Επεξεργασία Εικόνων Ληξιαρχικού Αρχείου και Εξαγωγή Πληροφορίας

Η Διπλωματική Εργασία
παρουσιάστηκε ενώπιον
του Διδακτικού Προσωπικού του
Πανεπιστημίου Αιγαίου

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

της
ΜΑΤΘΑΙΟΥ ΕΙΡΗΝΗΣ
ΚΑΡΛΟΒΑΣΙ - ΦΕΒΡΟΥΑΡΙΟΣ 2012

Η ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΔΙΔΑΣΚΟΝΤΩΝ ΕΠΙΚΥΡΩΝΕΙ
ΤΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ ΜΑΤΘΑΙΟΥ ΕΙΡΗΝΗΣ:

ΚΑΒΑΛΛΙΕΡΑΤΟΥ ΕΡΓΙΝΑ , Επιβλέπων

Ημερομηνία 17 ΦΕΒΡΟΥΑΡΙΟΥ 2012
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΣΤΑΜΑΤΑΤΟΣ ΕΥΣΤΑΘΙΟΣ, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΜΑΡΑΓΚΟΥΔΑΚΗΣ ΕΜΜΑΝΟΥΗΛ, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
2012

ΠΕΡΙΛΗΨΗ

Την τελευταία δεκαετία η ψηφιοποίηση παλαιών και σπάνιων αρχείων έχει υιοθετηθεί από πολλές Ευρωπαϊκές χώρες αλλά και από την Ελλάδα προκειμένου το υλικό αυτό να διατηρηθεί , να διευθετηθεί και να προωθηθεί. Προκύπτει λοιπόν , η ανάγκη για αυτοματοποιημένα συστήματα αναζήτησης , ταξινόμησης και εύρεσης πληροφορίας. Ιστορικά αρχεία θα μπορούσαν να είναι προσιτά σε όλους αλλά και διάφορες υπηρεσίες θα μπορούσαν εύκολα να προσαρμόσουν τα παλιό τους αρχείο στα σύγχρονα πληροφοριακά συστήματα.

Έτσι στα πλαίσια της εργασίας αυτής , ασχοληθήκαμε με την εξαγωγή πληροφορίας από ληξιαρχικά αρχεία , μέσω τεχνικών επεξεργασίας εικόνας. Η πληροφορία που εξάγεται είναι εικόνες – λέξεις που αναφέρουν το όνομα του εμπλεκόμενου ατόμου , το όνομα του πατέρα και της μητέρας του και την ημερομηνία και τον τόπο που τελείται το γεγονός (γέννηση , θάνατος κ.α.). Το σύστημα αυτό μπορεί να θεωρηθεί στάδιο προ-επεξεργασίας για οπτική αναγνώριση χαρακτήρων αλλά και είσοδος για άλλα συστήματα συλλογής , ταξινόμησης ή διάθεσης.

ΜΑΤΘΑΙΟΥ ΕΙΡΗΝΗΣ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ABSTRACT

During the last decade the method of digitation of old and rare archives has been adapted not only many European countries but also from Greece, in order to preserve, organize and promote them. Therefore, the need of authorized search and classification systems as well as information retrieval systems, comes to the forefront. In this way, historical records could be accessible to everyone and at the same time several public services could easily adjust their old archives to the contemporary information systems.

Thus, during this dissertation we focused on extracting information through registration records via image processing. The information extracted is text images regarding the name of the persons involved, father's and mother's name and finally the date and the place of the event described (birth, death e.t.c.). This system can be regarded as pre-processing stage of Optical Character Recognition (OCR) and in addition, input to other selection, classification or disposal systems

ΕΥΧΑΡΙΣΤΙΕΣ - ΑΦΙΕΡΩΣΕΙΣ

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

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

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

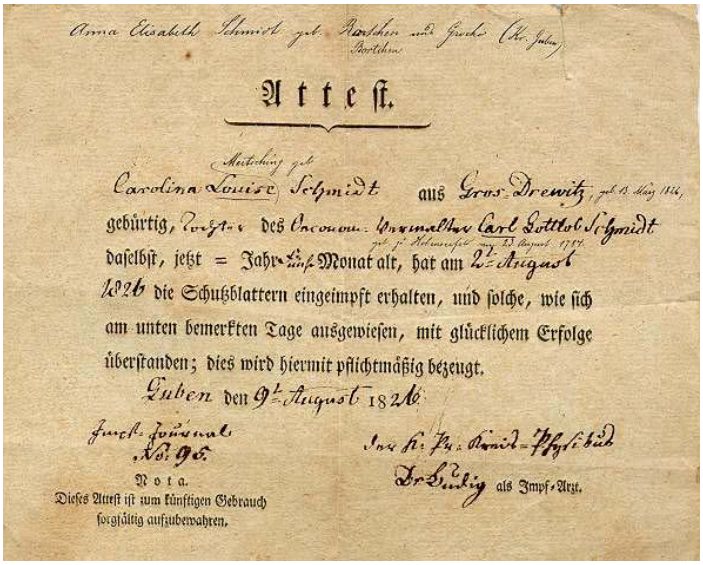
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Κεφάλαιο 1

Εισαγωγή

Τα ιστορικά έγγραφα, οι εφημερίδες, τα χειρόγραφα έγγραφα αποτελούν πολιτιστικά αγαθά. Για την διατήρηση της πολύτιμης πληροφορίας που έχουν αλλά και για την προβολή τους, τα τελευταία χρόνια έγινε μια πανευρωπαϊκή προσπάθεια ψηφιοποίηση τους.

Οι συλλογές των εικόνων – εγγράφων είναι τεράστιες και προκύπτει η ανάγκη περαιτέρω επεξεργασίας τους, προκειμένου να μπορούν να ταξινομηθούν και δεικτοδοτηθούν. Επίσης πολλές φορές τα έγγραφα αυτά είναι δυσανάγνωστα και η μελέτη τους από το ευρύ κοινό είναι αδύνατη. Ο λόγος που καθιστά ένα τέτοιου είδους έγγραφο δυσανάγνωστο συνήθως είναι οι φθορές λόγω παλαιότητας του αλλά και ο τρόπος γραφής των εκάστοτε εποχών. Βέβαια υπάρχουν και έγγραφα πιο σύγχρονα όπως τα ληξιαρχικά αρχεία τα οποία είναι μεικτά (μεικτά έγγραφα συναντάμε και παλαιότερα) δηλαδή αποτελούνται και από χειρόγραφο και από τυπωμένο κείμενο, τα οποία πρέπει να προσαρμοστούν στα νέα πληροφοριακά συστήματα φορέων και οργανισμών. Ένα αξιόπιστο αυτοματοποιημένο σύστημα ταξινόμησης ή αναζήτησης θα μείωνε κατά πολύ τα λειτουργικά έξοδα που απαιτούνται για την χειρωνακτική διαδικασία.



Attest.

Caroline Louise Schmidt aus Gros Drewitz, gebürtig, des Ackermann Gottlob Schmidt daselbst, jetzt 1 Jahr 5 Monat alt, hat am 2. August 1826 die Schutzblattern eingeimpft erhalten, und solche, wie sich am unten bemerkten Tage ausgewiesen, mit glücklichem Erfolge überstanden; dies wird hiermit pflichtmäßig bezeugt.

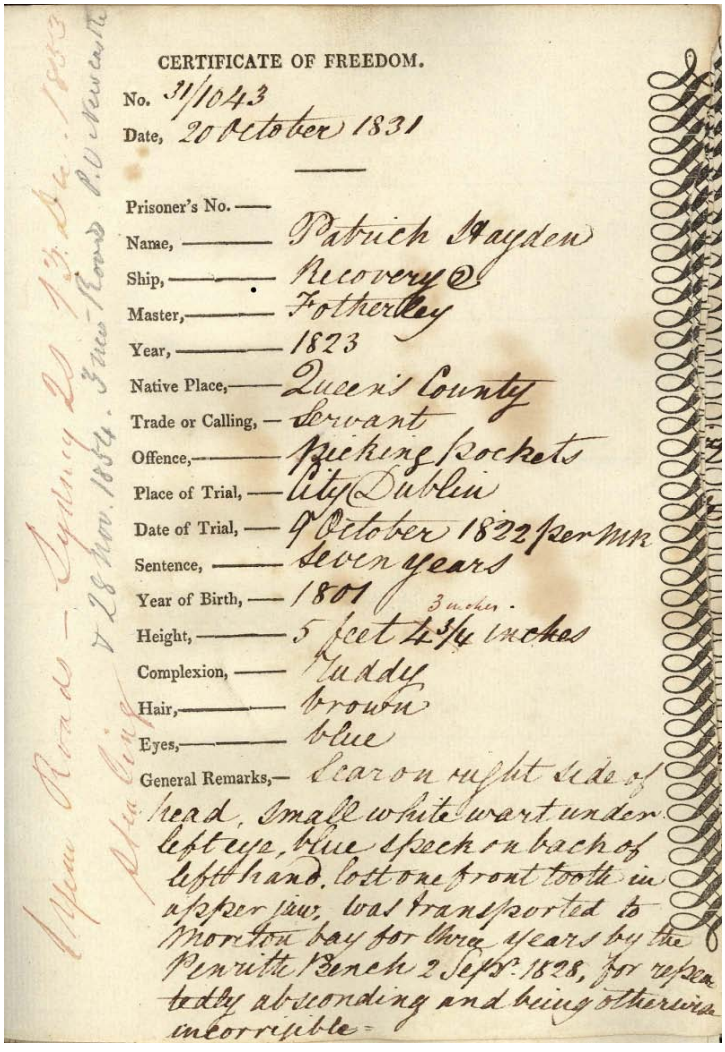
Luben den 9. August 1826.

Journal No. 95.

Nota.
Dieses Attest ist zum künftigen Gebrauch sorgfältig aufzubewahren.

Königl. Preuß. Kreis-Physikus
Dr. Lilig als Impf-Arzt.

