπως στον CHEN [CHEN K., 2017] όπου η εξαγωγή χρωμάτων και υλικού προέρχονται από χειροποίητες αναπαραστάσεις χαρακτηριστικών έγχρωμων εικόνων τα οποία δεδομένα στη συνέχεια προωθούνται σε έναν ταξινομητή SVM (Support Vector Machine). Η παραπάνω μέθοδος δεν απαιτεί καμία προεπεξεργασία όπως η διαδικοποίηση για την λειτουργία της σε έγχρωμες εικόνες.

Η χρήση της βαθιάς μάθησης ήταν μια εξαιρετικά επιτυχημένη μέθοδος για αρκετές από τις προσπάθειες που είχαν να κάνουν με την υπολογιστική όραση και την αναγνώριση χαρακτηριστικών προτύπων ανάλυση εγγράφων.

Πρόσφατες προσεγγίσεις στην τμηματοποίηση ιστορικών εγγράφων χρησιμοποιούν χαρακτηριστικά που εξάγονται από αρχιτεκτονικές βάσης μάθησης όπως οι αυτόματοι κωδικοποιητές. Χάρη σε αυτού του τύπου τις αρχιτεκτονικές τα χαρακτηριστικά από δεδομένα που δεν φέρουν ετικέτα τα μαθαίνουν αυτόματα. Τα μελλοντικά έργα στοχεύουν στο σχεδιασμό πιο αποτελεσματικών μεθόδων, με μεγαλύτερη ακρίβεια στο πλαίσιο της τμηματοποίησης και με πιο δίκαιες μετρήσεις αξιολόγησης.

ΚΕΦ. 4ο: ΕΦΑΡΜΟΓΕΣ-ΠΛΑΤΦΟΡΜΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΚΕΙΜΕΝΟΥ

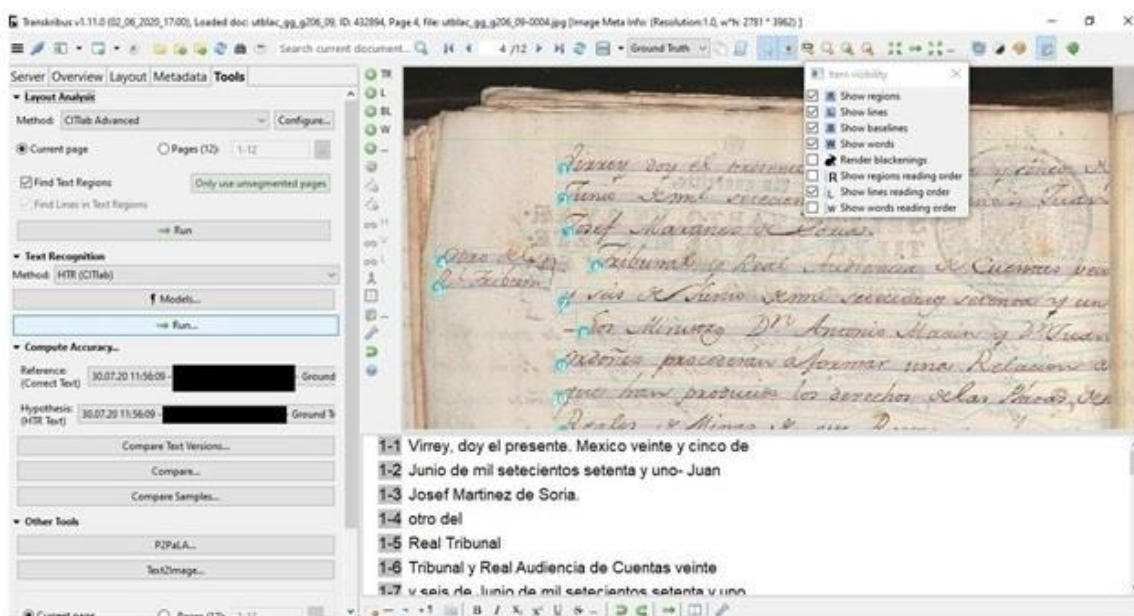
4.1 ΠΛΑΤΦΟΡΜΑ ΑΥΤΟΜΑΤΗΣ ΜΕΤΑΓΡΑΦΗΣ ΧΕΙΡΟΓΡΑΦΩΝ TRANSCIBUS

Στο παρόν κεφάλαιο θα γίνει μια προσπάθεια περιγραφής από μια σειρά από πλατφόρμες αυτόματης αναγνώρισης χειρόγραφου υλικού και ο τρόπος λειτουργίας και χρήσης τους όπως το transkribus η οποία αποτελεί ένα τέτοιο εξαιρετικό εργαλείο. Περιλαμβάνει φιλικό περιβάλλον εργασίας στο διαδίκτυο.

Το Transcribus αναπτύχθηκε σαν γλωσσικό εργαλείο για αυτόματη παραγωγή χειρόγραφου κειμένου το οποίο περιλαμβάνει μια σειρά από εργαλεία των οποίων ο συνδυασμός αποσκοπεί στο πλέον βελτιστοποιημένο αποτέλεσμα μεταγραφής ενός χειρόγραφου κειμένου του μεσαιώνα. Τα εργαλεία που περιλαμβάνει είναι: αναγνώριση κειμένου HTR, ανάλυση διάταξης, κατανόηση κειμένου, αναγνώριση συγγραφέα και τέλος εργαλείο οπτικής αναγνώρισης χαρακτήρων OCR χρησιμοποιώντας το ABBYY Finereader Engine 11. Σημαντικό στοιχείο αναφοράς είναι πως το Transcribus χρηματοδοτείτο από την Ευρωπαϊκή Επιτροπή ως μέρος του project READ <https://readcoop.eu/> αλλά και TranScriptorium. Μέσω αυτών των προγραμμάτων ευρωπαϊκής χρηματοδότησης ερευνητές και επιστήμονες της πληροφορικής κατόρθωσαν να δημιουργήσουν και να διαθέσουν ελεύθερα, καινοτόμες εφαρμογές αυτόματης μεταγραφής και ευρετηρίου, ενισχύοντας τις ερευνητικές δυνατότητες της Ευρώπης. Η πλατφόρμα Transcribus που έδωσε τη δυνατότητα στους χρήστες του να επιτελέσουν εργασίες μεταγραφής κειμένων ήταν διαθέσιμη για εγκατάσταση από την ιστοσελίδα tranScriptorium <http://transcriptorium.eu> μόλις το 2015. [MILIONI, 2020]

Ειδικότερα δε αφορά μια ολοκληρωμένη πλατφόρμα αναγνώρισης κειμένου, ανάλυσης εικόνων και αναγνώρισης χειρόγραφων ιστορικών εγγράφων. Πρόκειται για την πρώτη διεπιστημονική πλατφόρμα στην οποία οι Βιβλιοθήκες και οι επιστήμονες ερευνητές των ανθρωπιστικών σπουδών αντλούν τα οφέλη των μοντέλων βαθέως νευρωνικών δικτύων που αναπτύσσονται από επιστήμονες της πληροφορικής. Αποτέλεσμα των παραπάνω δεν είναι μόνο η επεξεργασία μόνο των πρώιμων τυπωμένων βιβλίων, αλλά και κάθε είδους χειρόγραφο. Ιδίως δε σε ότι έχει να κάνει με την αυτόματη μεταγραφή, υπάρχει πληθώρα δεδομένων κυρίως για τις Δυτικές

γλώσσες, παρόλα ταύτα όμως σταδιακά οι δυνατότητες επεκτείνονται και στις μη Δυτικές Γλώσσες επί παραδείγματι στο Κυριλλικό αλφάβητο.



Εικ. 32. Εμφανίζεται η δυνατότητα ανάλυσης διάταξης στο πάνω μέρος του πίνακα στη δεξιά πλευρά της οθόνης. Δίνεται η δυνατότητα στους χρήστες να εκτελούν ανάλυση διάταξης μία σελίδα για κάθε φορά, ή έγγραφο κάθε φορά.

Η πλατφόρμα έχει χρησιμοποιηθεί σε πολλές έρευνες αυτόματης μεταγραφής και οι χρήστες του μπορούν να αντλήσουν τη χρήση αυτών των δεδομένων μέσω αυτής. Μάλιστα σε απαιτητικά κείμενα και σε Λατινική Γλώσσα του Μεσαίωνα τα ποσοστά σφάλματος χαρακτήρων με βάση το μοντέλο εκπαίδευσης, είναι σε χαμηλά ποσοστά αυξάνοντας την αξιοπιστία της παραγόμενης μεταγραφής. Ένα από τα πλέον αξιόλογα μέρη της πλατφόρμας είναι πως ενσωματώνει μια πλήρη ροή εργασίας κειμένου πράγμα το οποίο σημαίνει πως η διαδικασία της μεταγραφής μέσω της πλατφόρμας δεν απαιτεί την συμπληρωματική χρήση επιπλέον διαδικασιών και εργαλείων απλουστεύοντάς την τόσο από άποψη χρόνου αλλά και αξιοπιστίας αποτελέσματος.

Η τεχνολογία HTR που παρέχεται μέσω της πλατφόρμας Transcribus, χρησιμοποιεί γλώσσα προγραμματισμού JAVA και το εργαλείο ανοικτού κώδικα SWT. Η μεταγραφή ενός χειρογράφου μπορεί να επιτευχθεί για δύο κύριους βασικούς λόγους. Ο πρώτος είναι μια απλή μεταγραφή, η οποία επιτρέπει στους χρήστες να εκπαιδευτούν στην αυτόματη μηχανή του HTR και να

μελετήσουν ιστορικά χειρόγραφα έγγραφα. Το δεύτερο είναι μια προηγμένου τύπου μεταγραφή, η οποία σε αυτή την περίπτωση παρέχει στους χρήστες την δυνατότητα δημιουργίας της μεταγραφής ενός εγγράφου που μπορεί να χρησιμοποιηθεί ως βάση μιας ψηφιακής έκδοσης. Η μηχανή HTR βασίζεται σε αλγόριθμους της μηχανής μάθησης. Η εκπαίδευση κρίνεται απαραίτητη στην τεχνολογία με την παρουσίαση παραδειγμάτων τουλάχιστον 30 σελίδων μεταγραφέντος υλικού. Έτσι υποβοηθάτε η κατανόηση των σχημάτων που αποτελούν λέξεις και χαρακτήρες.

Το Transcribus μπορεί να δημιουργήσει αυτόματα μεταγραφές με ποσοστό σφάλματος χαρακτήρα 5-10%. Πρακτικά αυτό σημαίνει πως το 90-95% των χαρακτήρων στο επεξεργασμένο κείμενο θα είναι σωστό. [SEAWARD, 2017]

4.2 ΕΛΛΗΝΙΚΗ ΠΛΑΤΦΟΡΜΑ ΑΥΤΟΜΑΤΗΣ ΜΕΤΑΓΡΑΦΗΣ ΧΕΙΡΟΓΡΑΦΩΝ μDoc.ts

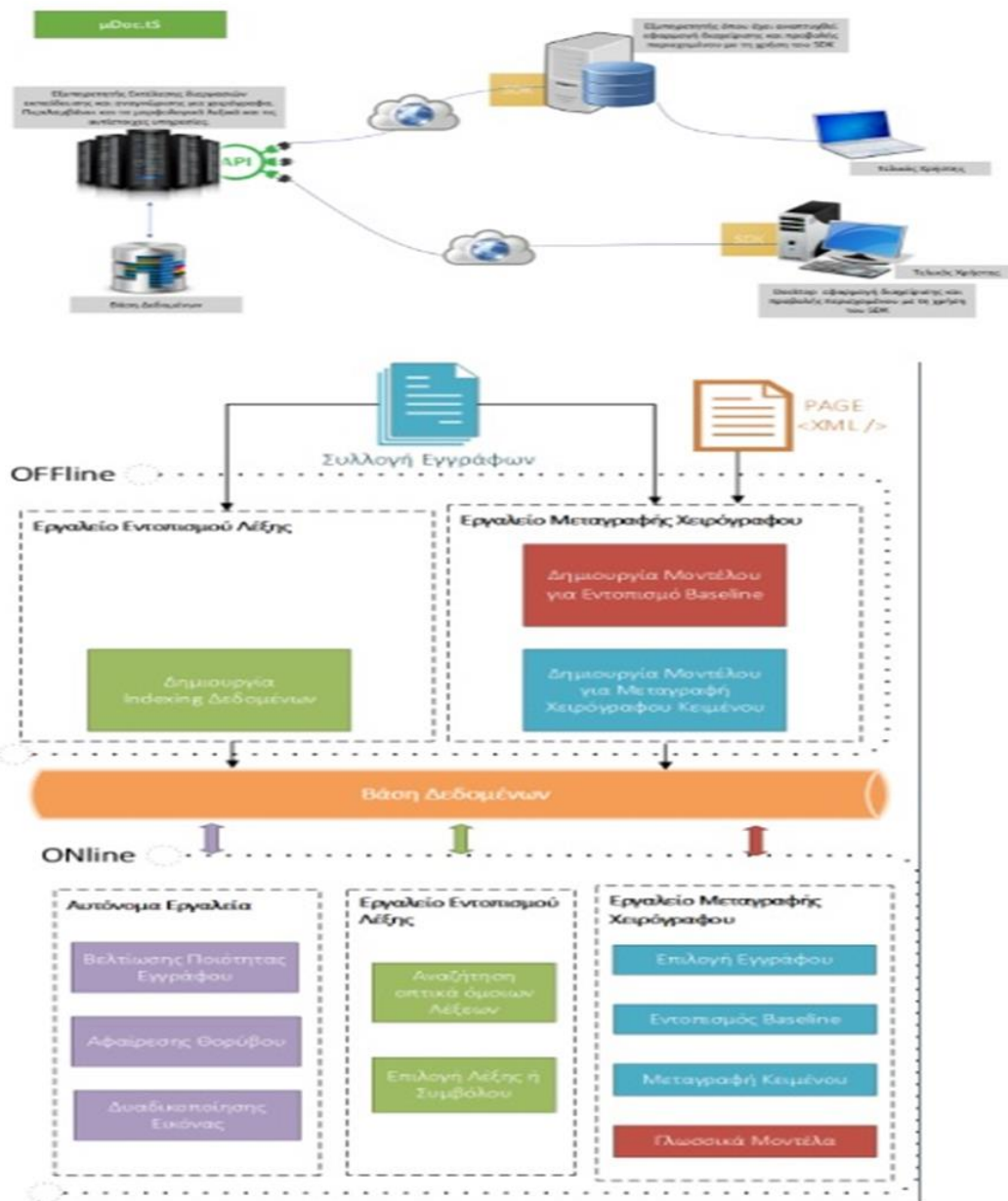
Στο προηγούμενο κεφάλαιο έγινε μια σύντομη παρουσίαση μιας από τις κορυφαίες και πιο γνωστές πλατφόρμες αυτόματης μεταγραφής χειρόγραφων κειμένων της Transcribus και εδώ θα γίνει μια αναφορά στην προσπάθεια δημιουργίας μιας αντίστοιχης ελληνικής πλατφόρμας μεταγραφής από το Δημοκρίτειο Πανεπιστήμιο Θράκης με την ονομασία μDoc.ts <https://mdoc-ts.ee.duth.gr/>. Πρόκειται για ένα πρόγραμμα το οποίο βρίσκεται σε στάδιο παράδοσης. Το έργο ξεκίνησε τον Ιούνιο του 2018 με ορίζοντα ολοκλήρωσης τον Ιούνιο του 2021. Το έργο είναι αποτέλεσμα σύμπραξης του Δημοκρίτειου Πανεπιστημίου μαζί με κρατικούς και ιδιωτικούς φορείς που δραστηριοποιούνται στον τομέα της Πληροφορικής.