Σχήμα 1: Μεικτό έγγραφο, Ιατρικής βεβαίωσης του 1826

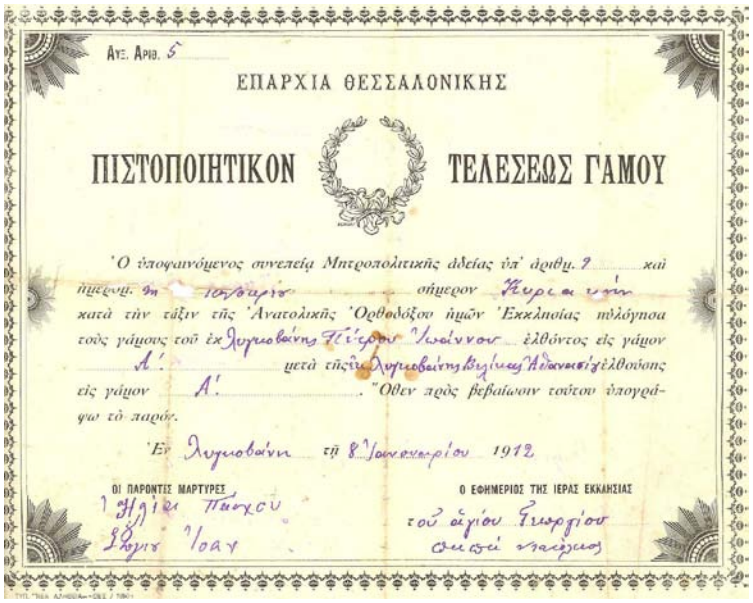


CERTIFICATE OF FREEDOM.

No. 31/1043
Date, 20 October 1831

Prisoner's No. —
Name, Patrick Hayden
Ship, Recovery
Master, Fotherley
Year, 1823
Native Place, Queens County
Trade or Calling, Servant
Offence, Picking Sockets
Place of Trial, City Dublin
Date of Trial, 9 October 1822 per mer
Sentence, seven years
Year of Birth, 1801
Height, 5 feet 4 3/4 inches
Complexion, Ruddy
Hair, brown
Eyes, blue
General Remarks, Scar on right side of head, small white wart under left eye, blue speck on back of left hand, lost one front tooth in upper jaw, was transported to Moreton bay for three years by the Penrith Bench 2 Sessions 1828, for repeatedly absconding and being otherwise incorrigible.

Σχήμα 2: Μεικτό έγγραφο, Πιστοποιητικού ελευθερίας 1831



Αντ. Αριθ. 5

ΕΠΑΡΧΙΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΙΣΤΟΠΟΙΗΤΙΚΟΝ ΤΕΛΕΣΕΩΣ ΓΑΜΟΥ

'Ο ὑποφαινόμενος συνεπείᾳ Μητροπολιτικῆς ἀδείας ὑπ' ἀριθμ. 9 καὶ ἡμερομ. 9η Ιανουαρίου σήμερον Κυριακήν κατὰ τὴν τάξιν τῆς 'Ανατολικῆς 'Ορθοδόξου ἡμῶν 'Εκκλησίας ἠλόγησα τοὺς γάμους τοῦ ἐκ Λυμπόβας Πέτρου Ἰωάννου ἐλθόντος εἰς γάμον Α'. μετὰ τῆς ἐκ Λυμπόβας Βιγίνας Ἀθανασίου ἐλθούσης εἰς γάμον Α'. "Οθεν πρὸς βεβαίωσιν τούτου ὑπογράφω τὸ παρόν.

'Εν Λυμπόβᾳ τῇ 8 Ἰανουαρίου 1912

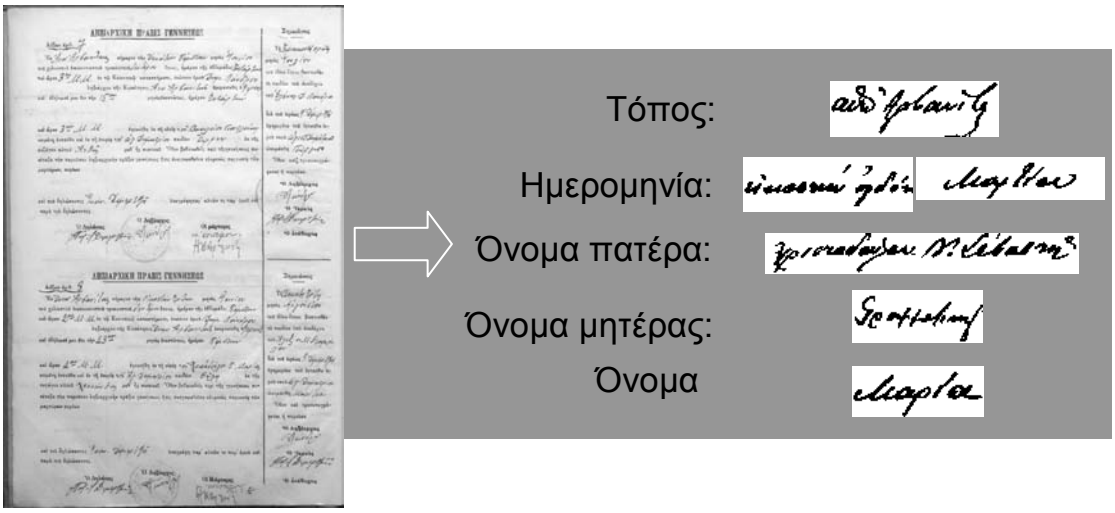
ΟΙ ΠΑΡΟΝΤΕΣ ΜΑΡΤΥΡΕΣ

Ὁ ΕΦΗΜΕΡΙΟΣ ΤΗΣ ΙΕΡΑΣ ΕΚΚΛΗΣΙΑΣ

Σχήμα 3: Μεικτό έγγραφο, Πιστοποιητικού γάμου του 1912

1.1 Στόχος

Στόχος αυτής της εργασίας είναι η δημιουργία ενός συστήματος εξαγωγής συγκεκριμένης πληροφορίας από ληξιαρχικά έγγραφα. Επίσης, το προτεινόμενο σύστημα περιέχει τμήμα προ-επεξεργασίας κατάλληλο για διάφορα είδη ιστορικών μεικτών εγγράφων. Με άλλα λόγια, η έξοδος του συστήματος θα είναι εικόνες κειμένου με συγκεκριμένη ερμηνεία που σκοπό θα έχουν να ορίσουν το περιεχόμενο του κειμένου π.χ. όνομα, ημερομηνία κ.α.



Σχήμα 4: Εξαγωγή ζητούμενης εικόνας-πληροφορίας

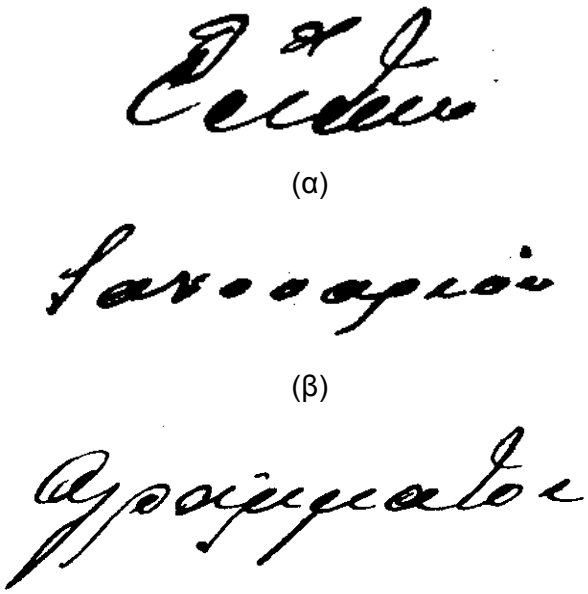
1.2 Προβλήματα ιστορικών εγγράφων

Γενικά τα ιστορικά έγγραφα έχουν πολλές φθορές λόγω παλαιότητας. Ένα πολύ συχνό φαινόμενο που έχει παρατηρηθεί είναι οι κηλίδες από υγρασία σε διάφορες θέσεις του εγγράφου. Κηλίδες μπορεί επίσης να βρει κανείς και από μελάνι. Με το πέρασμα των χρόνων το χαρτί απορροφά το μελάνι με αποτέλεσμα σε ένα έγγραφο να διακρίνεται και η πίσω γραμμένη ή τυπωμένη του πλευρά. Επιπλέον πολλές φορές συναντά κανείς σχίσματα είτε λόγω τσακίσματος της σελίδας, είτε λόγω φθοράς στις άκρες της. Τέλος η φθορά του εγγράφου μπορεί να προέλθει από ένα είδος εντόμου.

Εκτός από τις υλικές φθορές που έχει κανείς να αντιμετωπίσει σε ένα ιστορικό έγγραφο, πρέπει να λάβει υπόψη του και τις ιδιαιτερότητες της γραφής του. Στα τυπωμένα παλαιότερα έγγραφα τα γράμματα δεν είναι σωστά στοιχισμένα ούτε κατά ύψος αλλά ούτε κατά πλάτος. Επίσης κάποια κείμενα είναι τυπωμένα με κλίση είτε προς όλη την γραμμή, είτε προς τα γράμματα. Ακόμα και η ένταση που είναι τυπωμένα τα γράμματα σε κάθε έγγραφο ποικίλει. Τέλος τα διαστήματα ανάμεσα στα γράμματα και στις λέξεις δεν είναι σταθερά.

Πολλά προβλήματα καλείται να αντιμετωπίσει κανείς και όσον αφορά το χειρόγραφο κείμενο. Ένα από αυτά έγκειται στο γεγονός της γραφής κειμένου ανεξαρτήτου συγγραφέα. Αυτό σημαίνει ότι το κείμενο μπορεί να είναι με καλλιγραφική γραφή ή με

μεμονωμένους χαρακτήρες ή με γραφή χωρίς περιορισμούς. Στο χειρόγραφο επίσης, διαφέρει το μέγεθος των γραμμάτων, στη γραφή αν είναι αν είναι κεφαλαία ή πεζά, στην ένταση γραφής, στην κλίση των γραμμών αλλά και των μεμονωμένων χαρακτήρων. Επιπλέον ένας άνθρωπος μπορεί να γράψει εκτός των ορίων που του έχουν τεθεί από το τυπωμένο (στην περίπτωση των μεικτών εγγράφων) ή μπορεί να μουτζουρώσει ακόμα και να γράψει επιπλέον σημειώσεις σε ακαθόριστα σημεία του εγγράφου.



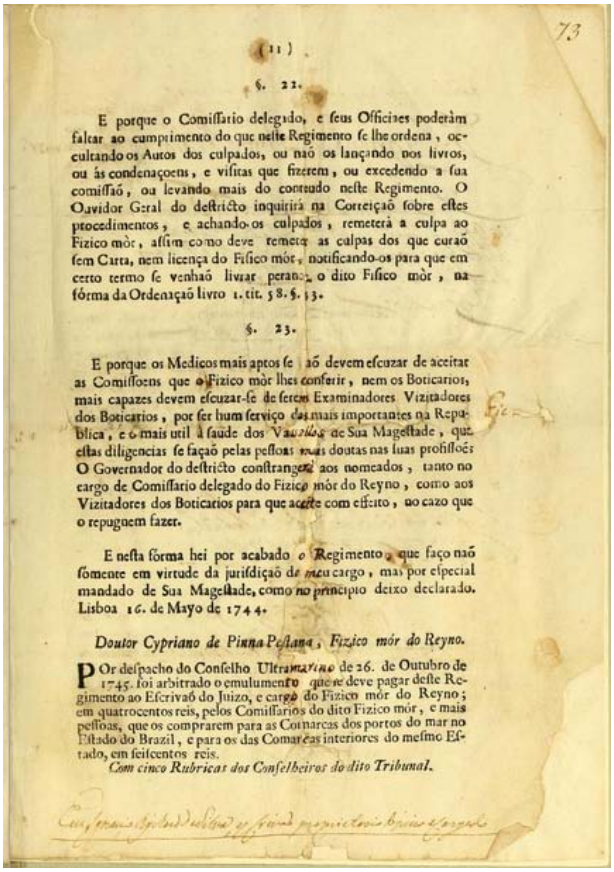
(α)

(β)

(γ)

Σχήμα 5: Είδη γραφής (α) καλλιγραφική, (β) μεμονωμένοι χαρακτήρες,(γ) χωρίς περιορισμούς

Πέρα από τα προβλήματα που δημιουργούν οι φθορές ή οι ιδιαιτερότητες των εγγράφων ως προς τον τρόπο γραφής τους γίνονται πολλά λάθη κατά την ψηφιοποίηση τους τα οποία πρέπει να αντιμετωπιστούν. Η ανομοιομορφία στο φωτισμό είναι συχνό φαινόμενο αλλά και η κλίση του εγγράφου. Τακτικά συμβαίνει να εμφανίζεται στις άκρες του εγγράφου κενό το οποίο προκύπτει από το κακό σάρωμα της εικόνας. Ωστόσο το πιο σοβαρό θέμα που καλείται να φέρει εις πέρας κανείς είναι η επιλογή κακής ανάλυσης και ποιότητας εικόνας κατά την ψηφιοποίηση.



(11)

§. 22.

E porque o Comiſſario delegado, e ſeus Officiaes poderão faltar ao cumprimento do que neſte Regimento ſe lhe ordena, occultando os Autos dos culpados, ou naõ os lançando nos livros, ou ás condenaçoens, e viſitas que fizerem, ou excedendo a ſua comiſſaõ, ou levando mais do contudo neſte Regimento. O Ouvidor Geral do deſtricto inquirirá na Correiçaõ ſobre eſtes procedimentos, e achando os culpados, remeterá a culpa ao Fizico mòr, aſſim como deve remeter as culpas dos que curaõ ſem Carta, nem licença do Fiſico mór, notificando-os para que em certo termo ſe venhaõ livrar perante o dito Fiſico mòr, na fórma da Ordenaçaõ livro 1. tit. 58. §. 13.

§. 23.

E porque os Medicos mais aptos ſe naõ devem eſcuzar de aceitar as Comiſſoens que o Fizico mòr lhes conferir, nem os Boticarios, mais capazes devem eſcuzar-ſe de ſerem Examinadores Vizitadores dos Boticarios, por ſer hum ſerviço dos mais importantes na Republica, e o mais util á ſaude dos Vaſſallos de Sua Mageſtade, que eſtas diligencias ſe façaõ pelas peſſoas mais doutas nas ſuas profiſſoẽs O Governador do deſtricto conſtrangerá aos nomeados, tanto no cargo de Comiſſario delegado do Fizico mór do Reyno, como aos Vizitadores dos Boticarios para que aceite com effeito, no cazo que o repugnem fazer.

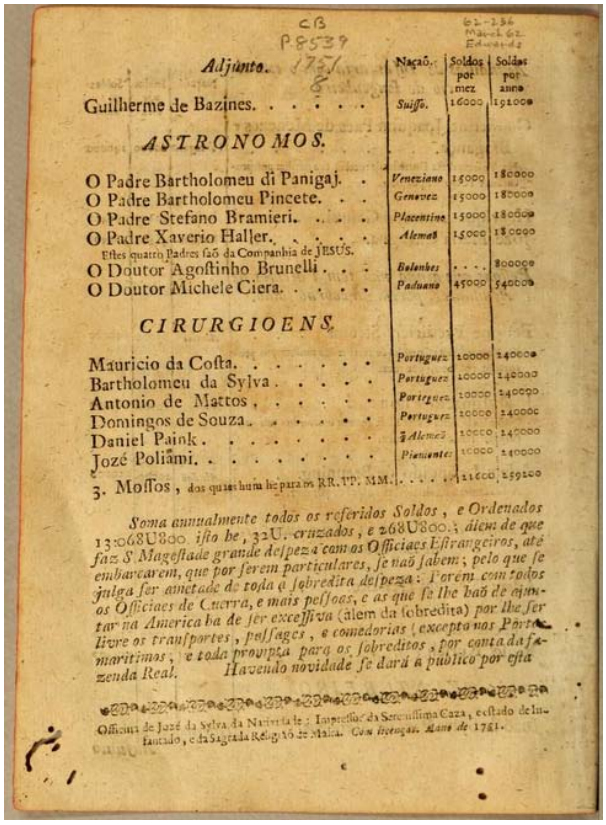
E neſta fórma hei por acabado o Regimento, que faço naõ ſomente em virtude da juriſdiçaõ de meu cargo, mas por eſpecial mandado de Sua Mageſtade, como no principio deixo declarado. Lisboa 16. de Mayo de 1744.

Doutor Cypriano de Pinna Peſſana, Fizico mór do Reyno.

POr deſpacho do Conſelho Ultramarino de 26. de Outubro de 1745. foi arbitrado o emolumento que ſe deve pagar deſte Regimento ao Eſcrivaõ do Juizo, e cargo do Fizico mór do Reyno; em quatrocentos reis, pelos Comiſſarios do dito Fizico mór, e mais peſſoas, que os comprarem para as Comarcas dos portos do mar no Eſtado do Brazil, e para os das Comarcas interiores do meſmo Eſtado, em ſeiſcentos reis.

Com cinco Rubricas dos Conſelheiros do dito Tribunal.

Σχήμα 6: Παράδειγμα ιστορικού εγγράφου με ύπαρξη θορύβου λόγω τσακίσματος

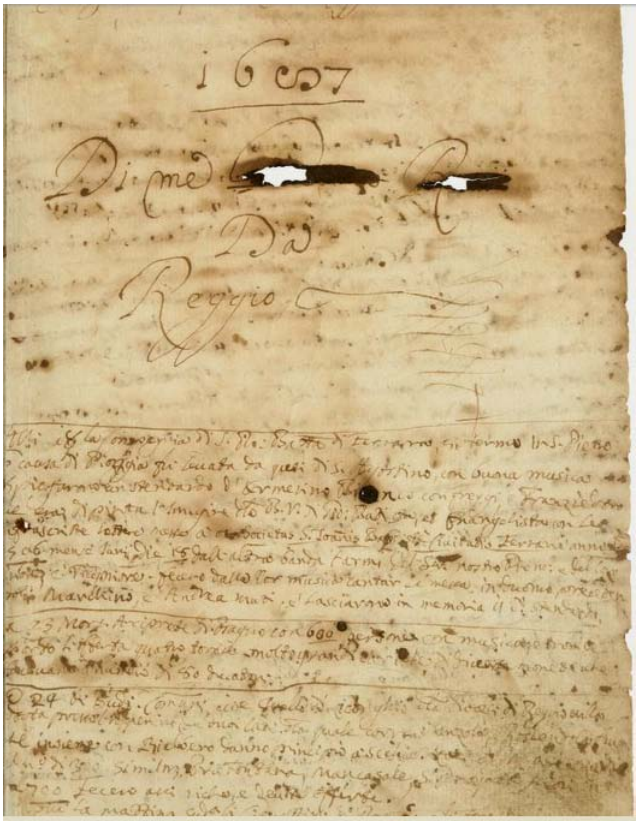


Adjunto.	Nação.	Soldos por mez	Soldos por anno
Guilherme de Bazines.	Suisso.	16000	192000
ASTRONOMOS.			
O Padre Bartholomeu di Panigaj.	Veneziano	15000	180000
O Padre Bartholomeu Pincete.	Genovez	15000	180000
O Padre Stefano Bramieri.	Placentino	15000	180000
O Padre Xaverio Haller.	Alemaõ	15000	180000
Estes quatro Padres saõ da Companhia de JESUS.			
O Doutor Agostinho Brunelli.	Bolonhez	....	800000
O Doutor Michele Ciera.	Paduano	45000	540000
CIRURGIOENS.			
Mauricio da Costa.	Portuguez	20000	240000
Bartholomeu da Sylva.	Portuguez	20000	240000
Antonio de Mattos.	Portuguez	20000	240000
Domingos de Souza.	Portuguez	20000	240000
Daniel Paink.	Alemez	20000	240000
Jozé Poliami.	Piemontez	20000	240000
3. Mossos, dos quaes huma he para os RR. PP. MM.		21600	259200

Soma annualmente todos os referidos Soldos, e Ordenados 13:068U800. isto he, 32U. cruzados, e 268U800.; alem de que faz S. Magestade grande despeza com os Officiaes Estrangeiros, até embarcarem, que por serem particulares, se naõ sabem; pelo que se julga ser ametade de toda a sobredita despeza: Porem com todos os Officiaes de Guerra, e mais pessoas, e as que se lhe baõ de ajuntar na America ha de ser excessiva (alem da sobredita) por lhe ser livre os transportes, passages, e comedorias (excepta nos Portos maritimos), e toda provizaõ para os sobreditos, por conta da fazenda Real. Havendo novidade se dará a publico por esta.