Από τα χειρόγραφα θα μπορούν οι ερευνητές να χρησιμοποιήσουν υλικό αυτόματης επεξεργασίας και μεταγραφής, δίχως να χρειαστεί να φτάσουν στα ίδια τα πρωτότυπα, είτε να δουν εικόνες τους για τις οποίες θα πρέπει να έχουν ιδιαίτερες γλωσσολογικές παλαιογραφικές και άλλες ανάλογες γνώσεις για να μπορούν να ερμηνεύσουν τα γραφόμενα. Επομένως είναι ένα έργο ιδιαίτερος σημαντικό για τον ελληνικό χώρο καθώς η παραγωγή χειρογράφων και υλικού είναι τεράστια και πολύ περισσότερος ακαδημαϊκός και ερευνητικός κόσμος θα μπορεί να έχει πολύπλευρη πρόσβαση στο υπερπολύτιμο αυτό υλικό.

Πέρα από θέματα γλωσσολογικών γνώσεων και ικανοτήτων για κατανόηση των κειμένων πολλές φορές αυτά είναι δυσανάγνωστα και λόγω φθοράς του χρόνου αλλά και λόγω των διαφόρων ιδιαιτεροτήτων των κειμένων με την συνένωση γραμμών κειμένου, λέξεων, εμπότιση μελανιών αλλά και φθορά των υλικών πάνω στα οποία είναι αποτυπωμένα τα χειρόγραφα, αλλά και στις ιδιαιτερότητες των χαρακτήρων των γραμμάτων αναλόγως της χρονικής περιόδου που αυτά αποτυπώθηκαν αλλά και τα γραφικά στυλ αναλόγως του συγγραφέα.

Η δομή του περιλαμβάνει την δυνατότητα αυτόματης εξαγωγής κειμένου μέσω ενός συστήματος μεταγραφής με τη χρήση γλωσσικών μοντέλων. Τα εργαλεία που θα χρησιμοποιούνται θα έχουν ως στόχο να παρέχουν την μεγαλύτερη δυνατή ποιότητα στην μεταγραφή και την ποιότητα σε ψηφιοποιημένα χειρόγραφα κείμενα με μέσα που θα μπορούν να αφαιρέσουν τον θόρυβο, την δυαδικοποίηση αλλά και μέσω της ανάδειξης συγκεκριμένων περιοχών του κειμένου. Σημαντικό ρόλο θα έχει και η τεχνική της απευθείας αναζήτησης λέξεων χρησιμοποιώντας τις πλέον σύγχρονες μεθόδους εντοπισμού λέξεων (Keyword Spotting).

Σημαντικό επίσης ρόλο για την επιτυχία του έργου είναι η φιλικότητα προς το χρήστη και η δυνατότητα διαχείρισης των δεδομένων με τον πλέον εύκολο και αποτελεσματικό τρόπο. Μια λεπτομερής περιγραφή της πλατφόρμας αποτυπώνεται με τον καλύτερο τρόπο στην εικόνα που ακολουθεί.



Εικ. 33. Πλατφόρμα αναγνώρισης μ.Doc.ts

ΚΕΦΑΛΑΙΟ 5ο: ΠΙΝΑΚΕΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΥΓΚΡΙΤΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΕΦΑΡΜΟΓΩΝ ΣΕ ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΚΕΙΜΕΝΟΥ ΕΥΡΕΣΗΣ ΛΕΞΕΩΝ (WORD SPOTTING) ΚΑΙ ΑΝΑΓΝΩΡΙΣΗΣ ΧΕΙΡΟΓΡΑΦΟΥ ΚΕΙΜΕΝΟΥ (HANDWRITING TEXT RECOGNITION HTR)

5. Στο παρόν κεφάλαιο θα εξεταστούν και θα γίνουν αναφορές σε δύο ξεχωριστές ενότητες για την ακρίβεια και την αποτελεσματικότητά εφαρμογών και μεθόδων για την αναγνώριση κειμένου όπως στον εντοπισμό λέξεων σε χειρόγραφο κείμενο όσο και στην αναγνώριση χειρόγραφου κειμένου.

5.1 ΕΝΤΟΠΙΣΜΟΣ ΛΕΞΕΩΝ ΣΕ ΧΕΙΡΟΓΡΑΦΟ ΚΕΙΜΕΝΟ. ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 2016

Εντοπισμός λέξεων κλειδιών σε χειρόγραφο κείμενο είναι το έργο της εύρεσης των αναζητούμενων λέξεων σε χειρόγραφες συλλογές εικόνων οι οποίες δεν περιλαμβάνουν την διαδικασία μέσω οπτικής αναγνώρισης κειμένου. Τα τελευταία χρόνια η εύρεση λέξεων σε χειρόγραφο κείμενο φαίνεται να αποτελεί μια εφικτή λύση στον τομέα της ευρετηρίασης και ανάκτησης εγγράφων στην περίπτωση που οι μέθοδοι βασίζονται στην οπτική αναγνώριση δεν αποδίδουν ικανοποιητικά αποτελέσματα.

Οι μέθοδοι και τα αποτελέσματα που θα παρατεθούν θα αφορούν συνέδρια και διαγωνισμούς που διεξήχθησαν από τις πλέον πρόσφατες προς παλαιότερες χρονολογίες.

Αρχικά θα γίνει αναφορά σε συνέδριο του 2016 (ICFHR2016) [PRATIKAKIS I., 2016] στο οποίο κεντρικός στόχος είναι η δημιουργία μιας αξιολόγησης εντός του πλαισίου για συγκριτική αξιολόγηση χειρογράφου στον εντοπισμό λέξεων κλειδιών με βάση τόσο το ερώτημα με το παράδειγμα (Query by Example, QbE) όσο και το ερώτημα με προσεγγίσεις σειράς. (Query by String, QbS)

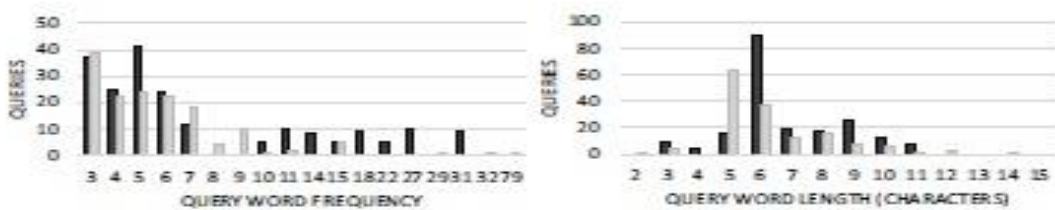
Το ερώτημα ανά παράδειγμα είναι μια μέθοδος δημιουργίας ερωτημάτων η οποία επιτρέπει στο χρήστη να αναζητά λέξεις σε έγγραφα βάσει ενός παραδείγματος με τη μορφή επιλεγμένης συμβολοσειράς κειμένου, λέξης, ή μορφή ονόματος εγγράφου. Από την άλλη στο ερώτημα με προσέγγιση σειράς, οι χαρακτήρες σε ένα αίτημα αναζήτησης αποστέλλονται σε ένα πρόγραμμα ερωτήσεων. Εκεί περιέχεται συνήθως η εντολή του ερωτήματος και τα δεδομένα αναζήτησης.

Οι συμβολοσειρές των ερωτημάτων χρησιμοποιούνται ευρέως στον ιστό του διαδικτύου για ερωτήματα μηχανών αναζήτησης καθώς και σε συνδέσμους με βάσεις δεδομένων.

Η κάθε μια από τις παραπάνω διακριτές μεθόδους εμπεριέχει τον δικό της βαθμό δυσκολίας και στοχεύουν σε διαφορετικούς στόχους. Για παράδειγμα στο ερώτημα με προσέγγιση σειράς η μέθοδος είναι πολύ σημαντική και απαιτείται σε εφαρμογές που περιλαμβάνεται ευρείας κλίμακας χειρόγραφη ευρετηρίαση και αναζήτηση εικόνων. Πλεονέκτημα επίσης μπορεί να αποτελέσει η εκπαιδευτική χρήση βασιζόμενη σε εύρεση λέξεων. Σε άλλο είδος εφαρμογών περιλαμβάνεται βοήθεια από ανθρώπινους μεταγραφείς στους οποίους ενυπάρχει η δυνατότητα να εξεύρουν λέξεις σε ένα έγγραφο οι οποίες έχουν σχήμα παρόμοιο με μια λέξη ή μέρος μιας λέξης στο ίδιο ή σε άλλο έγγραφο. Σε αυτή την περίπτωση το σύστημα αναζήτησης ερωτήματος με παράδειγμα δίχως εκπαίδευση, είναι το πλέον κατάλληλο. [PRATIKAKIS I., 2016]

Παρά το γεγονός πως οι μέθοδοι αναζήτησης με παράδειγμα και σειράς αντιμετωπίζουν διαφορετικά βασικά ζητήματα, σε τεχνικό επίπεδο είναι και τα δύο ενοποιημένα και επίσης μπορούν να είναι εξαρτώμενα σε ζητήματα βασισμένα σε κατάτμηση κειμένου ή όχι, να περιλαμβάνουν εκπαίδευση δεδομένων υπό επίβλεψη ή όχι.

Οι βάσεις δεδομένων πάνω στις οποίες στηρίχτηκαν οι εν λόγω μέθοδοι που θα περιγραφούν στη συνέχεια όσο και τα μεταξύ τους αποτελέσματα κατάταξης αποτελούνται από δύο διαφορετικές συλλογές που προετοιμάστηκαν στο ευρωπαϊκό έργο READ. Το Alvermann Konzilsprotokolle και η Βοτανική συλλογή της Βρετανικής Ινδίας. Η πρώτη συλλογή αφορά 18.000 σελίδες αντίγραφα πρακτικών των επίσημων συνεδριάσεων της κεντρικής διοίκησης μεταξύ των ετών 1794-1797 και η δεύτερη συλλογή από τα Office Records της Βρετανικής Ινδίας και παρέχεται από την Βρετανική Βιβλιοθήκη.



Εικ. 34. Στατιστικά ερωτήματος λέξης για Βοτανική (ανοιχτόχρωμο) και Konzilsprotokolle (σκούρο): αριστερά συχνότητα, δεξιά μήκος

Η εννοούμενη μέση ακρίβεια (mean average Precision mAP,) είναι ο αλγόριθμος που χρησιμοποιήθηκε για την αξιολόγηση της εμφάνισης των αποτελεσμάτων των συμμετεχόντων και των μεθόδων. Για κάθε ερώτημα στο σύνολο Q , η (παρεμβλλόμενη) μέση ακρίβεια υπολογίστηκε χρησιμοποιώντας την (παρεμβολή) ακρίβεια στην κορυφή k , $\pi(k)$ ($\hat{\pi}(k)$), και την ανάκληση στην κορυφή- k , $\rho(k)$.

Σε ότι αφορά στον ανταγωνισμό με την χωρίς τμηματοποίηση πρόκληση, εκεί υπάρχει ένα εντοπισμένο οριοθετημένο πλαίσιο που μπορεί να μην ταιριάζει με τις αναφορές. Το εντοπιζόμενο πλαίσιο που θεωρείται πως έχει σωστή αντιστοίχιση είναι όταν η σχετική αλληλεπικάλυψη περιοχής έχει πλαίσιο αναφοράς που είναι μεγαλύτερο ή ίσο με 0,5 και έχει την ίδια ετικέτα με την αναφορά.

Ακολουθούν σύντομες περιγραφές των μεθόδων που ακολούθησαν οι συμμετέχοντες και στη συνέχεια τα αποτελέσματα και οι πίνακες.

Η ομάδα του Computer Vision Center (CVCDAG), Universitat Autònoma de Barcelona, Spain ακολουθεί την εξής μέθοδο. Εικόνες λέξεων κωδικοποιούνται αρχικά σε χαρακτηριστικά διανύσματα χρησιμοποιώντας το Διάνυσμα Fischer. Εν συνεχεία τα διανύσματα αυτών των χαρακτηριστικών χρησιμοποιούνται μαζί με ένα πυραμιδικό ιστόγραμμα ετικετών χαρακτήρων (PHOC) για να γίνει μάθηση του μοντέλου των χαρακτηριστικών με βάση μηχανή διανυσματικής υποστήριξης. (Support Vector Machine, SVM)

Συγκεκριμένα πρόκειται για γραμμικούς ταξινομητές οι οποίοι βασίζονται στην αρχή μεγιστοποίησης του περιθωρίου. Με την ελαχιστοποίηση του διαρθρωτικού κινδύνου βελτιώνεται η πολυπλοκότητα του ταξινομητή με στόχο την επίτευξη εξαιρετικής απόδοσης γενίκευσης. Το (PHOC) το πυραμιδικό ιστόγραμμα χαρακτήρων είναι ένα δυαδικό ιστόγραμμα που κωδικοποιεί το εάν ένας συγκεκριμένος χαρακτήρας εμφανίζεται στην αναπαριστώμενη λέξη ή όχι. Η αναπαράσταση της χωρικής πυραμίδας διασφαλίζει ότι διατηρούνται οι πληροφορίες της σειράς χαρακτήρων. Η αναπαράσταση PHOC αποτελεί την συνένωση μερικών εκ των ιστογραμμάτων.