Officina de Jozé da Sylva da Natividade: Impressor da Serenissima Caza, e Estado de Infantado, e da Sagrada Religiaõ de Malta. Com licenças. Anno de 1751.

Σχήμα 7: Παράδειγμα ιστορικού εγγράφου με ύπαρξη θορύβου λόγω εντόμου αλλά και ύπαρξη γωνίας εκτροπής κειμένου



Σχήμα 8: Παράδειγμα ιστορικού εγγράφου με ύπαρξη θορύβου λόγω σκισίματος, λεκέδων και απορρόφησης μελανιού από το πίσω μέρος της σελίδας

1.3 Περιγραφή αρχείων

Στην παρούσα εργασία ασχοληθήκαμε με δυο ληξιαρχικά έγγραφα. Το πρώτο είναι το αρχείο των Άνω Αρβανιτών της Σάμου της δεκαετίας του ΄70. Σε αυτό το αρχείο έγινε και εξαγωγή πληροφορίας. Το δεύτερο αρχείο είναι του Αργοστολίου της Κεφαλονιάς και αναφέρεται στο χρονικό διάστημα 1900-1910. Σε αυτό το αρχείο δοκιμάστηκε το σύστημα όσον αφορά το στάδιο προ-επεξεργασίας. Και τα δυο παραπάνω αρχεία είναι μεικτά (χειρόγραφο και τυπωμένο κείμενο μαζί) .

Το αρχείο των Άνω Αρβανιτών έχει αρκετές διαφορές σε σχέση με αυτό του Αργοστολίου. Κατ' αρχάς η ανάλυση του είναι 72 dpi, το μέγεθος της εικόνας του εγγράφου είναι 810x1200 και το είδος του αρχείου είναι jpeg. Ο όρος dpi (dots per inch) αναφέρεται στην ανάλυση μιας μονάδας εξόδου και δείχνει πόσα σημεία (dots) θα τοποθετηθούν σε μια συγκεκριμένη διάσταση (inch). Το μέγεθος των dpi είναι ανάλογο με την ποσότητα των δεδομένων της εικόνας, δηλαδή για την περαιτέρω επεξεργασία της εικόνας, θα πρέπει το μέγεθος αυτό να κυμαίνεται στις τιμές 300 με 600 dpi. Για το αρχείο των Άνω Αρβανιτών το μέγεθος αυτό (72 dpi) θεωρείται πολύ μικρό. Επίσης το jpeg είναι ένα πρότυπο απωλεστικής συμπίεσης εικόνων, το οποίο μπορεί να έχει μικρό μέγεθος ως αρχείο αλλά έχει αρκετά μειονεκτήματα. Δηλαδή εμφανίζονται ατέλειες στην εικόνα, χρωματική παραμόρφωση των άκρων κ.α. .

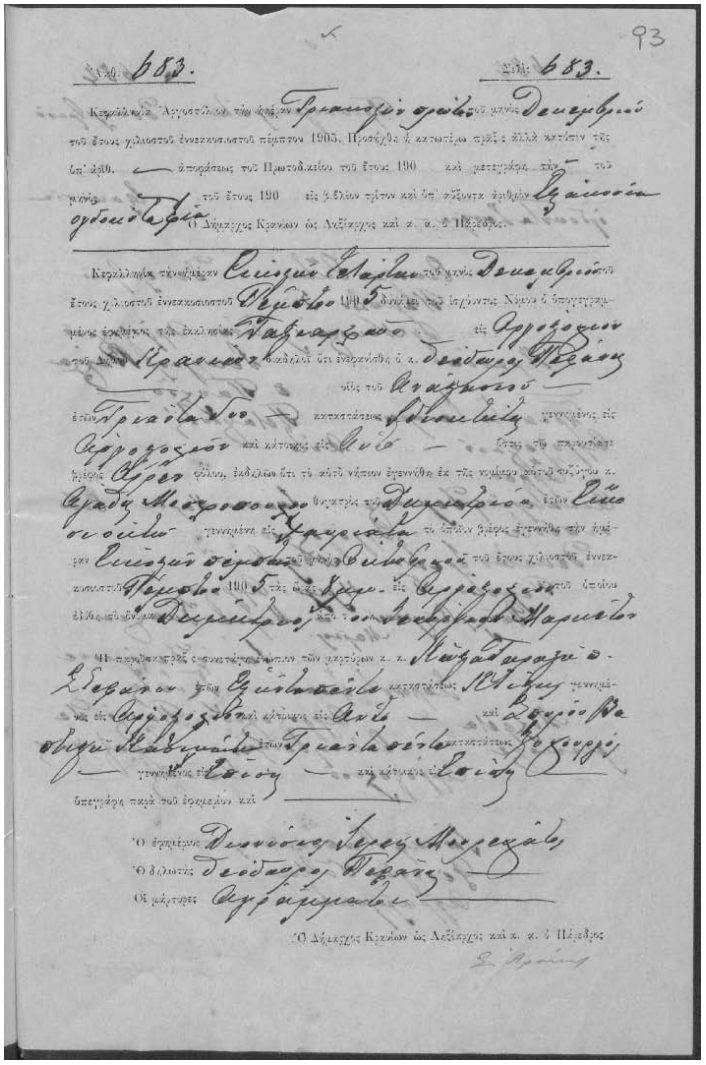
(α)

(β)

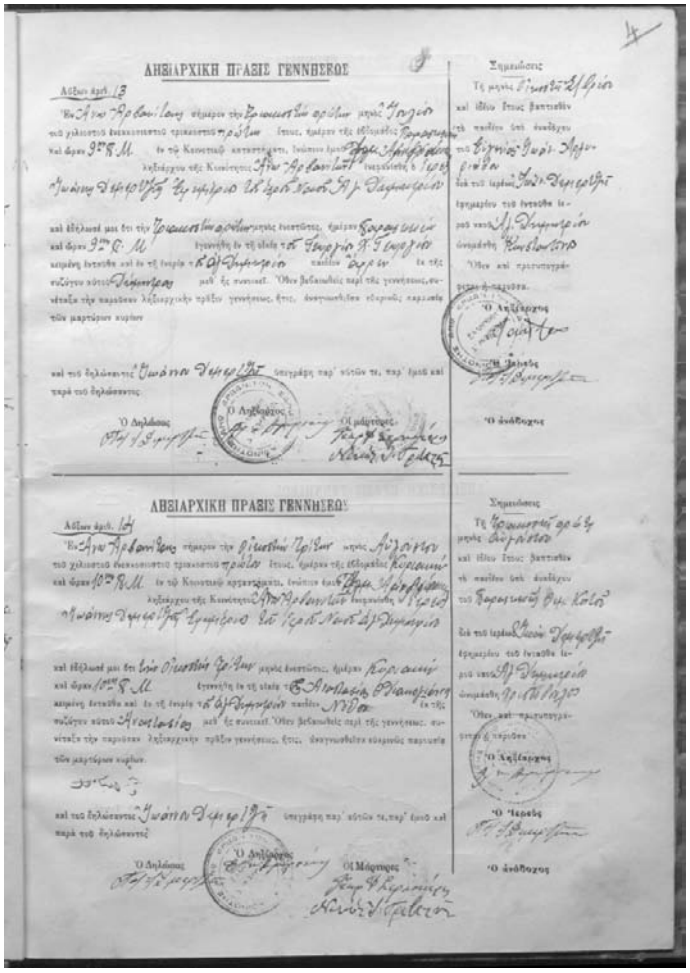
Σχήμα 9: Παράδειγμα ανάλυσης εικόνας: (α) 60 dpi, (β) 150 dpi

Επομένως το jpeg αρχείο πρέπει να αποφεύγεται σε περιπτώσεις που η ακρίβεια και η ποιότητα της εικόνας παίζει καθοριστικό ρόλο. Παρόλα αυτά το αρχείο της Σάμου έχει αυτή τη μορφή. Ένα άλλο μειονέκτημα του αρχείου είναι το μέγεθος της εικόνας εγγράφου. Όπως και η ανάλυση, το μέγεθος της εικόνας προσδιορίζει την ποσότητα των δεδομένων της. Επίσης στο αρχείο Άνω Αρβανιτών σε κάθε εικόνα έχουμε δυο εγγραφές οι οποίες αφορούν είτε πράξη γεννήσεως, είτε θανάτου, είτε γάμου.

Σε αντίθεση με το προηγούμενο αρχείο, αυτό της Κεφαλονιάς έχει υψηλότερη ανάλυση, 300dpi, μεγαλύτερο μέγεθος 2430x3680 περίπου και ο τύπος του αρχείου είναι tiff. Το πρότυπο tiff έχει υψηλότερο μέγεθος από το jpeg, αλλά χωρίς απώλεια ποιότητας. Ακόμη κάθε εικόνα εγγράφου έχει μια εγγραφή. Συμπερασματικά το αρχείο αυτό είναι καταλληλότερο για περαιτέρω επεξεργασία. Παρόλα αυτά η εργασία αντιμετωπίζει και τις δυο περιπτώσεις στο στάδιο της προ-επεξεργασίας.



Σχήμα 10: Έγγραφο-εικόνας αρχείου Κεφαλλονιάς



ΔΗΞΙΑΡΧΙΚΗ ΠΡΑΞΙΣ ΓΕΝΝΗΣΕΩΣ

ΔΗΞΙΑΡΧΙΚΗ ΠΡΑΞΙΣ ΓΕΝΝΗΣΕΩΣ

Σχήμα 11: Έγγραφο-εικόνας αρχείου Σάμου

1.4 Δομή εργασίας

Στα επόμενα κεφάλαια θα παρουσιαστούν και θα αναλυθούν λεπτομερώς τα διάφορα στάδια του προτεινόμενου συστήματος και θα δοθούν πειραματικά αποτελέσματα και συμπεράσματα.

Ποιο συγκεκριμένα στο δεύτερο κεφάλαιο θα παρουσιαστεί συνοπτικά το σύστημα και θα γίνει μια σύντομη περιγραφή του κάθε σταδίου του. Στο τρίτο κεφάλαιο θα γίνει αναλυτική περιγραφή της κάθε ενότητας του συστήματος. Στο τέταρτο κεφάλαιο παρατίθενται τα αποτελέσματα της εργασίας και στο τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα.

1.5 Τεχνικά Στοιχεία

Όλα τα στάδια του συστήματος υλοποιήθηκαν σε Matlab. Για την υλοποίηση χρησιμοποιήθηκε η 7.12 (R2011a) Matlab έκδοση.

Ο υπολογιστής που χρησιμοποιήθηκε για τα διάφορα πειράματα ήταν Intel Core i7 στα 2.2GHz και με μνήμη 4 GB, που κρίθηκε επαρκής για το πλήθος των εικόνων που χρησιμοποιήθηκαν.

Κεφάλαιο 2

Παρουσίαση συστήματος

Στο κεφάλαιο αυτό που ακολουθεί θα γίνει μια σύντομη περιγραφή του συστήματος εξαγωγής συγκεκριμένης πληροφορίας από ληξιαρχικά αρχεία αλλά και των σταδίων του, που είναι απαραίτητη για την κατανόηση της εργασίας.

Πιο συγκεκριμένα θα γίνει μια συνοπτική περιγραφή του συστήματος, θα αναφερθεί ένα σενάριο χρήσης του με τη βοήθεια ενός σχήματος και στη συνέχεια θα παρουσιαστούν τα στάδια του συστήματος μαζί με μια αναφορά των προβλημάτων που αντιμετωπίζει το καθένα.

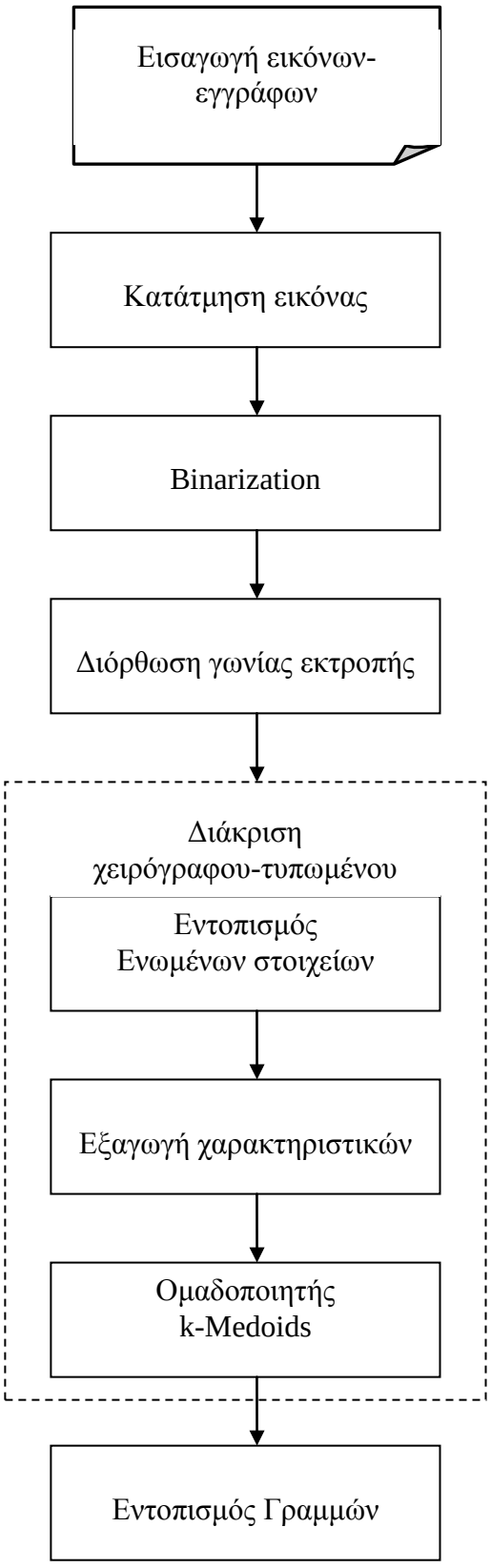
2.1 Σύντομη παρουσίαση του συστήματος

Στα πλαίσια της εργασίας παρουσιάζεται ένα σύστημα για την εξαγωγή συγκεκριμένης πληροφορίας από ληξιαρχικά αρχεία. Στη συνέχεια θα γίνει μια σύντομη περιγραφή του συστήματος αυτού.

Το προτεινόμενο σύστημα αποτελείται από πέντε βασικά στάδια, τα οποία μπορεί να θεωρηθούν και ως στάδιο προ-επεξεργασίας ιστορικών μεικτών εγγράφων – εικόνων. Στο τελευταίο στάδιο γίνεται η εξαγωγή της συγκεκριμένη πληροφορίας (όνομα, ημερομηνία κτλ.), η οποία αναφέρεται σε ένα είδους αρχείο. Στην εργασία αυτή χρησιμοποιήθηκε το αρχείο της Σάμου για όλα τα στάδια του συστήματος ενώ

το αρχείο της Κεφαλλονιάς στα στάδια 2-5 αφού το πρώτο βήμα είναι περιττό για αυτά τα αρχεία.

Η είσοδος του συστήματος είναι εικόνα εγγράφου και η έξοδος εικόνες λέξεων που περιέχουν τη ζητούμενη πληροφορία.



Διάγραμμα 1: Προτεινόμενο σύστημα

2.2 Σύντομη παρουσίαση σταδίων του συστήματος

Το πρώτο στάδιο είναι η κατάτμηση της εικόνας (image segmentation). Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο οι εικόνες των αρχείων της Σάμου αποτελούνται από δυο εγγραφές. Επίσης όπως φαίνεται στο Σχήμα 11 κάποια στοιχεία εμφανίζονται σε άλλο πλαίσιο στα δεξιά της εικόνας. Έτσι κρίθηκε απαραίτητη η κατάτμηση της εικόνας ώστε να χωριστούν οι δυο διαφορετικές εγγραφές αλλά και η κάθετη κατάτμηση της για να γίνει ξεχωριστή διαχείριση του πλαισίου.

Το επόμενο στάδιο αποτελείται από την ψηφιοποίηση του συστήματος (binarization). Σε αυτό το στάδιο οι τιμές των εικονοστοιχείων (pixels) της εικόνας μετατρέπονται σε 0 ή 1 (μαύρο ή άσπρο αντίστοιχα) και αυτό γίνεται γιατί σε τέτοιου είδους εικόνα η ανάδειξη ουσιωδών χαρακτηριστικών είναι πού πιο απλή. Επίσης με τη διαδικασία που έχει επιλεχθεί για binarization αφαιρούμε από την εικόνα πολλούς θορύβους. Όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο τα ιστορικά έγγραφα λόγω παλαιότητας πάσχουν από κάποια προβλήματα. Το στάδιο αυτό λοιπόν αυτό μπορεί να αντιμετωπίσει πολλά από αυτά. Για παράδειγμα απαλείφονται οι κηλίδες που έχουν προκληθεί από μελάνι ή υγρασία, επίσης σβήνονται τα μελάνια που έχουν απορροφηθεί από την πίσω πλευρά και τα τσακίσματα ή σκισίματα ενός εγγράφου συνήθως γίνονται άσπρα. Όταν γενικά αναφέρεται σβήσιμο ή απαλοιφή θορύβου εννοείται σε εκείνα τα σημεία της εικόνας τα εικονοστοιχεία μετατρέπονται σε άσπρα.
Ακολουθεί το στάδιο της διόρθωσης γωνίας εκτροπής, κατά την οποία διορθώνεται η κλίση του εγγράφου. Έχει παρατηρηθεί ότι ένα έγγραφο μπορεί να είναι υπό κλίση είτε από κακά ψηφιοποίηση είτε από λάθος τυπογραφικό ή ανθρώπινο. Αυτή η διαδικασία κρίνεται απαραίτητη ώστε να μπορεί κανείς στην συνέχεια να εξάγει αντικειμενικότερα χαρακτηριστικά από την εικόνα.

Το σύστημα συνεχίζει με το στάδιο της διάκρισης τυπωμένου- χειρόγραφου το οποίο αποτελείται από τρία τμήματα: αναγνώριση συνδεδεμένων στοιχείων, εξαγωγή χαρακτηριστικών, ομαδοποιητής k-medoids.

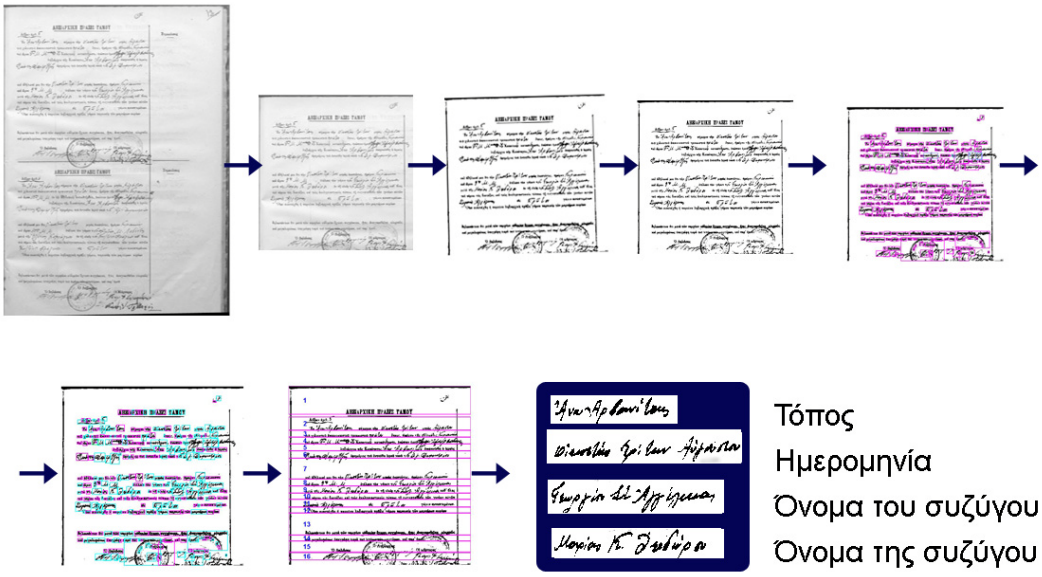
Στην αναγνώριση συνδεδεμένων στοιχείων (connected components) εντοπίζονται connected components και καθαρίζεται η εικόνα από κάποιους θορύβους. Πιο συγκεκριμένα απαλείφονται οι γραμμές που βρίσκονται συνήθως αριστερά της

εικόνας (από κακό scanning) αλλά και οι γραμμές που βρίσκονται στο έγγραφο.