Ως βασική αναπαράσταση είναι ένα δυαδικό ιστόγραμμα χαρακτήρων, με κωδικοποίηση για το ποιοι χαρακτήρες εμφανίζονται στην συμβολοσειρά. Τα νέα επίπεδα προστίθενται στο ιστόγραμμα με πυραμιδικό τρόπο, όπου σε κάθε νέο επίπεδο η λέξη χωρίζεται περεταίρω και ένα νέο ιστόγραμμα χαρακτήρων προστίθεται για κάθε νέο τμήμα σε διαφορετικά μέρη της λέξης. Τελικά, γίνεται χρήση 5 επιπέδων με αναπαράσταση της λέξης σε 604 διαστάσεις. Στο τέλος της διαδικασίας έρχεται να υποστηριχθεί από μηχανή διανυσματικής υποστήριξης.

Η ομάδα Pattern Recognition Group (PRG), TU Dortmund University, Germany η μέθοδος που χρησιμοποίησε τον εντοπισμό λέξεων είναι του PHOCNet. Πρόκειται για ένα συνελικτικό νευρωνικό δίκτυο 19 επιπέδων το οποίο είναι ειδικά σχεδιασμένο για εκμάθηση χαρακτηριστικών εικόνας εγγράφου. Το βαθύ νευρωνικό δίκτυο εκπαιδεύτηκε σε συνδυασμό με το PHOC για να υπάρχει η δυνατότητα αναπαράστασης της κάθε εικόνας λέξης. Το PHOCNet μπορεί να προβλέψει μια δεδομένη λέξη και εικόνα δίχως να χρειάζεται να επανασυνδέσει ή να περικόψει την εικόνα. Για την κατάτμηση εργασίας, οι λέξεις υποβάλλονται σε επεξεργασία από το συνελικτικό νευρωνικό δίκτυο. Για τη χωρίς τμηματοποίηση μέθοδο, ένα συρόμενο παράθυρο πάνω από τις εικόνες εγγράφων χρησιμοποιούνται για να εξάγουν το PHOC από τη θέση κάθε παραθύρου.

Για το ερώτημα με παράδειγμα, κάθε ερώτημα αντιστοιχεί στο πλησιέστερο προεπιλεγμένο μέγεθος και η ανάκτηση πραγματοποιείται με αυτό το μέγεθος. Για ερώτημα με σειρά η εικόνα λέξης της εκπαίδευσης χρησιμοποιήθηκε ως ερώτημα για ανάκτηση.

Σε επόμενη ομάδα εκείνης του Visual Information and Interaction (QTOB), Uppsala University, Sweden, ένα συνελικτικό δίκτυο έχει λάβει εκπαίδευση για να είναι ικανό να εξαγάγει μια αναπαράσταση εικόνας χρησιμοποιώντας προσέγγιση τριπλού δικτύου [6] κατά την οποία ο περιγραφέας για μια εικόνα λέξης, προσπαθεί να μαθαίνει το να μπορεί να προβλέψει το αν λέξεις ανήκουν ή δεν ανήκουν στην ίδια κατηγορία τάξης. Στην αρχιτεκτονική ConvNet χρησιμοποιείτε το δίκτυο ResNet με 34 διαστρωματώσεις επιπέδων. Το ResNet συντομογραφία για το Residual Network είναι συγκεκριμένος τύπος νευρωνικού δικτύου που παρουσιάστηκε το 2015 σε εργασία από τους HE, ZHANG, REN, SUN. Η επιτυχία του εν λόγω μοντέλου απεδείχθη εξαιρετικά επιτυχημένη. Μόλις οι εικόνες των λέξεων ενσωματωθούν είτε με τη μέθοδο ερώτημα με παράδειγμα, είτε ερώτημα προς λέξη συμβολοσειρά, ο εντοπισμός της λέξης μπορεί να εκτελεστεί στην λέξη που είναι ενσωματωμένη στο χώρο από μέσα του συνημιτονου της απόστασης. Το συνημίτονο της απόστασης αφορά μόνο θετικές αξίες. Η αύξηση δεδομένων χρησιμοποιήθηκε έτσι ώστε να μεγαλώσει το μέγεθος του σετ κατάρτισης εφαρμόζοντας απλή γεωμετρική και μορφολογική πράξη στα ήδη υπάρχοντα εκπαιδευτικά δεδομένα. Όλα τα μοντέλα εκπαιδεύτηκαν σε NVIDIA GTX.

Στο ICFHR2016 ακολουθεί μια τελευταία ομάδα εκείνης του Tel Aviv University (TAU), Israel όπου και εδώ χρησιμοποιείται και στις δύο προκλήσεις του ερωτήματος με παράδειγμα (QbE). Στην τμηματοποίηση δίχως βάση, το πρώτο βήμα είναι η εξαγωγή υποψήφιας λέξης από τις σελίδες των εγγράφων που γίνεται η επεξεργασία. Οι εικόνες είναι δυαδικές και τα συνδεδεμένα στοιχεία φιλτράρονται σε μεγαλύτερα ή μικρότερα μέρη. Η ίδια διαδικασία επίσης ακολουθείται και στην πρόκληση με βάση την τμηματοποίηση. Για να βελτιωθεί η απόδοση της μεθόδου λαμβάνεται υπόψη η κάθε εικόνα του ερωτήματος, περισσότερες από μία φορές, αλλάζοντας την αρχική εικόνα σε σταθερό αριθμό pixel προς τις τέσσερις κατευθύνσεις.

Πίνακας 1: Τα καλύτερα αποτελέσματα με μέση ακρίβεια μέσης ακρίβειας (PmAP) σε κάθε κομμάτι: (α) Ερώτημα με παράδειγμα και (β) ερώτημα με σειρά. Τελευταία στήλη τελικό σκορ, καλύτερο αποτέλεσμα καθε προκλήσης με έντονη γραφή.

(a) Track I: Query by Example							
Team	Segm. based			Segm. free			Final Score
	Botany	Konzil.	Average	Botany	Konzil.	Average	
CVCDAG	75.77	77.91	76.84	0.21	0.0	0.10	76.86
PRG	46.61	88.14	67.38	15.89	52.20	34.05	74.18
TAU	50.64	71.11	60.87	37.48	61.78	49.63	70.80
QTOB	54.95	82.15	68.55	—	—	—	68.55

(b) Track II: Query by String							
Team	Segm. based			Segm. free			Final Score
	Botany	Konzil.	Average	Botany	Konzil.	Average	
PRG	36.47	76.93	56.70	11.80	48.41	30.10	62.72
CVCDAG	65.69	55.27	60.48	—	—	—	60.48
QTOB	3.40	12.19	7.79	—	—	—	7.79

Πίνακας 2:Βλεπουμε τα καλύτερα αποτελέσματα μέσης ακρίβειας από κάθε ομάδα, η περίοδος υποβολής και το καλύτερο αποτέλεσμα επισημαίνεται με έντονη γραφή. Μέρος α το Query by Example, και β το Query by String.

(a) Track I: Query by Example								
Team	Botany	Segm. based			Botany	Segm. free		
		Konzil.	Average	Period		Konzil.	Average	Period
CVCDAG	75.77	77.91	76.84	I	0.42	0.0	0.21	III
PRG	89.69	96.05	92.87	III	15.89	52.20	34.05	I
TAU	50.64	71.11	60.87	I	37.48	61.78	49.63	I
QTOB	54.95	82.15	68.55	I	—	—	—	—

(b) Track II: Query by String								
Team	Botany	Segm. based			Botany	Segm. free		
		Konzil.	Average	Period		Konzil.	Average	Period
PRG	74.47	94.20	84.34	III	11.80	48.41	30.10	I
CVCDAG	65.69	82.91	74.30	II	—	—	—	—
QTOB	3.40	12.19	7.79	I	—	—	—	—

Στις παραπάνω μεθόδους χρησιμοποιήθηκαν οι ίδιες λέξεις ερωτήματος για την αξιολόγηση QbE και QbS. Ειδικότερα για την περίπτωση QbE για εντοπισμό αρκετών από αυτές τις λέξεις, πολλά ερωτήματα εικόνων ανα ερώτημα λέξης επιλέχτηκαν χειροκίνητα και χρησιμοποιήθηκαν ως παραδείγματα ερωτημάτων. Στο ζήτημα που αφορά το QbS δεν υπάρχει νόημα να γίνεται επανάληψη του ίδιου ερωτήματος. Ως αποτέλεσμα στο μέρος του QbS όλες οι λέξεις του ερωτήματος έχουν τον ίδιο αντίκτυπο. Αντίθετα στο μέρος του QbE οι λέξεις με περισσότερα παραδείγματα έχουν τον μεγαλύτερο αντίκτυπο.

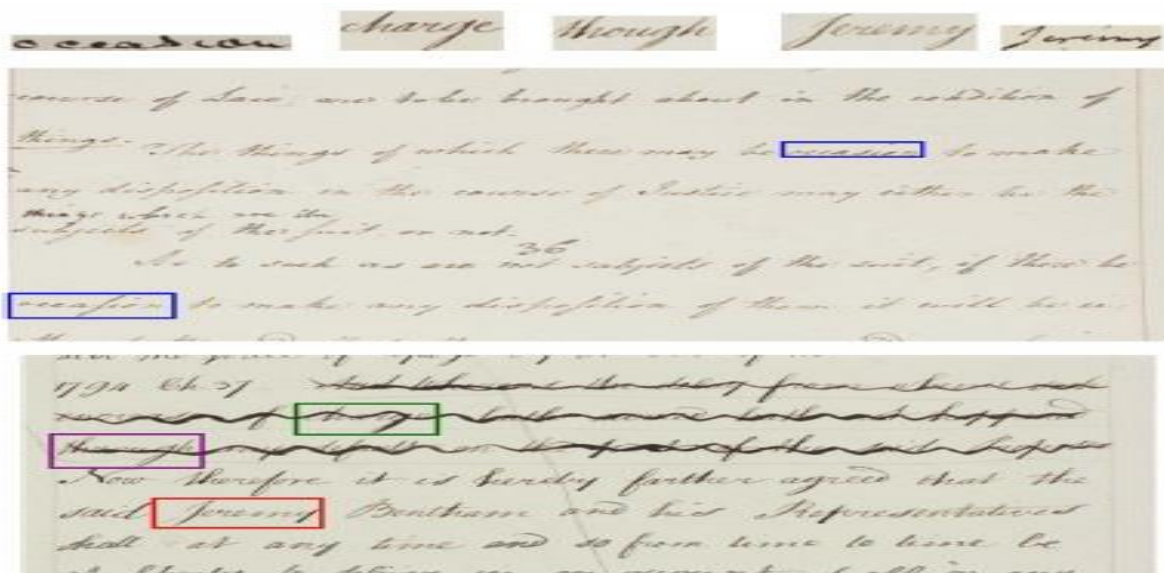
Όπως φαίνεται πιο πάνω στην εικόνα 6* στο 90% περίπου του ερωτήματος, οι εικόνες αντιστοιχούν σε μεγάλες λέξεις με 5 ή περισσότερους χαρακτήρες οι οποίοι μάλιστα εντοπίζονται ευκολότερα εκτός των συστημάτων QbE τόσο και από τα QbS. Συγκρινόμενη είναι υψηλότερη θετική επίδραση στα συστήματα QbE παρά στα QbS. Με βάση τα αποτελέσματα και τα δεδομένα που χρησιμοποιήθηκαν θεωρήθηκε πως για να είναι πιο ρεαλιστικές οι συνθήκες προσομοίωσης και δυσκολίας η ποσότητα των λέξεων ερωτήματος πρέπει να κατά πολύ πιο μεγαλύτερη. Με χρήση μερικών χιλιάδων μη επιλεγμένων λέξεων θα υπάρχει η δυνατότητα να εξασφαλιστεί τόσο ο απαιτούμενος βαθμός μεταβλητότητας όσο και οι αντικειμενικές δυσκολίες. [PRATIKAKIS I., 2016]

5.1.2 ΕΝΤΟΠΙΣΜΟΣ ΛΕΞΕΩΝ ΣΕ ΧΕΙΡΟΓΡΑΦΟ ΚΕΙΜΕΝΟ. ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 2015

Στη συνέχεια θα καταγραφούν οι μέθοδοι τα αποτελέσματα και οι βάσεις δεδομένων που χρησιμοποιήθηκαν σε διαγωνισμό συνεδρίου του 2015 (ICDAR2015). Τα κύρια χαρακτηριστικά αυτής της ερευνητικής διαδικασίας αφορούσαν τον έλεγχο των μεθόδων σχετικά με ελεύθερη εκπαίδευση βασισμένη τόσο στην τμηματοποίηση και στον εντοπισμό λέξεων κλειδιών με βάση το ερώτημα με το παράδειγμα (Query by Example, QbE) όσο και δίχως να βασίζεται στην κατάτμηση κειμένου. [PUIGCERVER J., 2015]

Η δε επόμενη έρευνητική διαδικασία αφορά τις μεθόδους που βασίζονται στην εκπαίδευση και στην μη κατάτμηση κειμένου τόσο με βάση το ερώτημα εντοπισμού λέξης με βάση το παράδειγμα (Query by Example, QbE) όσο και στο ερώτημα με προσεγγίσεις σειράς. (Query by String, QbS,).