Στη διαδικασία εξαγωγής χαρακτηριστικών, εξάγονται χαρακτηριστικά από κάθε συνδεδεμένο στοιχείο του προηγούμενο βήματος. Τα χαρακτηριστικά αυτά εκμεταλλεύονται τις διαφοροποιήσεις και τα προβλήματα του χειρόγραφου σε σχέση με το τυπωμένο κείμενο. Για παράδειγμα το ύψος των γραμμάτων, όπως επίσης και η κλίση του χειρόγραφου σε σχέση με το τυπωμένο.

Το τρίτο τμήμα του σταδίου αποτελείται από τον ομαδοποιητή k-medoids για τη διάκριση τυπωμένου- χειρόγραφου. Τα χαρακτηριστικά που εκμεταλλεύεται ο αλγόριθμος είναι αυτά του προηγούμενου τμήματος. Η έξοδος του σταδίου αυτού του συστήματος είναι ο χαρακτηρισμός των συνδεδεμένων στοιχείων ως χειρόγραφου ή τυπωμένου.

Επόμενο στάδιο είναι αυτό της κατάτμησης σε γραμμές. Με αυτό το στάδιο επισημαίνονται οι γραμμές της εικόνας-εγγράφου. Τα δεδομένα των γραμμών βοηθούν μετέπειτα το σύστημα να εντοπίζει την ζητούμενη πληροφορία. Αντιμετωπίζονται δηλαδή προβλήματα των μεικτών εγγράφων όπως είναι οι περιττές χειρόγραφες σημειώσεις, η μη τήρηση των ορίων από το συγγραφέα κ.α..



Σχήμα 12: Σενάριο συστήματος

Κεφάλαιο 3

Περιγραφή συστήματος

Όπως έχει ήδη τονιστεί η σημερινή εποχή κατακλύζεται από τεράστιες ποσότητες ψηφιακών εγγράφων. Για να αξιοποιηθούν από διάφορα συστήματα όπως η ανάκτηση εγγράφων ή η Οπτική Αναγνώριση Χαρακτήρων (OCR) θα πρέπει να εντοπιστεί το κείμενο που περιέχουν.

Έχουν γίνει πολλές ερευνητικές προσπάθειες για να αντιμετωπιστεί το πρόβλημα με ποικίλους τρόπους. Υπάρχουν δυο ειδών τεχνικές :

- οι από κάτω προς τα πάνω [1, 2] οι οποίες πρώτα τμηματοποιούν την εικόνα και μετά την ενώνουν με κάποια κριτήρια και,
- οι από πάνω προς τα κάω [3] οι οποίες με επαναληπτικές μεθόδους τμηματοποιούν την εικόνα.

Στην προκειμένη εργασία εντοπίζεται συγκεκριμένο κείμενο σε μεικτά έγγραφα και στο παρόν κεφάλαιο θα αναλυθούν όλα τα βήματα που ακολουθεί το σύστημα.

3.1 Κατάτμηση εικόνας-εγγράφου

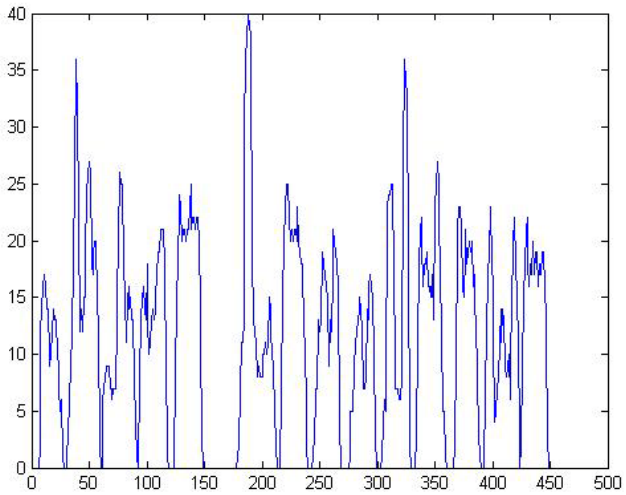
Το στάδιο κατάτμησης εικόνων είναι αναγκαίο για το αρχείο των Άνω Αρβανιτών. Οι εικόνες εγγράφου του αρχείου περιέχουν δυο εγγραφές και συμπληρωματικές πληροφορίες σε πλαίσιο στα δεξιά τους.

Η διαδικασία κατατμήζει την εικόνα οριζόντια για να διαχωρίσει τις δυο εγγραφές και στην συνέχεια κατατμήζει και το πλαίσιο σε κάθε μια από τις εγγραφές. Για την παραπάνω διαδικασία θεωρήθηκε απαραίτητη η εξαγωγή κάθετου και οριζόντιου ιστογράμματος (horizontal / vertical histogram ή projection profile).

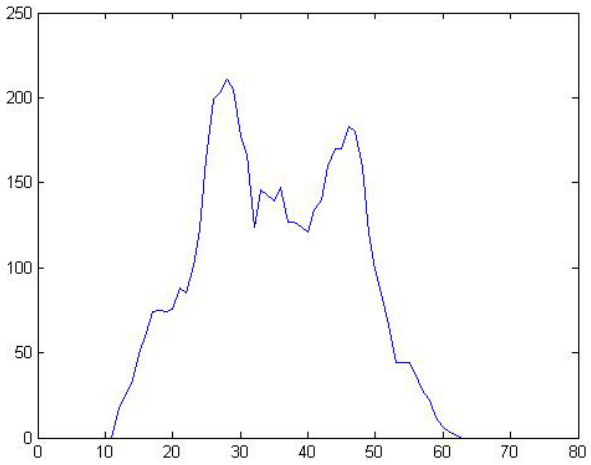
Τα ιστογράμματα χρησιμοποιούνται από την βιβλιογραφία για διόρθωση της γωνίας εκτροπής, της κατάτμησης χαρακτήρων ή λέξεων και σε πολλά άλλα στάδια στην Επεξεργασία Εικόνας

Αν θεωρηθεί η εικόνα ως πίνακας διαστάσεων m x n, ως οριζόντιο ιστόγραμμα ορίζεται το πλήθος των μαύρων pixel που υπάρχουν σε κάθε γραμμή της εικόνας και ως κάθετο ιστόγραμμα το πλήθος των μαύρων pixel που υπάρχουν σε κάθε στήλη της εικόνας.

(α)



(β)



(γ)

Σχήμα 13:Παράδειγμα ιστογραμμάτων: (α) εικόνα λέξης, (β) κάθετο ιστόγραμμα, (γ) οριζόντιο ιστόγραμμα

Έτσι μπορεί κανείς να έχει την πληροφορία για την κατανομή του μελανιού σε κάθε στήλη και γραμμή της εικόνας εγγράφου. Στο Σχήμα 13 βλέπουμε τα ιστογράμματα για μια εικόνα αρχείου.

Τα βήματα της κατάτμησης της εικόνας είναι τα εξής :

1. κατάτμηση εγγραφών :

- εξαγωγή οριζόντιου ιστογράμματος
- εύρεση μέγιστης τιμής
- οριζόντια κατάτμηση

2. κατάτμηση πλαισίων για κάθε εγγραφή :

- εξαγωγή κάθετου ιστογράμματος
- εύρεση μέγιστης τιμής
- κάθετη κατάτμηση.

Η παραπάνω ακολουθία βημάτων έχει ως αποτέλεσμα τέσσερις εικόνες όπως φαίνεται στο παρακάτω Σχήμα 14.

Σχήμα 14: Παράδειγμα κατάτμησης εικόνας εγγράφου και παρουσίαση ιστογραμμάτων

3.2 Binarization

Η δυαδικοποίηση των εικόνων είναι από τα πιο βασικά βήματα του συστήματος, αφού είναι από τα πρώτα στάδια του συστήματος και οποιοδήποτε σφάλμα στο επίπεδο αυτό εξαπλώνεται και στα επόμενα επίπεδα επεξεργασίας. Επίσης όπως έχει ήδη τονιστεί σε προηγούμενο κεφάλαιο σε αυτό το βήμα θα πρέπει να αντιμετωπιστούν οι φθορές, ο θόρυβος και άλλα προβλήματα που έχουν οι εικόνες εγγράφων.

Οι μέθοδοι δυαδικοποίησης χωρίζονται στις εξής κατηγορίες :

- καθολικές (global)

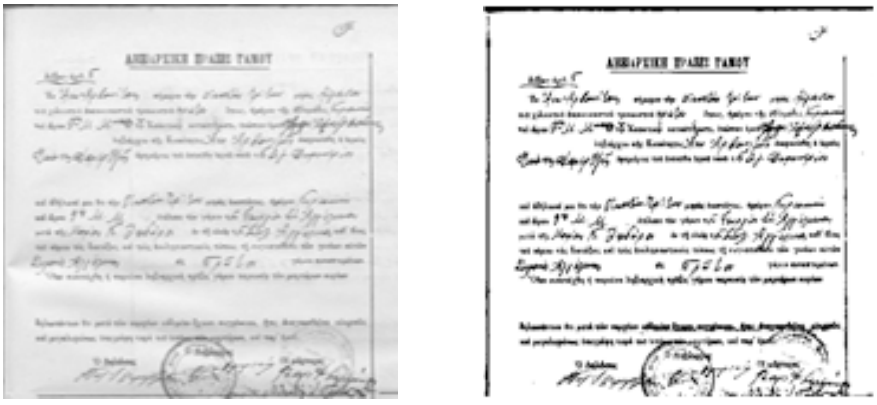
- τοπικές (local)
- υβριδικές (hybrid).