Η βάση δεδομένων πάνω στις οποίες εκπαιδεύτηκαν και προσαρμόστηκαν οι μέθοδοι των όσων συμμετείχαν στην διαδικασία προήλθε από μια σειρά χειρόγραφων εγγράφων με συγγραφέα τον Άγγλο φιλόσοφο Jeremy Bentham χρονολογίας του 16^{ου} και 17^{ου} αιώνα και τα οποία έγγραφα ψηφιοποιήθηκαν στο πλαίσιο του έργου transcriptorium. Από τα παραπάνω έγγραφα αντλήθηκαν 15.419 τμηματοποιημένες εικόνες λέξεων. [PUIGCERVER J., 2015]



Εικ. 35. Παραδείγματα εικόνων ερωτημάτων και τμημάτων δύο εικόνων που περιέχουν τις λέξεις "Occasion" σε μπλέ περίγραμμα, "charge" σε πράσινο, "through" σε μωβ και "Jeremy" με κόκκινο.

Η αξιολόγηση των δεδομένων γίνεται επίσης με βάση την εννοούμενη μέση ακρίβεια (mean average Precision mAP,) καθώς πρόκειται για τον χρησιμοποιούμενο αλγόριθμο.

Το Pattern Recognition Group (PRG), TU Dortmund University, Germany, στην εργασία χωρίς τμηματοποίηση την διενέργησε με δυαδικοποίηση Otsu. Με την μέθοδο αυτή ο αλγόριθμος επιστρέφει ένα όριο έντασης με το διαχωρισμό του εικονοστοιχείου σε δύο κατηγορίες. Στο προσκήνιο και στο φόντο. Οι περιγραφείς κάθε λέξης ποσοτικοποιούνται με βάση το βιβλίο κωδικών και εξάγεται μια χωρική πυραμίδα σε δύο επίπεδα. Το πρώτο επίπεδο χωρίζει την εικόνα σε 3 ίσες διατομές κατά το μήκος της κατεύθυνσης της γραφής ενώ στο δεύτερο η εικόνα χωρίζεται σε 9 τμήματα ίσου μεγέθους κατά την ίδια κατεύθυνση. Για κάθε ερώτημα οι δοκιμαστικές λέξεις ταξινομούνται με βάση την απόσταση Bray Curtis στις αντίστοιχες χωρικές πυραμίδες.

Από την άλλη η ομάδα του Computer Vision Center (CVC), Universitat Autònoma de Barcelona, Spain, χρησιμοποιεί εικόνες λέξεων που αντιπροσωπεύονται από έναν περιγραφέα χρησιμοποιώντας το πλαίσιο του «σάκου των λέξεων» (Bag of Visual Words, BoVW). Η βασική ιδέα του μοντέλου είναι στην αντιπροσώπευση των εγγράφων, είτε αυτά είναι ιστοσελίδες

αρχεία word κλπ, ως μία συλλογή σημαντικών σημείων, ενώ από την άλλη αγνοεί παντελώς τη σειρά εμφάνισης των λέξεων. Ειδικότερα χρησιμοποιήθηκε μια οπτική αναπαράσταση όπου οι τοπικές περιοχές του κειμένου είναι εικόνα πυκνής δειγματοληψίας σε ένα σταθερό βήμα της τάξεως των 3 εικονοστοιχείων και σε 4 διαφορετικές κλίμακες μεγεθών των 16, 24, 32 και 40 εικονοστοιχείων.

Σε επόμενο στάδιο για επόμενα βήματα η επεξεργασία γίνεται ανεξάρτητα. Οι τοπικές περιοχές χαρακτηρίζονται από το ολοκληρωμένο ιστόγραμμα του τοπικού περιγραφέα κλίσης χρησιμοποιώντας διαμόρφωση $8 \times 4 \times 4$ 128 διαστάσεων. Οι περιγραφείς μετατρέπονται σε οπτικές λέξεις χρησιμοποιώντας k ως μέσα διανυσματικής κβαντοποίησης με μέγεθος βιβλίου κωδικών 2048. Οι τοπικοί περιγραφείς κωδικοποιούνται σε οπτικές λέξεις με χρήση γραμμικής κωδικοποίησης περιορισμένης τοποθεσίας.

Ο οπτικός περιγραφέας λαμβάνεται συνενώνοντας τα ιστογράμματα των οπτικών λέξεων κάθε πυραμιδοειδούς επιπέδου. Αυτή η διαμόρφωση οδηγεί σε οπτική περιγραφή 43008 διαστάσεων ανά κλίμακα. Οι εικόνες λέξεων στη συλλογή και οι εικόνες ερωτημάτων όπως και η απόσταση του Ευκλείδη συνηθίζεται να δίνει μια κατάταξη με αποτελέσματα. Σε κάθε ερώτημα επιστρέφονται 1400 αναφορές.

Τέλος η ομάδα του «Computational Intelligence Technology Lab. (CITlab), University of Rostock, Germany» το οποίο βασίζει την μέθοδό του στο σύστημα CITlab το οποίο βασίζεται σε ένα επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) και στη μηχανή αναγνώρισης που ονομάζεται ARGUS το οποίο σύστημα αναπτύχθηκε από κοινού από την PLANET intelligent systems GmbH. Και στις δύο περιπτώσεις των εργασιών η προσέγγιση είναι κατά βάση παρόμοια. Η μόνη διαφορά εστιάζεται στο δεύτερο σκέλος της 2^{ης} εργασίας όπου στον εντοπισμό με βάση το παράδειγμα (QbE), όπου η αναπαράσταση συμβολοσειράς της εικόνας του ερωτήματος εξάγεται πρώτη και ακολουθείται η προσέγγιση με βάση τον εντοπισμό της συμβολοσειράς (QbS).

Σε πρώτο στάδιο οι γραμμές κειμένου εξάγονται από τις εικόνες του παρεχόμενου εγγράφου. Οι εικόνες γραμμής του κειμένου έχουν υποβληθεί σε επεξεργασία μέσω του συστήματος ομαλοποίησης γραφής του CITlab σε συνδυασμό με την ομαλοποίηση του κειμένου ως προς το μέγεθος στις διορθώσεις και στην κλίση των γραμμών. Έτσι οι εικόνες πλέον είναι σε μέγεθος 96 εικονοστοιχείων με το κυρίως σώμα του κειμένου να είναι κατάλληλα τοποθετημένο αλλά και τραβηγμένο ώστε να καλύψει το απαραίτητα κεντρικό τμήμα της γραμμής της εικόνας. Οι εικόνες γραμμής κειμένου που προκύπτουν τροφοδοτούνται σε ένα επαναλαμβανόμενο νευρωνικό δίκτυο χωρίς όποια περεταίρω επεξεργασία τμηματοποίησης.

Το αποτέλεσμα της επεξεργασίας μέσω της χρήσης επαναλαμβανόμενου νευρωνικού δικτύου, υπολογίζει το πίσω μέρος της κατανομής των χαρακτήρων του αλφαβήτου. Το αλφάβητο περιέχει όλα εκείνα τα ψηφία, πεζά αλλά και κεφαλαία του τυπικού λατινικού αλφαβήτου, σημεία στίξης και άλλων ειδικών χαρακτήρων και επίσης κενό διάστημα αλλά και ένα ειδικό σύμβολο δίχως χαρακτήρες για να εντοπίζει τα όρια των χαρακτήρων στην έξοδο του RNN. Αυτή, τροφοδοτείται σε έναν αποκωδικοποιητή ο οποίος κάνει αναζητήσεις για μέχρι τέσσερις περιοχές της γραμμής και είναι πιθανό να περιλαμβάνει και τη λέξη κλειδί του ερωτήματος.

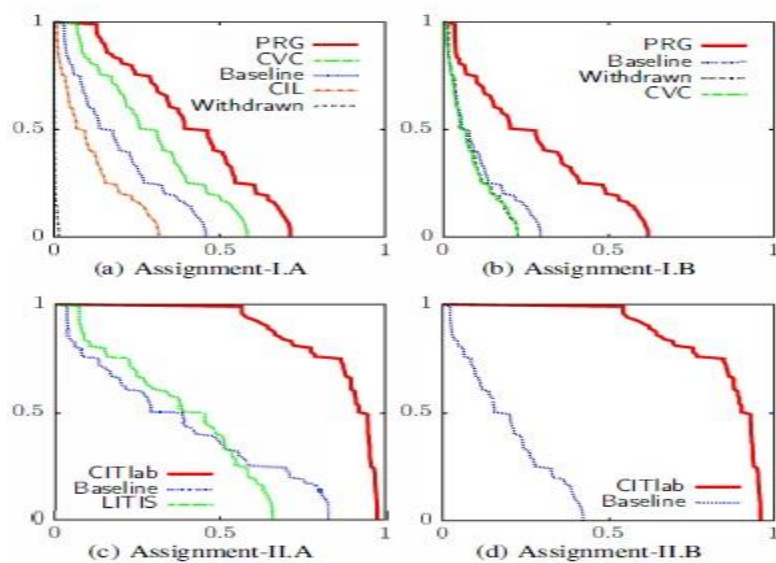
Αυτή η λέξη μπορεί να περιβάλλεται από οποιοδήποτε σημείο στίξης, κενό διάστημα, ή να βρίσκεται στην αρχή ή στο τέλος της γραμμής του κειμένου. Τα πλαίσια οριοθέτησης για τις λέξεις προσεγγίζονται από ευρετήρια τα οποία σχετίζονται με την περιοχή λέξεων και βελτιώθηκαν χρησιμοποιώντας μέθοδες επεξεργασίας εικόνας.

Τα αποτελέσματα των παραπάνω μεθόδων αποδεικνύουν αν μη τι άλλο πως εκείνες οι μέθοδοι που βασίζονται στην εκπαίδευση επιτυγχάνουν πολύ καλύτερη απόδοση από τις προσεγγίσεις χωρίς εκπαίδευση. Τα συστήματα αυτά ανταπεξήλθαν σε περίπλοκες καταστάσεις και παρείχαν πολύ πιο ακριβή οριοθετημένα κουτιά ακόμα και σε περιοχές που ήταν περισσότερο έντονα διαγραμμένες οι εικόνες εγγράφων. Επίσης τα συστήματα CITlab αναδείχτηκαν επίσης για την αποτελεσματικότητά τους όπως επίσης και εκείνοι που έκαναν χρήση πιο εξελιγμένων κρυφών

μοντέλων Markov σε σχέση με εκείνα που χρησιμοποίησαν πιο απλουστευμένα. [PUIGCERVER J., 2015]

Σύμφωνα με τα αποτελέσματα που αποτυπώνονται και στους πίνακες που ακολουθούν, είναι δεδομένο πως αν είναι διαθέσιμα τα δεδομένα εκπαίδευσης, οι μέθοδοι που βασίζονται στην εκπαίδευση για την κατασκευή των συστημάτων επιτυγχάνουν περισσότερο χρήσιμες αποδόσεις. Ερώτημα εξακολουθεί να αποτελεί μέχρι πόσα δεδομένα εκπαιδευτικού υλικού θα πρέπει να γίνουν χρήση για να επιτευχθεί μια αισθητή διαφορά στην απόδοση αυτών των μεθόδων στον τομέα του εντοπισμού λέξεων κλειδιών. [PUIGCERVER J., 2015]

Στην εικόνα βλέπουμε πώς απεικονίζονται οι μέσες καμπύλες (mRP) ανάκλησης ακρίβειας (Recall Precision) των αποτελεσμάτων των συμμετεχόντων. Η περιοχή κάτω από τη μέση καμπύλη ακρίβειας είναι η περιοχή μέσης ακρίβειας. Οι καμπύλες δίνουν περισσότερο ακριβείς λεπτομέρειες για κάθε λύση όπως και τη μέγιστη ακρίβεια ανάκλησης που επιτυγχάνεται από τον κάθε συμμετέχοντα.



Εικ. 36. Μέσες καμπύλες ανάκλησης ακρίβειας και τα βασικά συστήματα ταξινομημένα σε φθίνουσα γραμμή. Η καμπύλη με το μεγαλύτερο πάχος αντιστοιχεί στην ομάδα με τα υψηλότερα αποτελέσματα.

Πίνακας 3: Αποτελέσματα ανάλογα με την διαδρομή α και τη διαδρομή β. Οι στήλες 2 & 4 δείχνουν τη μέση τιμή ακρίβειας (mAP) κάθε ανάθεσης και στη διαδρομή β στις στήλες 3 και 5 την ακρίβεια στο 5, n (5). Η τελευταία στήλη αποτυπώνει την βαθμολογία του κάθε συμμετέχοντα στο συγκεκριμένο μέρος.

(a) Track I: Training-free

Assignment Team	I.A: Segm. based		I.B: Segm. free		Track-I Score
	mAP	$\pi(5)$	mAP	$\pi(5)$	
PRG	0.4244	0.4605	0.2761	0.3434	1.2
CVC	0.3000	0.3427	0.0821	0.1087	0.7
Baseline	0.1935	0.2241	0.1023	0.1504	—
Withdrawn	0.0024	0.0028	0.0848	0.1088	0.0
CIL	0.1124	0.1475			0.0

(b) Track II: Training-based

Assignment Team	II.A: QbS		II.B: QbE		Track-II Score
	mAP	$\pi(5)$	mAP	$\pi(5)$	
CITlab	0.8711	0.8737	0.8521	0.8552	1.2
Baseline	0.3834	0.4831	0.1958	0.2356	—
LITS	0.3822	0.4864	—	—	0.0

5.1.3 ΕΝΤΟΠΙΣΜΟΣ ΛΕΞΕΩΝ ΣΕ ΧΕΙΡΟΓΡΑΦΟ ΚΕΙΜΕΝΟ. ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 2014

Σε διαγωνισμό του 2014 (ICFHR 2014), στόχο αποτελούσε η καταγραφή της προόδου σε αλγόριθμους εντοπισμού λέξεων κλειδιών χρησιμοποιώντας ως καθιερωμένο μέτρο αξιολόγησης της απόδοσης που συναντώνται συχνά στην ανάκτηση των πλέον σχετικών βιβλιογραφικών πληροφοριών. Ο διαγωνισμός αποτελείται από ξεχωριστά διακριτά μέρη η μία με βάση την τμηματοποίηση και η άλλη ελεύθερη από τμηματοποίηση. Υπάρχει συμμετοχή διαφορετικών ερευνητικών ομάδων με μεθόδους με βάση την κατάτμηση κειμένου και άλλες χωρίς κατάτμηση. Καταγράφονται οι τότε εξελίξεις στον εντοπισμό λέξεων κλειδιών και εξερεύνηση απόδοσης καθιερωμένων μέτρων αξιολόγησης που απαντώνται συχνά στην ανάκτηση σχετικών βιβλιογραφικών πληροφοριών. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την συγκριτική αξιολόγηση περιέχουν τόσο ιστορικά όσο και μοντέρνα έγγραφα από πολλαπλούς συγγραφείς. [PRATIKAKIS I., 2014]

Τα θέματα που εξετάζονται έχουν ως εξής: μια εικόνα λέξης ερωτήματος μαζί με τη συλλογή εικόνων χειρόγραφων εγγράφων που δίνονται ως είσοδος στο υπό εξέταση σύστημα. Η υπολογιζόμενη έξοδος αποτελείται από μια λίστα κατάταξης οριοθετημένων κουτιών τα οποία

Η βάση δεδομένων από την οποία αντλούνται τα δεδομένα είναι η βάση Bentham αποτελούμενη από χειρόγραφα ντοκουμέντα υψηλής ποιότητας ανάλυσης, 3000 εικονοστοιχείων σε πλάτος και αντίστοιχα 4000 εικονοστοιχεία σε ύψος. Τα χειρόγραφα έγγραφα έχουν συγγραφεί από τον Bentham Jeremy όσο και από το προσωπικό της γραμματείας του στα τέλη του 18^{ου} αιώνα και στο πρώτο ήμισυ περίπου του 19^{ου}. Επίσης υπάρχει και μια σύγχρονη βάση αποτελούμενη από διάφορους συγγραφείς και είναι σε 4 γλώσσες λατινογενείς και στα ελληνικά.

[illegible]

Εικ. 37. Απεικόνιση εικόνας εγγράφων από α) βάση Bentham και β) Modern dataset

Ως μέτρο αξιολόγησης των μεθόδων των αλγόριθμων που χρησιμοποιήθηκαν για τον εντοπισμό λέξεων, ήταν το Precision at Top, δηλαδή στην ανάκτηση με ακρίβεια 5 λέξεων στην κορυφή (P@5), η μέση ακρίβεια (mAP) και το λεγόμενο κανονικοποιημένο μειωμένο αθροιστικό κέρδος τόσο για δυαδικές όσο και για μη δυαδικές συνάφειες κρίσης. Οι καμπυλες ακρίβειας ανάκλησης παρουσιάζουν την απόδοση των μεθόδων σε όλο το εύρος της ανάκλησης.

Όσο για την ακρίβεια, αυτή είναι το κλάσμα των ανακτημένων λέξεων οι οποίες είναι σχετικές με το ερώτημα, ενώ την ίδια ώρα αυτή πρέπει να είναι ακριβής και να προσδιορίζεται για τις λέξεις k που ανακτώνται και υπολογίζεται με βάση το P @ k. Ειδικότερα, χρησιμοποιείται το P@5 όπου η ακρίβεια αφορά τις 5 ανακτώμενες λέξεις. Με βάση τα παραπάνω η μέτρηση ορίζει με πόση επιτυχία οι αλγόριθμοι παράγουν τα καλύτερα αποτελέσματα στις 5 πρώτες θέσεις της λίστας. Επίσης η εννοούμενη μέση ακρίβεια (mAP) χρησιμοποιείται για την αξιολόγηση των μεθόδων αποτελεί τυπικό μέτρο της απόδοσης των συστημάτων για ανάκτηση πληροφοριών. Έτσι ορίζεται ο μέσος όρος της τιμής ακριβείας που λαμβάνεται κάθε φορά μετά από κάθε ανακτώμενη σχετική λέξη. [PRATIKAKIS I., 2014]

Ειδικότερα στις χρησιμοποιούμενες μεθόδους η ομάδα του «The Blavatnik School of Computer Science, Tel- Aviv University, Israel» το σύνολο δεδομένων των εικόνων προς υποβολή ερώτησης, υποβάλλονται σε προεπεξεργασία με μια απλή λειτουργία δυαδικοποίησης ακολοθούμενη από την ανάκτηση πολλαπλών υποψήφιων αλληλεπικαλυπτόμενων στόχων. Κάθε δυαδικός στόχος όπως και το δυαδικό ερώτημα, μεταβάλλει το μέγεθος προκειμένου να προσαρμόζεται σε ένα ορθογώνιο σταθερού μεγέθους και να παρουσιάζεται από συμβατικούς περιγραφείς εικόνας. Η μεταβαλλόμενη σε μέγεθος εικόνα 160X56 είχε χωριστεί σε ένα πλέγμα 20X7 κελιών καθένα εκ των οποίων αποτελείται από 8x8 εικονοστοιχεία. Για την βελτίωση των αποτελεσμάτων η εικόνα του ερωτήματος μεταφέρθηκε σε κάθε μια από τις 4 κατευθύνσεις και έγινε μέγιστη συγκέντρωση των αποτελεσμάτων. Η ανάκτηση των αποτελεσμάτων πραγματοποιείται σε κλάσμα του δευτερολέπτου.

Η ομάδα “Computer Vision Center, Universitat Autònoma de Barcelona, Spain” με βάση την μέθοδο Almazan (βλέπε περιγραφή στην εικόνα), όπου ο εντοπισμός και η αναγνώριση ζητημάτων αντιμετωπίστηκαν μέσω της μάθησης μιας κοινής αναπαράστασης τόσο για εικόνες κειμένου όσο και συμβολοσειρές κειμένου. Τα διάφορα χαρακτηριστικά χρησιμοποιούνται για να μάθουν μια σημασιολογική αναπαράσταση αποτελούμενη από εικόνες έξεων και στη συνέχεια πραγματοποιεί βαθμονόμηση των αποτελεσμάτων με CCA τα οποία τοποθετεί σε εικόνες και συμβολοσειρές σε κοινό χώρο. Συμβολοσειρές κειμένου ενσωματώνονται σε έναν δισδιάστατο δυαδικό χώρο μεταφερόμενο σε ένα πυραμιδικό ιστόγραμμα χαρακτήρων PHOC. Η ενσωμάτωση αυτή χρησιμοποιείται ως η προέλευση πηγής χαρακτηριστικών με προβολή εικόνων λέξεων σε άλλη χωρική διάσταση, όπου η κάθε διάσταση κωδικοποιεί το πόσο πιθανό είναι αυτή η λέξη να εμπεριέχει ένα συγκεκριμένο χαρακτήρα σε μια συγκεκριμένη περιοχή, πάντα σε σχέση με τον περιγραφέα του πυραμιδικού ιστογράμματος PHOC. Η μέθοδος αυτή παρέχει διπλό πλεονέκτημα. Αφενός κάνει άμεση σύγκριση μεταξύ εικόνων λέξεων και συμβολοσειρών που βγάζει νόημα και αφετέρου τα αποτελέσματα των εικόνων της ίδιας λέξης που συγκεντρώθηκαν καθοδηγούνται από το κοινόχρηστο PHOC.

Επόμενη χρησιμοποιούμενη μέθοδος από την ομάδα «Smith College, Department of Computer Science, Northampton MA, USA» και για τις δύο εργασίες, τόσο με τμηματοποίηση όσο και χωρίς τμηματοποίηση, οι καταχωρήσεις ακολουθούν έναν σχετικά ευέλικτο μηχανισμό προτύπων. Τα μοντέλα των ερωτημάτων ταιριάζουν στον στόχο των εικόνων της σελίδας για να εντοπιστούν τοποθεσίες όπου υπάρχει δυνατή καλή αντιστοίχιση (χαμηλής παραμόρφωσης). Περιοχές με καλύτερη αντιστοίχιση μετατράπηκαν σε ευέλικτα μοντέλα με μελάνι και ταιριάζουν στην εικόνα ερωτήματος για αντίστροφη επαλήθευση. Ο αλγόριθμος δίχως τμηματοποίηση ταιριάζει περισσότερο με το μοντέλο ερωτήματος σε πλήρεις σελίδες στόχους δίχως τυχόν προσαρμογές σε κλίμακα. Το κείμενο εισάγει την πιθανότητα σφάλματος, αν η εκτίμηση κλίμακας είναι λάθος. Το βήμα αντίστροφης αντιστοίχισης χρησιμοποιεί το μελάνι στον άμεσο χώρο τη αντιστοίχισης του μοντέλου ερωτήματος, δίχως να υπάρχει απόπειρα τμηματοποίησης των λέξεων, οπότε μπορεί να επιτρέπεται αντιστοίχιση σε μέρος της λέξης. Με βάση όμως την τμηματοποίηση ο αλγόριθμος προσπαθεί να στοχεύσει κλιμακωτά το μοντέλο ερωτήματος για κάθε λέξη στόχο με

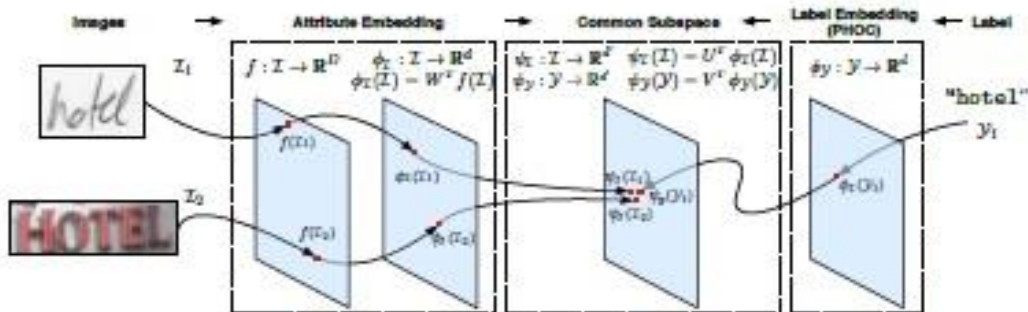
το να κάνει την αντιστοίχιση της διακριτικής απόστασης και των δύο οριζόντιων και κάθετων προβολών του μελανιού, με τελικό σκοπό τον περιορισμό και την αποφυγή ακραίων διατάσεων.

Η γαλλική ομάδα «Universit  de Lyon, CNRS, INSA-Lyon, LIRIS, France» ακολουθεί μεθόδους χωρίς μάθηση και ελεύθερες από τμηματοποίηση και δίνεται η δυνατότητα απευθείας εγναρμογής της σε έγχρωμες εικόνες και σε γκρι κλίμακα χωρίς δυαδικοποίηση. Τα αποτελέσματα της εν λόγω μεθόδου έχουν δοκιμαστεί με επιτυχία σε αρκετά είδη χειρογράφων, μεσαιωνικά, λατινικά, αραβικά, σανσκριτικά, ιερογλυφικά κ.λ.π. Μειονέκτημα της μεθόδου αποτελούσε το ότι ήταν σχετική αργή. Οι ζώνες ενδιαφέροντος εξάγονται από τη λέξη του ερωτήματος. Οι ζώνες αποτλεούνται από τοποθεσίες υψηλής καμπυλότητας με αντιστοίχιση στα άκρα και σε διασταυρώσεις. Το ερώτημα της λέξης μπορεί να παραμορφωθεί συνεκτικά για να ταιριάζει στη μεταβλητότητα των εγγράφων καθώς οι γεωμετρικοί σύνδεσμοι μεταξύ των ζωνών είναι ελαστικοί. Το ερώτημα γραφήματος των ζωνών ενδιαφέροντος συγκρίνεται και αντιστοιχεί σε τοποθεσίες που υπάρχουν χτυπήματα. Ιστογράμματα διαβάθμισης δεν χρησιμοποιούνται με αποτέλεσμα η μέθοδος αυτή χαρακτηριζόταν για την ακρίβειά της αλλά ο χρόνος απόκρισης μεγαλύτερος.

Τέλος, η ομάδα «Institute for Communications Technology (IfN) of Technische Universitat Braunschweig, Braunschweig, Germany» στην τυπική προσπάθεια για εντοπισμό των λέξεων, είναι η εύρεση πρότυπων λέξεων που εμφανίζονται σε μεγάλα ιστορικά έγγραφα τα οποία μπορεί να μην έχουν συγγραφεί από τον ίδιο συγγραφέα αλλά χαρακτηρίζονται από παρόμοιο στυλ γραφής. Οι παλαιογράφοι που θα πρεπε να έχουν γνώση στο στυλ γραφής θα μπορούσαν να χρησιμοποιήσουν αυτό το σύστημα για αναζήτηση γραμμένων λέξεων σε αυτό το παρόμοιο στυλ γραφής. Η μέθοδος εντοπισμού λειτουργεί σε εικόνες δίχως τμηματοποίηση εμπεριέχοντας τις πληροφορίες έγχρωμες ή σε χρώμα γκρι. Η περιγραφή της μεθοδολογίας με λίγα λόγια έχει ως εξής: αρχικά εφαρμόζεται μια σκιασμένη διόρθωση και στην περίπτωση του χρώματος οι εικόνες μετατρέπονται σε κλίμακα του γκρι χρησιμοποιώντας μια ψευδο φωτεινή προσέγγιση. Η γωνία κλίσης και το μέγεθος αξιοποιούνται ως χαρακτηριστικά τα οποία στη συνέχεια συγκρίνονται αυτόματα σε προσδιορισμένες ζώνες ενδιαφέροντος.

Αυτή η τεχνική παρότι θεωρείται ισχυρό εργαλείο για τυπικές παραλλαγές που μπορούν να εντοπιστούν σε χειρόγραφο κείμενο, εντούτοις δυσκολεύεται να προσεγγίσει και να ανιχνεύσει διαφορετικές παραλλαγές γραφής. Για να μειωθεί το υπολογιστικό κόστος το πρότυπο είχε προβλεφθεί έτσι ώστε να ταιριάζει με περιοχές εγγράφων που θεωρούνταν ότι περιείχε λέξεις. Η έξοδος του αλγόριθμου η θεωρείται η η λίστα καλύτερης κατάταξης περιοχών στο σύνολο των δεδομένων των εικόνων εγγράφων τα οποία προσομοιώνουν με το σχήμα του προτύπου της εικόνας, όπου το n , υποδηλώνει τον αριθμό των αποτελεσμάτων.