Οι καθολικές μέθοδοι [4] ορίζουν ένα κατώφλι, αναλύοντας τα στατιστικά χαρακτηριστικά της εικόνας. Με το κατώφλι αυτό χωρίζονται οι τιμές έντασης σε δυο, το φόντο και το κείμενο. Αυτό το είδος μεθόδων δεν μπορούν να αντιμετωπίσουν ικανοποιητικά κάποια χαρακτηριστικά των εικόνων εγγράφων, όπως για παράδειγμα το μεταβαλλόμενο φόντο.

Οι τοπικές μέθοδοι [5] καθορίζουν κατώφλι για κάθε pixel βάση των τοπικών στατιστικών χαρακτηριστικών της εικόνας. Αυτές οι μέθοδοι αντιμετωπίζουν κάποια ήδη θορύβου, αλλά δεν μπορούν συνήθως να διαχειριστούν καλά τις περιοχές με πολύ έντονη σκίαση.

Τέλος, οι υβριδικές μέθοδοι συνδυάζουν και τις δυο παραπάνω μεθόδους. Στο στάδιο του binarization θα χρησιμοποιήσουμε υβριδική μέθοδο η οποία θα παρουσιαστεί λεπτομερώς στη συνέχεια.

Για το binarization των εικόνων χρησιμοποιήθηκε ο Hybrid Iterative Global Thresholding (Hybrid IGT) αλγόριθμος [6], ο οποίος είναι συνδυασμός τοπικού και καθολικού κατωφλίου. Επιλέχθηκε να εφαρμοστεί ο παραπάνω αλγόριθμος γιατί αντιμετωπίζει επιτυχώς τα περισσότερα προβλήματα των ιστορικών εγγράφων και μπορεί να προσαρμοστεί στις ιδιαιτερότητες κάθε αρχείου που χρησιμοποιείται στην εργασία. Αρχικά χρησιμοποιείται ο καθολικός αλγόριθμος Hybrid IGT και στη συνέχεια εντοπίζονται τα μέρη που παραμένει ο θόρυβος και επεξεργάζονται χωριστά.



Σχήμα 15: Binarization

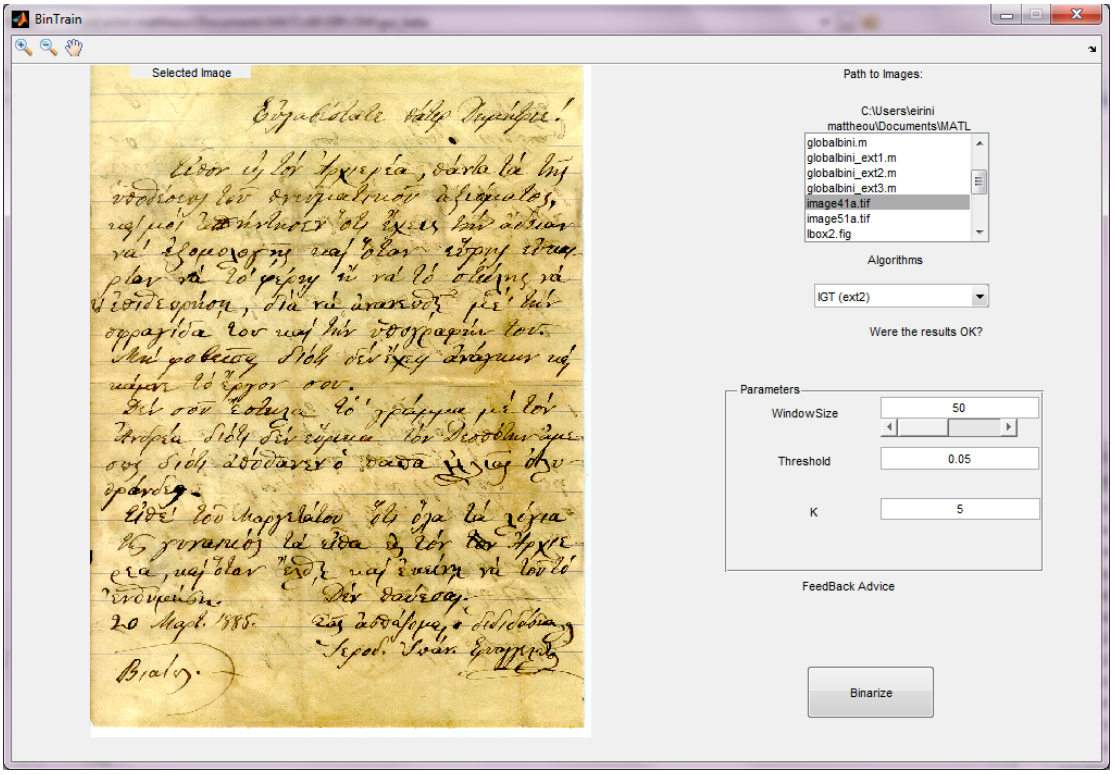
Αναλυτικότερα ο αλγόριθμος αποτελείται από τα παρακάτω βήματα :

- Επαναληπτική διαδικασία
  - υπολογίζει το μέσο όρο τιμών των pixel
  - τον αφαιρεί από κάθε pixel και κάνει stretch ώστε τα pixel που απομένουν να διανεμηθούν στην κλίμακα του γκρι.
  - Η επανάληψη σταματάει όταν $\left|T_i - T_{i-1}\right| < TH$ αν T είναι ένα όριο (threshold) σε κάθε επανάληψη.

- Στη συνέχεια εντοπίζονται οι περιοχές που παραμένει ο θόρυβος.
  - Χωρίζεται η εικόνα σε τμήματα μεγέθους *nxn* και υπολογίζεται η συχνότητα μαύρων pixel.
  - Τα τμήματα που ικανοποιούν τη σχέση $f(S) > m + ks$ επιλέγονται και επαναλαμβάνεται η παραπάνω διαδικασία.

Όπου m και s είναι το μέσο και η τυπική απόκλιση της συχνότητας των μαύρων pixel.

Επιπλέον για να επιλεχθούν με ακρίβεια και ευκολία οι παράμετροι που δόθηκαν στον αλγόριθμο χρησιμοποιήθηκε μια αλληλεπιδραστική εφαρμογή ‘ A tool for Tuning Binarization Techniques ’ [7] . Η εφαρμογή αυτή δίνει την δυνατότητα στο χρήστη να εισάγει την εικόνα που επιθυμεί να κάνει binarization και να επιλέξει τις παραμέτρους. Μπορεί στη συνέχεια να δει το αποτέλεσμα του αλγορίθμου και αλληλεπιδρώντας με το πρόγραμμα μπορεί να καταλήξει με τις κατάλληλες παραμέτρους για το είδος της εικόνας που επιθυμεί.



Σχήμα 16: Στιγμιότυπο εφαρμογής

Για το αρχείο των Άνω Αρβανιτών χρησιμοποιήθηκαν οι εξής τιμές :

- n=25
- k=5
- Threshold=1.8

Ενώ για το αρχείο της Κεφαλλονιάς χρησιμοποιήθηκαν οι παρακάτω τιμές :

- n=80
- k=3
- Threshold=0.12

3.3 Διόρθωση γωνίας εκτροπής

Σε ένα έγγραφο μπορεί να εμφανιστεί εσφαλμένη γωνία κλίσης είτε από κακή ψηφιοποίηση είτε από τυπογραφικό λάθος. Ωστόσο οι διαδικασίες όπως η κατάτμηση, η ταξινόμηση και άλλες είναι ευαίσθητες στην ύπαρξη κλίσεων στο έγγραφο. Για αυτό απαραίτητο βήμα του συστήματος είναι και η διόρθωση γωνίας εκτροπής.

Στην βιβλιογραφία υπάρχουν αρκετές μελέτες που προσπαθούν να διαχειριστούν επιτυχώς το πρόβλημα οι οποίες χωρίζονται στις εξής βασικές κατηγορίες :

- προβολές προφίλ (projection profile)
- μετασχηματισμός Hough (Hough transform)
- ομαδοποίηση κοντινότερου γείτονα (nearest neighbor clustering)
- υβριδικά συστήματα.

Στην πρώτη κατηγορία έχουν χρησιμοποιηθεί προβολές προφίλ σε συνδυασμό με μετασχηματισμούς Fourier [8], και καθορίζεται η γωνία εκτροπής από την πυκνότητα του διαστήματος Fourier. Επίσης έχουν χρησιμοποιηθεί οριζόντιες και κάθετες προβολές και συνδυασμοί τους με τεχνικές συνδεδεμένων στοιχείων [9].

Οι τεχνικές που ανήκουν στην δεύτερη κατηγορία χρησιμοποιούν το μετασχηματισμό Hough [10] για να βρουν την κλίση του εγγράφου σε συνδυασμό με τεχνικές συνδεδεμένων στοιχείων αλλά και άλλες παραλλαγές.

Στην τρίτη κατηγορία ανήκουν αυτές που χρησιμοποιούν ομαδοποιητή κοντινότερου γείτονα που συνίστανται για μικρές γωνίες εκτροπής.

Στη τελευταία κατηγορία χρησιμοποιούνται παραπάνω από μια από τις προηγούμενες τρεις κατηγορίες.