Συγκριτικά με βάση τους πίνακες αποτελεσμάτων αναφορικά με την εργασία που βασίζεται στην τμηματοποίηση η καλύτερη επίδοση είναι εκείνη της ομάδας του Αλμαζάν, όπως αποτυπώνεται στην εικόνα του Πανεπιστημίου της Βαρκελώνης «Computer Vision Center» (G2) και για την δεύτερη εργασία ελεύθερη από τμηματοποίηση κειμένου η καλύτερη επίδοση έχει επιτευχθεί από το «Blatavnik School» του Πανεπιστημίου του Τελ Αβίβ του Ισραήλ (G1). [PRATIKAKIS I., 2014]



Εικ. 38. Αποτύπωση μεθόδου Almazan.

Πίνακας 4: Αποτελέσματα μεθόδων με βάση την τμηματοποίηση

Method	BENTHAM DATASET				MODERN DATASET			
	P@5	MAP	NDCG (Binary)	NDCG	P@5	MAP	NDCG (Binary)	NDCG
G1	0.738 (1)	0.524 (1)	0.742 (2)	0.762 (2)	0.588 (2)	0.338 (2)	0.611 (2)	0.612 (2)
G2	0.724 (2)	0.513 (2)	0.744 (1)	0.764 (1)	0.706 (1)	0.523 (1)	0.757 (1)	0.757 (1)
G3	0.718 (3)	0.462 (3)	0.638 (3)	0.657 (3)	0.569 (3)	0.278 (3)	0.484 (3)	0.485 (3)

Πίνακας 5: Αποτελέσματα μεθόδων δίχως τμηματοποίηση και στα δύο σύνολα βάσεων δεδομένων.

BENTHAM DATASET																
	P@5				MAP				NDCG (Binary)				NDCG			
Method	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.617	0.611	0.599	0.609 (1)	0.428	0.419	0.402	0.416 (1)	0.653	0.640	0.621	0.638 (1)	0.671	0.657	0.640	0.56 (1)
G3	0.596	0.568	0.506	0.556 (2)	0.397	0.372	0.321	0.363 (2)	0.551	0.518	0.457	0.509 (2)	0.569	0.536	0.474	0.526 (2)
G4	0.351	0.341	0.313	0.335 (4)	0.219	0.209	0.187	0.205 (4)	0.386	0.363	0.319	0.356 (4)	0.400	0.376	0.331	0.369 (4)
G5	0.597	0.55	0.477	0.543 (3)	0.385	0.347	0.280	0.337 (3)	0.569	0.513	0.424	0.502 (3)	0.586	0.531	0.440	0.519 (3)
MODERN DATASET																
	P@5				MAP				NDCG (Binary)				NDCG			
Method	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.541	0.541	0.535	0.539 (1)	0.265	0.265	0.259	0.263 (1)	0.491	0.484	0.473	0.483 (1)	0.491	0.485	0.474	0.483 (1)
G3	0.429	0.422	0.399	0.417 (2)	0.170	0.165	0.152	0.163 (2)	0.310	0.301	0.277	0.296 (2)	0.310	0.301	0.277	0.296 (2)
G4	0.250	0.241	0.211	0.234 (4)	0.095	0.089	0.077	0.087 (4)	0.218	0.195	0.161	0.191 (4)	0.218	0.195	0.161	0.191 (4)
G5	0.264	0.247	0.223	0.245 (3)	0.100	0.092	0.081	0.091 (3)	0.229	0.201	0.168	0.199 (3)	0.229	0.202	0.168	0.200 (3)

Πίνακας 6: Τελική κατάταξη μεθόδων και στις 2 εργασίες.

TRACK I: SEGMENTATION-BASED		
Rank	Method	Score
1	G2	10
2	G1	14
3	G3	24
TRACK II: SEGMENTATION-FREE		
Rank	Method	Score
1	G1	8
2	G3	16
3	G5	24
4	G4	32

5.2 ΠΙΝΑΚΕΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΥΓΚΡΙΤΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΕΦΑΡΜΟΓΩΝ ΣΕ ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΚΕΙΜΕΝΟΥ (HANDWRITING TEXT RECOGNITION, HTR). ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ.

Στο 1^ο κεφάλαιο έχει γίνει μια αρχική πρώτη αναφορά σχετικά με την τεχνική αναγνώρισης χειρόγραφων κειμένων και σε λίγες γραμμές θα θυμήσουμε μερικά γενικά πράγματα. Μια από τις γενικές εφαρμογές χρήσης της τεχνικής αναγνώρισης κειμένου σχετίζεται με τις βιβλιοθήκες που παρέχουν πρόσβαση σε υλικό από το διαδίκτυο. Έτσι έγγραφα πολύ σημαντικής ιστορικής και άλλες αξίας ψηφιοποιούνται και παρέχεται δυνατότητα μαζικής πρόσβασης online. Λόγω του ότι δυσανάγνωστα κείμενα μπορούν να χρησιμοποιηθούν από κοινό με ειδικές γνώσεις, η

παροχή της δυνατότητας αναγνώρισης του κειμένου μέσω της τεχνικής αναγνώρισης κειμένου μέσω των σαρώσεων των εικόνων και τόσο η δυνατότητα ευρετηρίασης αλλά και ερωτημάτων προσφέρει πολύ καλά αποτελέσματα.

Μια από τις βασικές τεχνικές αναγνώρισης χειρογράφου είναι μέσω της οπτικής αναγνώρισης χαρακτήρων (OCR) το οποίο προορίζεται να αναγνωρίσει χαρακτήρες και κείμενα αντί σχημάτων ή προσώπων. Η διαδικασία που ακολουθείται είναι μέσω της σάρωσης εικόνων και την μετατροπή τους σε έγγραφο κειμένου.

Υπάρχει η διάκριση ανάμεσα στην διαδικτυακή (online) και εκτός δικτύου (offline) αναγνώριση χειρογράφου όπως φαίνεται και στην εικόνα.

Στην διαδικτυακή αναγνώριση χειρογράφου περιλαμβάνεται η αυτόματη μετατροπή κειμένου καθώς αυτό είναι αποτυπωμένο σε ειδικό ψηφιοποιητή ή PDA, όπου ένας αισθητήρας λαμβάνει τις κινήσεις της άκρης της πένας καθώς και την εναλλαγή του στυλό. Αυτό αναφέρεται ως είδος ψηφιακής μελάνης και θεωρείται ως ψηφιακή αναπαράσταση της γραφής. Τα PDA είναι διαθέσιμα είτε σε γραφίδα είτε σε πληκτρολόγιο και αρχικά ξεκίνησαν με βάση την γραφίδα ως είσοδο.

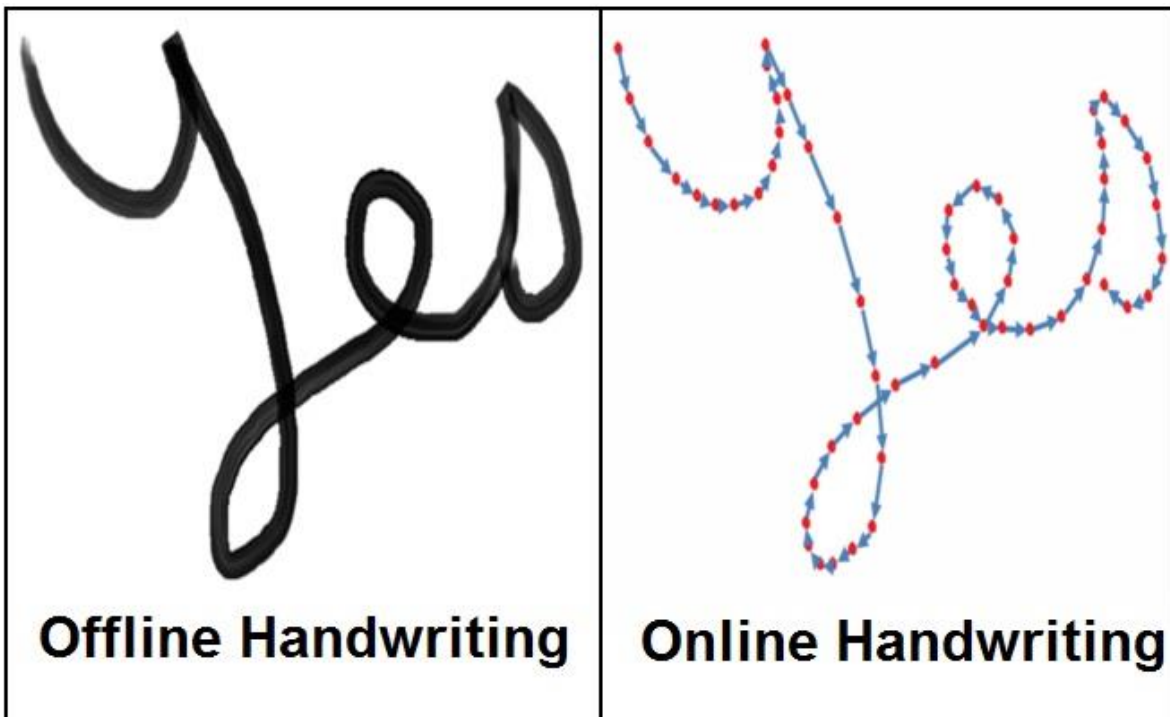
Η αναγνώριση ενός χειρόγραφου κειμένου μέσω συσκευής ως ένα πραγματικό αποτυπωμένο κείμενο αποτελεί μια πολύ σημαντική ικανότητα η οποία διαρκώς εξελίσσεται και βελτιώνεται.

Στην εκτός σύνδεσης αναγνώριση γραφής αυτή είναι συνώνυμη ως αναγνώριση οπτικού χαρακτήρα. Το OCR ιδίως σε χρήση βάσεων δεδομένων ψηφιακών βιβλιοθηκών παίζει σημαντικό ρόλο καθώς επιτρέπει την είσοδο πληροφοριών κειμένου εικόνας σε υπολογιστές με την ψηφιοποίηση και αποκατάσταση της εικόνας με μέθοδο αναγνώρισης. Συστήματα γραφής εκτός σύνδεσης αποτελούνται σε γενικές γραμμές από τα στάδια:

- Απόκτησης, ψηφιοποίηση.
- Τμηματοποίησης, χωρισμός εικόνας εγγράφου σε γραμμές λέξεις και χαρακτήρες.
- Αναγνώρισης, κάθε χαρακτήρας αναγνωρίζεται μέσω τεχνικής OCR.

- Μεταπεξεργασίας, με διορθώσεις σφαλμάτων χρησιμοποιώντας λεξικά ή μέσω ορθογραφικών ελέγχων.

Στην μέθοδο εκτός σύνδεσης τα στοιχεία στα οποία βασίζεται η αναγνώριση είναι μόνο σε αυτές τις παρεχόμενες πληροφορίες και σε τίποτα άλλο, όπως επίσης και ο θόρυβος που προέρχεται από το φθαρμένο υλικό ή και το ποτισμένο μελάνι δημιουργεί περεταίρω δυσκολίες πρόβλεψης.



Εικ. 39. Offline και Online μέθοδοι αναγνώρισης κειμένου.

Σε γενικές γραμμές τα συστήματα αναγνώρισης γραφής εκτός σύνδεσης, είναι λιγότερο ακριβή από τα συστήματα διαδικτυακής αναγνώρισης, για το λόγο ότι μόνο χωρικές πληροφορίες είναι διαθέσιμες για τα εκτός σύνδεσης ενώ τα διαδικτυακά έχουν πληροφορίες τόσο χωρικές όσο και χρονικές. [CHANDRA Anil, 2021]

5.2.1 ΠΙΝΑΚΕΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΥΓΚΡΙΤΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΕΦΑΡΜΟΓΩΝ ΣΕ ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΚΕΙΜΕΝΟΥ (HANDWRITING TEXT RECOGNITION, HTR). ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 2016.

Στο κεφάλαιο 5.2 προηγήθηκε μια σχετικά σύντομη αναφορά στο τί είναι αναγνώριση κειμένου σε ποιες διαδικασίες και τεχνικές επεξεργασίας απαιτεί για να υπάρχει μια συσχέτιση με την

συνέχεια καθώς στο παρόν κεφάλαιο θα γίνει ενασχόληση αναφορικά με την εξέλιξη των τεχνικών των μεθόδων και των αποτελεσμάτων τους σε συνέδρια του 2016 και του 2014.

Ειδικότερα εδώ γίνεται αναφορά σε διαγωνισμό του 2016 (ICFHR2016) στα πλαίσια διεθνούς συνεδρίου αναφορικά με την αναγνώριση κειμένου και ως βάση δεδομένων χρησιμοποιήθηκε το project READ. Η έρευνα αφορούσε το να εξεταστούν διαδικασίες τεχνικές και μέθοδοι στην αναγνώριση κειμένου εκτός σύνδεσης (off line) και να γίνει μεταγραφή ιστορικών εγγράφων. [SANCHEZ J. A., 2016]

Το έργο Recognition and Enrichment of Archival Documents (READ) εστιάζει στο να καταστήσει πιο προσβάσιμο και διαχειρίσιμο το αρχειακό υλικό με τη χρήση τεχνολογιών αιχμής. Στόχος είναι να τεθούν νέα πρότυπα στην αναγνώριση χειρόγραφου, στον εντοπισμό λέξεων, στην ανάλυση διάταξης, αυτόματη αναγνώριση συγγραφέα κλπ, με βάση έρευνα του έργου transcriptorium.

Επίσης στόχος του READ είναι η παροχή πλατφόρμας υπηρεσιών αυτόματης μεταγραφής όπως το transcribes.

Στο συγκεκριμένο πλαίσιο τα αρχεία για εξυπηρέτηση της έρευνας επιλέχθηκαν να είναι στη γερμανική γλώσσα και πιο συγκεκριμένα της συλλογής Ratsprotokolle με ημερομηνίες μεταξύ του 1470 ως το 1805.

Number of:	Train	Validation	Test	Total
Pages	350	50	50	450
Lines	8,367	1,043	1,140	10,550
Running words	35,169	3,994	4,297	43,460
Lexicon	6,985	1,526	1,656	8,120
Running OOV	-	669	633	-
OOV Lexicon	-	574	563	-
Character set size	92	80	83	92
Running Characters	208,595	26,654	25,179	260,428

Εικ. 40. Η βάση Ratsprotokolle πάνω στην οποία εξετάστηκαν οι μέθοδοι αναγνώρισης κειμένου στο Συνέδριο του 2016

Οι ανατιθέμενες εργασίες αφορούσαν εκείνες που θα ήταν περιορισμένο το εκπαιδευτικό υλικό που θα χρησιμοποιούνταν και ήταν υποχρεωτικό για λόγους ευκολότερης σύγκρισης των αποτελεσμάτων. Για τον ίδιο λόγο το βασικό σύστημα που είχε ετοιμαστεί ήταν τα κρυφά

μοντέλα Μάρκοβ (HMM). Το άλλο μέρος της εργασίας αφορούσε απεριόριστο εκπαιδευτικό υλικό καθώς έτσι επιτρεπόταν η χρήση δεδομένων της επιλογής των συμμετεχόντων.

Η μέτρηση αξιολόγησης ήταν ένας γραμμικός συνδυασμός του ποσοστού σφάλματος λέξης (WORD ERROR, WER) και του Ποσοστού Σφάλματος Χαρακτήρων (Character Error Rate, CER) (ποσοστού 50% για το κάθε ένα) μεταξύ της αναφοράς μεταγραφής και της παρεχόμενης μεταγραφής στο σύστημα από κάθε γραμμή. Σφάλματα χαρακτήρων συγκεντρωμένα σε λίγες λέξεις θεωρούνται καλύτερο αποτέλεσμα από συστήματα με σφάλματα χαρακτήρων σε όλες τις λέξεις. [SANCHEZ J. A., 2016]

Η ομάδα «Human Language Technology and Pattern Recognition Group, Germany (RWTH)» χρησιμοποιεί δίχως κλίμακα εικόνες αμέσως μετά την εφαρμογή τους και αγωγή ενίσχυσης. Τα χαρακτηριστικά που χρησιμοποιήθηκαν ήταν διασδιάστατο πλέγμα εικονοστοιχείων. Για την οπτική μοντελοποίηση χρησιμοποιήθηκε συνδυασμός ROVER από 16 πολλαπλής κατεύθυνσης Long-Short Term Memory δίκτυα (MLSTM) με περίπου 5 MLSTM διαστρωματώσεις συνελκτικών επιπέδων και 3 φορές συγκεντρώσεις max pooling ανά δίκτυο. Όλα τα δίκτυα χρησιμοποίησαν το είδος εξόδου νευρωνικού δικτύου εξαγωγής αποτελεσμάτων Connectionist Temporal Classification (CTC). Για αποκωδικοποίηση χρησιμοποιήθηκαν τα κρυφά μοντέλα Μάρκοβ (HMM) και για τη γλώσσα ένα μοντέλο 10-gram βασισμένο σε χαρακτήρες με υπολογισμό εξομάλυνσης kneser-ney από την εκπαιδευτική βάση. Η μέθοδος εξομάλυνσης Kneser-Ney χρησιμοποιείται κυρίως για τον υπολογισμό της πιθανότητας κατανομής των n-grams σε ένα έγγραφο. [SANCHEZ J. A., 2016]

Η ομάδα «Telecom ParisTech, France, and University of Balamand, Lebanon (ParisTech)» χρησιμοποιεί ως σύστημα αναγνώρισης ένα επαναλαμβανόμενο νευρωνικό δίκτυο αμφίδρομης μακράς και σύντομης μνήμης το οποίο συνίσταται στην σύζευξη 2 επαναλαμβανόμενων νευρωνικών δικτύων. Τα δύο κρυφά επίπεδα είναι κατασκευασμένα από 100 μπλοκ LSTM με ένα μπλοκ ανά κελί. Το στρώμα εξόδου, αποτελείται από 93 νευρώνες που αντιστοιχούν στους διαφορετικούς χαρακτήρες, σε αριθμούς και σημεία στίξης. Το σύστημα αναγνώρισης BLSTM εκπαιδεύτηκε με βάση τη μέθοδο της βαθμίδας που παρέχεται σε όλες τις παρεχόμενες εικόνες.

Στην περίπτωση που τα ποσοστά σφάλματος δεν βελτιωθούν για 20 επεισόδια (epochs) η εκπαίδευση σταματά και έτσι αποφεύγεται η υπερβολική προσαρμογή δεδομένων. Με βάση το σύστημα BLSTM υπολογίζεται για κάθε πλαίσιο οι αντίστοιχοι έξοδοι του δικτύου, όπου ο κάθε ένας από αυτούς σχετίζεται με μια τάξη χαρακτήρων. Ο αλγόριθμος μετάβασης παρέχει μια ακολουθία λέξεων που έχουν δοθεί στο λεξικό και στο bi-gram γλωσσικό μοντέλο που δημιουργήθηκε από τα δεδομένα μεταγραφής που έχουν παρατεθεί.

Στη συνέχεια ακολουθεί η περιγραφή της μεθόδου της ομάδας «Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, France (LITIS)» η οποία χρησιμοποίησε ως δεδομένα εκπαίδευσης και επικύρωσης αυτά που της παρατέθηκαν. Εφαρμόστηκαν σε αριθμό εικόνων για τεχνητή επέκτασή τους μικρές τροποποιήσεις και αλλαγές όπως περιστροφή και τεμαχισμός. Έγινε εξομάλυνση των εικόνων σε ύψος ανάλυσης 100 εικονοστοιχείων και 3 στρώματα με (100, 70, 120) BLSTM επαναλαμβανόμενου νευρωνικού δικτύου χρησιμοποιώντας RNNLIB που είναι ένα επαναλαμβανόμενο νευρωνικό δίκτυο βιβλιοθήκης για επίλυση προβλημάτων μάθησης ακολουθίας. Κανένα γλωσσικό μοντέλο δεν ακολουθήθηκε και ως λεξικό χρησιμοποιήθηκαν οι λέξεις που δόθηκαν με το εκπαιδευτικό υλικό. Η αποκωδικοποίηση βασίστηκε σε συνδυασμό πολλαπλών BLSTM και η έξοδος συμβολοσειράς επαληθεύθηκε λέξη προς λέξη με το λεξικό.

Από την άλλη η ομάδα «BYU Computer Science Department, USA (BYU)» προσέγγισε το ζήτημα με τη χρήση επαναλαμβανόμενου συνελκτικού νευρωνικού δικτύου, έναν δηλαδή συνδυασμό από CNN, RNN και ενός νευρωνικού δικτύου εξαγωγής αποτελεσμάτων (CTC) με απώλεια. Τα εκπαιδευτικά δεδομένα αυξήθηκαν με την ομοιόμορφη τοποθέτηση σημείων ελέγχου στην εικόνα και με την τυχαία μετατόπιση των σημείων ελέγχου. Για κάθε πρωτότυπη εικόνα γραμμής δημιουργήθηκαν 10 συνθετικές παρουσίες. Με την υποβολή τους σε προ-επεξεργασία χρησιμοποιήθηκαν οι τεχνικές του βασικού συστήματος που παρασχέθηκε, η διαδικοποίηση και η γκρίζα κλίμακα της εικόνας. Για κάθε μέθοδο οι τρεις εικόνες συνενώθηκαν ως ξεχωριστά κανάλια και παρουσιάστηκαν στο δίκτυο.

Τέλος η ομάδα «Artificial Intelligence and Image Analysis (**A2IA**)» προχώρησε χρησιμοποιώντας μεθόδους τόσο στη βάση του περιορισμένου εκπαιδευτικού υλικού, όσο και του εκπαιδευτικού υλικού δίχως περιορισμούς.

Στην διαδικασία βάση του περιορισμένου εκπαιδευτικού υλικού, εκπαίδευση και σετ επικύρωσης

χρησιμοποιήθηκαν για τα εκπαιδευτικά μοντέλα. Σε κάθε γραμμή χρησιμοποιήθηκαν δύο τμηματοποιήσεις. Στο αρχικό πολύγωνο, όπως δίνεται σε αρχείο PAGE xml, με μια επέκταση του αρχικού πολυγώνου προς τα όρια περιοχών γειτονικών πολυγώνων προκειμένου να χαθεί το πίσω πλαίσιο. Οι εικόνες των γραμμών είχαν μετατραπεί σε κλίμακα γκρι χρώματος. Ως προς την οπτική μοντελοποίηση χρησιμοποιήθηκε εκπαιδευτικό μοντέλο πολυδιάστατων επαναλαμβανόμενων νευρωνικών δικτύων (MDLSTM-RNN) εκπαιδευμένο με CTC, επεκτείνοντας την δυναμική ικανότητα εφαρμογής των επαναλαμβανόμενων νευρωνικών δικτύων (RNN) σε οπτικές απεικονίσεις επεξεργασίες κλπ με αποφυγή προβλημάτων κλιμάκωσης από άλλα πολυδιάστατα μοντέλα.

Οι πρωτότυπες μεταγραφές γραμμής περιλαμβάνουν τμήματα από ενωτικές λέξεις όπως και ολόκληρες προτάσεις με το σύμβολο του συλλαβισμού που τις διαχωρίζει με σημεία στίξης κλπ. Η αποκωδικοποίηση έγινε εφικτή με χρήση σταθμισμένων τελικών τιμών με αναζήτηση δέσμης χρησιμοποιώντας το εργαλείο Kaldi που ενδείκνυται για εκπαίδευση μοντέλων αναγνώρισης ομιλίας και αποκωδικοποίησης ήχου ομιλιών.

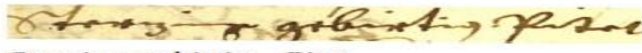
Η ομάδα «A2IA» στην εργασία δίχως περιορισμούς, χρησιμοποίησε επίσης για λόγους εκπαίδευσης και σετ επικύρωσης, γραμμές HWGL, όπου έχει εφαρμοστεί σε ένα εσωτερικό σύνολο δεδομένων από περίπου 500 χειρόγραφες επιστολές στα γερμανικά. Ένα πρώτο επίπεδο συνέλιξης έχει χρησιμοποιηθεί με δειγματοληψία 2X2 και εν συνεχεία επίπεδα LSTM και MDLST-RNN.

Τα καλύτερα αποτελέσματα των παραπάνω μεθόδων έχουν εξαχθεί και αποτυπώνονται στον παρακάτω πίνακα (Πίνακας 7). Στο μέρος της εργασίας δίχως περιορισμούς εκπαιδευτικού υλικού συμμετείχε η ομάδα «Α2ΙΑ» και τα αποτελέσματα σε σχέση με εκείνα της εργασίας με βάση το περιορισμένο εκπαιδευτικό υλικό είναι σχετικά ανώτερα. [SANCHEZ J. A., 2016]

Πίνακας 7: Βέλτιστα αποτελέσματα με βάση το ποσοστό σφάλματος λέξεων (Word Error Rate, WER) και ποσοστό σφάλματος χαρακτήρα (Character Error Rate, CER) των συμμετεχόντων και στις δύο διαδρομές με περιορισμό και χωρίς περιορισμό.

	Restricted track	Unrestricted track
RWTH	20.9±1.2 / 4.8±0.3	-
BYU	21.1±1.2 / 5.1±0.3	-
A2IA	22.1±1.2 / 5.4±0.3	21.0±1.2 / 5.1±0.3
LITIS	26.1±1.3 / 7.3±0.4	-
ParisTech	46.6±1.5 / 18.5±0.5	-

Η ομάδα RWTH όπως φαίνεται και από τα αποτελέσματα του πίνακα 7, αποτυπώνει στις μεταγραφόμενες γραμμές καλύτερα αποτελέσματα σύμφωνα με το φιλτράρισμα με το ποσοστό σφάλματος λέξεων (WER). Στην εικόνα... αναδεικνύονται τα ζητήματα μιας σωστής τμηματοποίησης καθώς αν κατά το στάδιο της αναγνώρισης αυτή δεν ισχύει, τότε τα αποτελέσματα στο ποσοστό σφάλματος λέξεων αυξάνονται, παρότι το ποσοστό σφάλματος λέξεων είναι χαμηλό.

	WER = 0/4 = 0% CER = 0/24 = 0%
	WER = 1/4 = 25% CER = 1/25 = 4%
	WER = 2/4 = 50% CER = 1/24 = 4,2%
	WER = 3/5 = 60% CER = 4/28 = 14,8%

Εικ. 41. Εικόνες γραμμών. Μεταγραφή αναφοράς και υπόθεσης ομάδας RWTH κάτω από κάθε εικόνα. Δεξιά αποτελέσματα με βάση το WER και CER.

5.2.2 ΠΙΝΑΚΕΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΥΓΚΡΙΤΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΕΦΑΡΜΟΓΩΝ ΣΕ ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΚΕΙΜΕΝΟΥ (HANDWRITTING TEXT RECOGNITION, HTR). ΜΕΘΟΔΟΙ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 2014.

Στο υποκεφάλαιο αυτό θα περιγραφούν μέθοδοι και τα αποτελέσματα που κρίθηκαν στα πλαίσια του 14^{ου} Διεθνούς Συνεδρίου σε σχέση με την αναγνώριση κειμένου. Περιλαμβάνονται δύο διαδρομές στις εργασίες των ομάδων δίχως περιορισμούς στη χρήση εκπαιδευτικού υλικού. Οι εικόνες των χειρογράφων αντλήθηκαν από αγγλική βάση δεδομένων που ήταν ενσωματωμένη στο έργο *transcriptorium*. Στόχος της έρευνητικής αυτής προσπάθειας ήταν να αντληθούν και να κριθούν αποτελεσματικότερες μέθοδοι και τεχνικές αλλά και από άποψη κόστους οφέλους στην μεταγραφή εικόνων κειμένων στα αγγλικά γερμανικά ισπανικά κλπ. Η αγγλική βάση που χρησιμοποιήθηκε ήταν η Bentham. [PRATIKAKIS I., 2014]

Οι διαδρομές των ανατιθέμενων εργασιών αφορούσαν την περιορισμένη διαδρομή όπου θα έπρεπε να γίνει χρήση μόνο των δεδομένων που παρασχέθηκαν για λόγους εκπαίδευσης και συντονισμού. Και επίσης στην μη περιορισμένη διαδρομή όπου οι συμμετέχοντες μπορούσαν να χρησιμοποιήσουν εκπαιδευτικό υλικό της αρεσκείας τους. Με τις δύο επιλογές διαδρομών μπορούσε να γίνει σύγκριση ως προς τα αποτελέσματα που θα επιτύγχαναν οι τεχνικές ανάλογα με την ποσότητα δεδομένων του υλικού εκπαίδευσης που χρησιμοποιήθηκε.

Τα καλύτερα αποτελέσματα τα οποία αντλήθηκαν από τους συμμετέχοντες στην διαδικασία, έχουν αποτυπωθεί στον πίνακα 8. Ήταν προφανές πως σημαντική βελτίωση αφορούσε υλικό που αφορούσε μεγαλύτερη ποσότητα εκπαιδευτικού υλικού όπως φαίνεται στη στήλη χωρίς περιορισμό. Η ομάδα A2IA (Artificial Intelligence and Image Analysis) στην εργασία δίχως περιορισμό έδειξε τα καλύτερα αποτελέσματα παρότι οι συμμετέχοντες στις εν λόγω διαδικασίες χρησιμοποίησαν παρόμοιες τεχνικές. Ο λόγος της διαφοροποίησης ίσως να αφορούσε την χρήση οπτικών μοντέλων πιο βελτιωμένων σε συνδυασμό με χρήση μεγαλύτερου όγκου από διάφορες βάσεις δεδομένων και από την άλλη το γλωσσικό μοντέλο περιελάμβανε χειρόγραφο από την αγγλική βάση Bentham όπου η αναγνώριση επιτύγχανε ίσως καλύτερα αποτελέσματα σε σχέση με τις υπόλοιπες γλώσσες. [PRATIKAKIS I., 2014]

ΚΕΦ. 6º: ΣΥΜΠΕΡΑΣΜΑΤΑ

Κάνοντας μια περιδιάβαση και μια αναδρομή στο θέμα της μεταγραφής κειμένων και των μεθόδων τους είναι πολύ εύκολο να προσέξει κανείς τις διαφοροποιήσεις και την αποτελεσματικότητα των μεθόδων και την εξέλιξη των τεχνικών και της ακρίβειας με την πάροδο των χρόνων. Η δυνατότητα της μεταγραφής κειμένων παλαιότερων χρόνων σε σύγχρονο κείμενο είτε έστω η κατανόησή τους, κρίθηκε ως πολύτιμο εργαλείο για την ερμηνεία γεγονότων του παρελθόντος και ως σημαντικότερη πηγή γνώσης. Μια πρώτη προσπάθεια ανάκτησης τέτοιων πληροφοριών έγινε με την επιστήμη της παλαιογραφίας γνώση όμως που την είχαν λίγοι.

Όμως στη συνέχεια στα μέσα περίπου του 20^{ου} αιώνα ήρθε η τεχνολογία με την ανάπτυξη υπολογιστικών συστημάτων να φέρει την δική της επανάσταση. Όσο η τεχνολογική εξέλιξη αυξανόταν τόσο επίσης αναπτύσσονταν και δημιουργούνταν νέες μέθοδοι και εμπλουτίζονταν άλλες.

Ενώ μέχρι την δεκαετία του 1980 οι προσπάθειες ακόμα θεωρούνταν σε εβρυϊκό στάδιο, ήδη στις δύο πρώτες δεκαετίες του 21^{ου} αιώνα παρατηρούμε μια πραγματική επανάσταση. Ειδικότερα καθώς εισάγονται νέες τεχνολογίες όπως η μηχανική μάθηση, η βαθιά μάθηση, τα μεγάλα δεδομένα, η τεχνητή νοημοσύνη, οι ολοένα και μεγαλύτεροι επεξεργαστές υπολογιστικών συστημάτων και οπτικής ανάλυσης εικόνων δημιουργούν προϋποθέσεις είτε να χρησιμοποιηθούν συνδυαστικά σε νέες μεθόδους και τεχνικές είτε να επαυξήσουν τις δυνατότητες ήδη υπάρχουσών μεθόδων. Ιδιαίτερα δε η εξέλιξη αρχιτεκτονικών που στηρίζονται σε συνελκτικά νευρωνικά δίκτυα (CNN) σε συνδυασμό με τις παραπάνω τεχνολογίες προσδίδουν εξαιρετικές δυνατότητες.

Αποτέλεσμα της χρήσης των παραπάνω τεχνολογιών, είναι να αποδίδονται με την εξέλιξη υπάρχοντων ή μελλοντικών τεχνολογιών, ολοένα και καλύτερα αποτελέσματα με μεγαλύτερη

ακρίβεια στην μεταγραφή κειμένων του παρελθόντος, τα οποία θα παρέχονται μέσω ψηφιακών Βιβλιοθηκών. Έτσι θα δίνεται η δυνατότητα μαζικής πρόσβασης σε πραγματικό χρόνο καθώς και δυνατότητα αναζήτησης ειδικής ευρετηριασμένης γνώσης σε ερευνητές ακαδημαϊκούς κλπ.

Βιβλιογραφία

[DIMOND T. L., 1957] IRE-ACM-AIEE '57 (Eastern): Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference: Computers with deadlines to meet December 1957, pp. 232–237, doi: 10.1145/1457720.1457765

[FISHER, 2009] Fischer *et al.*, "Automatic Transcription of Handwritten Medieval Documents," *2009 15th International Conference on Virtual Systems and Multimedia*, Vienna, 2009, pp. 137-142, doi: 10.1109/VSM.2009.26.

[MALAGON, 2012] Malagón, C., Rizky, R., Kim Y., Marzal F. and Izquierdo, L. "Automatic Abbreviation Detection in Medieval Medical Documents," *2012 International Conference on Frontiers in Handwriting Recognition*, Bari, 2012, pp. 337-342, doi: 10.1109/ICFHR.2012.184.

[SEAWARD, 2017] Seaward, L., & Kallio, M. "Transkribus: Handwritten Text Recognition technology for historical documents" *DH*.

[MUEHLBERGER, 2019] Muehlberger, G. et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study", *Journal of Documentation*, Vol. 75 No. 5, pp. 954-976, doi: <https://doi.org/10.1108/JD-07-2018-0114>

[PHILIPS, 2020] Philips, J. P. & Tabrizi N. "Historical Document Processing: A Survey of Techniques, Tools, and Trends", East Carolina University, United States of America, 2020

[CHANDNA, 2016] Chandna S., Rindone F., Dachsbacher C., and Stotzka R., "Quantitative exploration of large medieval manuscripts data for the codicological research," 2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV), Baltimore, MD, 2016, pp. 20-28, doi: 10.1109/LDAV.2016.7874306.

[SPRINGMANN, 2014] Springmann, U. et al. "OCR of historical printings of Latin texts: problems, prospects, progress." DATeCH '14 (2014).

[BAECHLER, 2010] Baechler M., Ingold R. "Medieval manuscript layout model" ACM Symposium on Document Engineering 2010: 275-278

[SIMISTIRA, 2016] Simistira F., Seuret M., Eichenberger N., Garz A., Liwicki M. and Ingold R, "DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 471-476, doi: 10.1109/ICFHR.2016.0093.

[KRIZHEVSKY, 2012] Krizhevsky A., Sutskever I., Hinton G. E. "ImageNet Classification with Deep Convolutional Neural Networks", NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, December 2012, pp. 1097–1105

[SUCH, 2018] Such F. P., Peri D., Brockler F., Paul H. and Ptucha R., "Fully Convolutional Networks for Handwriting Recognition," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 86-91, doi: 10.1109/ICFHR-2018.2018.00024.

[SUDANA, 2020] Sudana, Oka; Gunaya, I Wayan; Putra, I Ketut Gede Darma, "Handwriting identification using deep convolutional neural network method", Vol. 18, No. 4, August 2020, pp. 1934-1941, doi: 10.12928/TELKOMNIKA.v18i4.14864

[KESTEMONT, 2017] Kestemont, M. and D. Stutzmann. "Script Identification in Medieval Latin Manuscripts Using Convolutional Neural Networks." DH (2017).

[LOMBARDI, 2020] Lombardi Francesco, Marinai Simone, "Deep Learning for Historical Document Analysis and Recognition - A Survey". J. Imaging 6(10): 110 (2020)

[SUDHOLT, 2018] Sudholt, Sebastian and G. Fink. "Attribute CNNs for word spotting in handwritten documents." International Journal on Document Analysis and Recognition (IJDAR) 21 (2018): 199-218.

[CARNEIRO T., 2018] Carneiro T. , R. V. Medeiros Da Nobrega R. V. , Nepomuceno T., Bian G. B., De Albuquerque V. H. C. , and Filho P. P. R., "Performance Analysis of Google Colaboratory as a

Tool for Accelerating Deep Learning Applications,” IEEE Access, vol. 6, pp. 61677–61685, 2018. doi: 10.1109/ACCESS.2018.2874767.

[CHEN K., 2017] Chen, K.; Seuret, M.; Hennebert, J.; Ingold, R. Convolutional neural networks for page segmentation of historical document images. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 965–970.

[PASTOR-PELLICER J., 2016] Pastor-Pellicer, J.; Afzal, M.Z.; Liwicki, M.; Castro-Bleda, M.J. Complete system for text line extraction using convolutional neural networks and watershed transform. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 30–35

[ASI A., 2015] Asi A., Cohen R., Kedem K., and Elsana J., “Simplifying the reading of historical manuscripts,” in Proceedings of the 13th International Conference of Document Analysis and Recognition, pp. 826–830, 2015

[GARZ R., 2011] A. Garz, R. Sablatnig, and M. Diem, “Layout analysis for historical manuscripts using sift features,” in Proceedings of the 11th International Conference on Document Analysis and Recognition, pp. 508–512, 2011.

[SHI B., 2016] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11):2298–2304, 2016.

[KOLSCH A., 2018] Kölsch A., Mishra A., Varshneya S., Afzal M. Z., and Liwicki M., "Recognizing Challenging Handwritten Annotations with Fully Convolutional Networks," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 25-31, doi: 10.1109/ICFHR-2018.2018.00014.

[XU Y., 2017] Xu Y., He W., Yin F. and C. Liu, "Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 541-546, doi: 10.1109/ICDAR.2017.94.

[LONG J., 2015] Long J., Shelhamer E., and Darrell T., "Fully convolutional networks for semantic segmentation," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[ALMAZAN J., 2014] J. Almazán, A. Gordo, A. Fornés and E. Valveny, "Word Spotting and Recognition with Embedded Attributes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 12, pp. 2552-2566, 1 Dec. 2014, doi: 10.1109/TPAMI.2014.2339814.

[PRATIKAKIS I., 2016] Pratikakis I., Zagoris K., Gatos B., Puigcerver J., Toselli A. H., and Vidal, E., "ICFHR2016 Handwritten Keyword Spotting Competition (H-KWS 2016)," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 613-618, doi: 10.1109/ICFHR.2016.0117.

[PUIGCERVER J., 2015] Puigcerver J., Toselli A. H. and Vidal E., "ICDAR2015 Competition on Keyword Spotting for Handwritten Documents," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1176-1180, doi: 10.1109/ICDAR.2015.7333946.

[SANCHEZ J. A., 2016] Sanchez J. A., Romero V., Toselli A. H. and Vidal E. , "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 630-635, doi: 10.1109/ICFHR.2016.0120.

[PRATIKAKIS I., 2014] Pratikakis I., Zagoris K., Gatos B., Louloudis G., and Stamatopoulos N., "ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014)," 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014, pp. 814-819, doi: 10.1109/ICFHR.2014.142

Πηγές από Διαδικτυακές Ιστοσελίδες

«Handwritten text recognition». Geniza Lab. Princeton University., February 2021. Διαθέσιμο σε: <https://genizalab.princeton.edu/projects/handwritten-text-recognition>. (Ανακτήθηκε 24 Μαΐου, 2021)

“RESILIENCE Tool: eScriptorium”. 29 September 2020. RESILIENCE. Διαθέσιμο σε: <https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/>. (Ανακτήθηκε 24 Μαΐου, 2021)

[ROMAN Victor, 2020.]“Most Popular Convolutional Neural Networks Architectures: Learn about their structure and how to implement them! Διαθέσιμο σε: <https://towardsdatascience.com/convolutional-neural-networks-most-common-architectures-6a2b5d22479d> (Ανακτήθηκε 25 Μαΐου, 2021)

STUTZMANN, Dominique. ICFHR2016 competition on the classification of medieval handwritings in latin script. 2016. Διαθέσιμο σε: <https://orflamms.hypotheses.org/1388> (Ανακτήθηκε 20 Μαΐου 2021)

“VGG16 – Convolutional Network for Classification and Detection.” 20 November, 2018. Διαθέσιμο σε: <https://neurohive.io/en/popular-networks/vgg16/> (Ανακτήθηκε 18 Μαΐου, 2021)

“The evolution of ImageNet for Deep Learning in Computer vision.” 13 November 2020. <https://analyticsindiamag.com/imagenet-and-variants/> (Ανακτήθηκε 19 Ιουλίου, 2021)

CHANDRA Anil, MATCHA Naidu. “How to easily do Handwriting Recognition using Deep Learning”. May 2021. (Ανακτήθηκε Ιούνιος 2021)

Thesis

[MENGQIAO Z, 2020] Mengqiao, Zhao. Handwritten digit recognition based on segmentation-free method. Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden, October 2020.

[MILIONI, 2020] Milioni, N. “Automatic Transcription of Historical Documents. Transcribus a tool for Libraries, Archives and Scholars,” Uppsala Universitet, Department of ALM, 2020