Στο παρόν σύστημα για το βήμα της διόρθωσης γωνίας εκτροπής χρησιμοποιήθηκε ο αλγόριθμος που στηρίζεται στην κατανομή Wigner-Ville [11]. Πιο αναλυτικά τα βήματα του αλγορίθμου είναι :

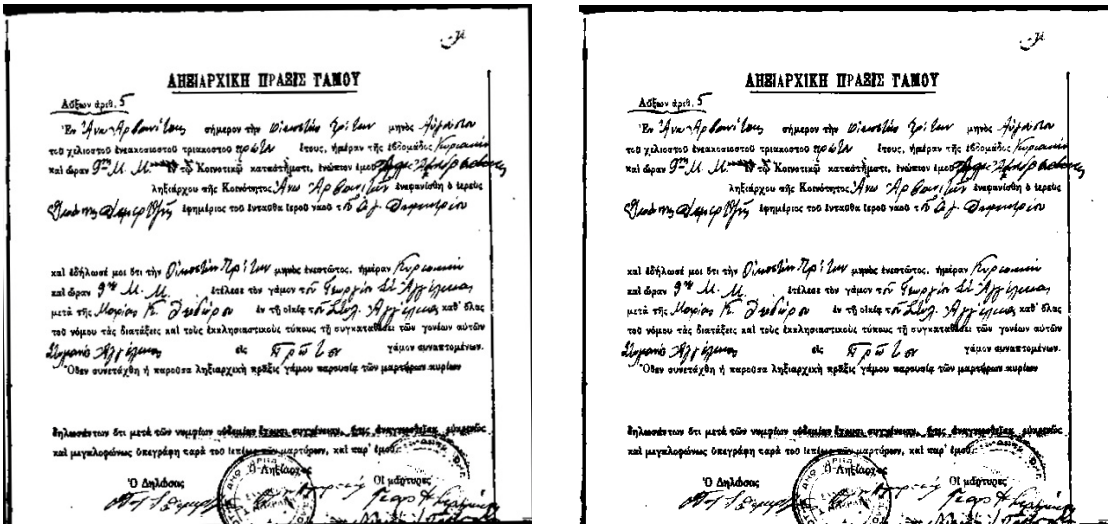
- Η εικόνα περιστρέφεται σε γωνία ±θ και υπολογίζεται η οριζόντια προβολή
- Για κάθε προβολή υπολογίζονται τα η WVD και η μέγιστη ένταση
- Η γωνία της οποίας η προβολή παρουσιάζει τη μέγιστη ένταση επιλέγεται και η εικόνα περιστρέφεται σε αυτή.
- Η διαδικασία επαναλαμβάνεται για μικρότερες γωνίες
- Στο σύστημα χρησιμοποιήθηκε η γωνία θ=10.

```
Angle=-84
while Angle<=84{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+12
}
select the maximum intensity curve[angle1] with the highest peak
rotate_page[angle1*12]

Angle=-6
while Angle<=6{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+1
}
select the maximum intensity curve[angle2] with the highest peak
rotate_page[angle2*1]

Angle=-0.5
while Angle<=0.5{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+0.1
}
select the maximum intensity curve[angle3] with the highest peak
rotate_page[angle3*0.1]
```

Σχήμα 17: Αλγόριθμος



Σχήμα 18: Διόρθωση γωνίας εκτροπής

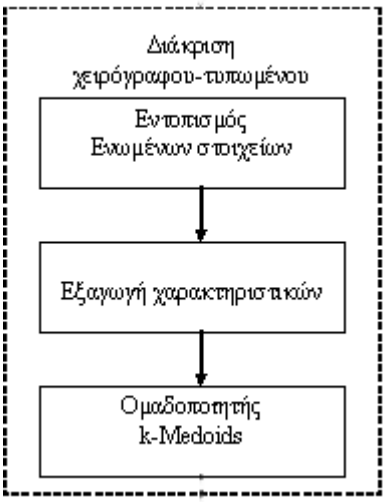
3.4 Διάκριση Χειρόγραφου – Τυπωμένου

Στο στάδιο αυτό το σύστημα ενημερώνει για το αν ένα τμήμα του κειμένου της εικόνας είναι χειρόγραφο ή όχι. Γνωρίζοντας αν ένα κείμενο είναι τυπωμένο ή μη, μπορεί κανείς σε επόμενο στάδιο να απορρίψει ή να δεχτεί το ένα από τα δυο είδη κειμένου. Για παράδειγμα, αν κάποιος αναζητεί εικόνα κειμένου που περιέχει την ημερομηνία και είναι χειρόγραφο, το σύστημα θα ψάξει μόνο στο χειρόγραφο κείμενο και όχι στο τυπωμένο.

Επίσης η γνώση του συστήματος για το είδος κειμένου είναι απαραίτητη αν το ίδιο ενσωματωθεί σε κάποιο σύστημα OCR, αφού η αντιμετώπιση των χειρόγραφων χαρακτήρων είναι τελείως διαφορετική από αυτήν των τυπωμένων για την αναγνώριση του.

Έχουν γίνει αρκετές έρευνες για την διάκριση χειρόγραφου – τυπωμένου και έχουν χρησιμοποιηθεί διάφορες τεχνικές κατηγοριοποίησης (classification). Στη βιβλιογραφία υπάρχουν μέθοδοι ταξινόμησης νευρωνικών δικτύων, ελάχιστης απόστασης, Hidden Markov Model (HMM) και άλλες. Στο προτεινόμενο σύστημα δεν έγινε ταξινόμηση αλλά ομαδοποίηση με τον αλγόριθμο k – Εσωτερικών Αντιπροσώπων (k-Medoids). Πριν από τη διαδικασία της ομαδοποίησης γίνεται ο εντοπισμός των ενωμένων στοιχείων, από αυτά η εξαγωγή κάποιων χαρακτηριστικών και στη συνέχεια αυτά εισάγονται στο k-Medoids. Στο υποκεφάλαιο

αυτό θα περιγραφούν αναλυτικά οι παραπάνω διαδικασίες.

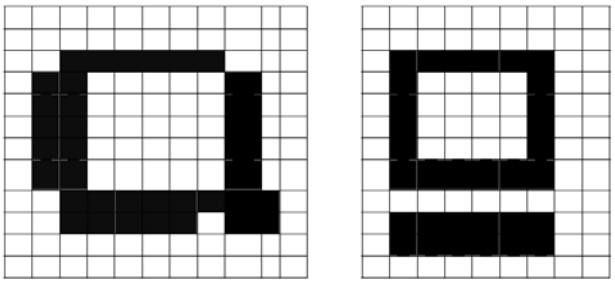


Διάγραμμα 2: το υποσύστημα Διάκρισης χειρόγραφου τυπωμένου

3.4.1 Εντοπισμός συνδεδεμένων στοιχείων (CC)

Τα συνδεδεμένα στοιχεία (converted components) στον κλάδο της επεξεργασίας εικόνας επιτρέπουν την κατηγοριοποίηση των εικονοστοιχείων της εικόνας σε ξεχωριστές ομάδες.

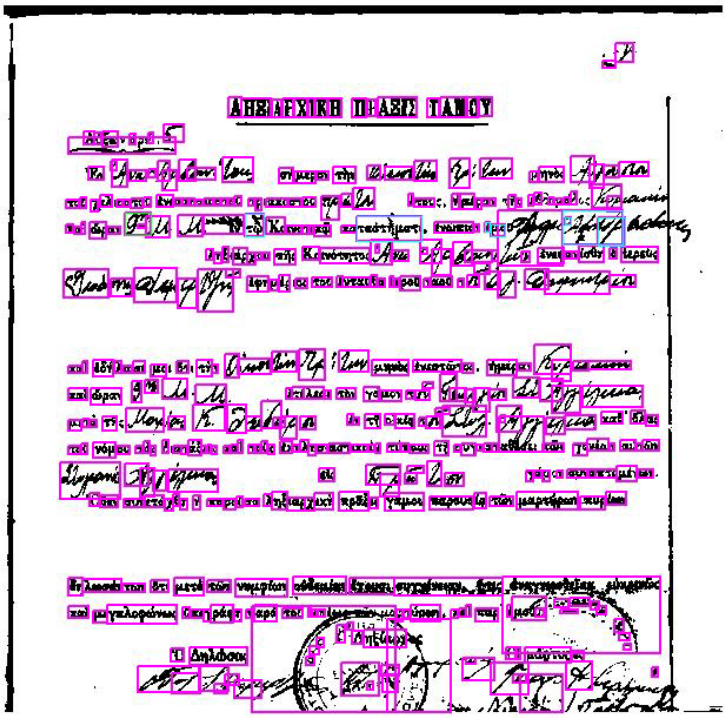
Πιο συγκεκριμένα αν θεωρήσουμε σύνολο S εικονοστοιχείων μια δυαδικής εικόνας με την ίδια τιμή (0 ή 1), δυο εικονοστοιχεία του S, p1 και p2 ονομάζονται συνδεδεμένα όταν υπάρχει διαδρομή από εικονοστοιχεία του S που οδηγεί από το p1 στο p2. Δηλαδή ένα σύνολο εικονοστοιχείων λέγεται συνδεδεμένο στοιχείο όταν όλα τα εικονοστοιχεία του είναι μεταξύ τους συνδεδεμένα.



(α) (β)

Σχήμα 19: (α) ένα ενωμένο στοιχείο, (β) δύο ενωμένα στοιχεία

Εντοπίζονται τα συνδεδεμένα στοιχεία της εικόνας-εγγράφου , το κείμενο έχει χωριστεί κυρίως σε γράμματα αλλά και συλλαβές. Τα γράμματα χωρίζονται καλύτερα στο τυπωμένο κείμενο αφού δεν παρατηρούνται συνεχή εικονοστοιχεία ανάμεσα στους χαρακτήρες. Αυτό δεν συμβαίνει συνήθως στο χειρόγραφο αφού εκεί έχουμε γραμμή χωρίς περιορισμούς.



Σχήμα 20: Εντοπισμός ενωμένων στοιχείων

Αφού έχουν εντοπιστεί τα συνδεδεμένα στοιχεία εκτελείται μια επιπλέον διαδικασία για να εντοπιστεί κάποιο είδος θορύβου αλλά και να αυξηθεί η ταχύτητα σε επόμενο στάδιο. Δηλαδή :

- Τα πολύ μεγάλα CC απορρίπτονται. Αυτό επιτυγχάνεται με την απόρριψη των CC που ικανοποιούν τις παρακάτω συνθήκες :

$$CC_h > \frac{\mathrm{Im}\, h}{3}$$

$$CC_w > \frac{\mathrm{Im}\, w}{3}$$

όπου CC_h και CC_w είναι το ύψος και το πλάτος αντίστοιχα του CC και Im_h, Im_w το ύψος και το πλάτος της εικόνας εγγράφου.

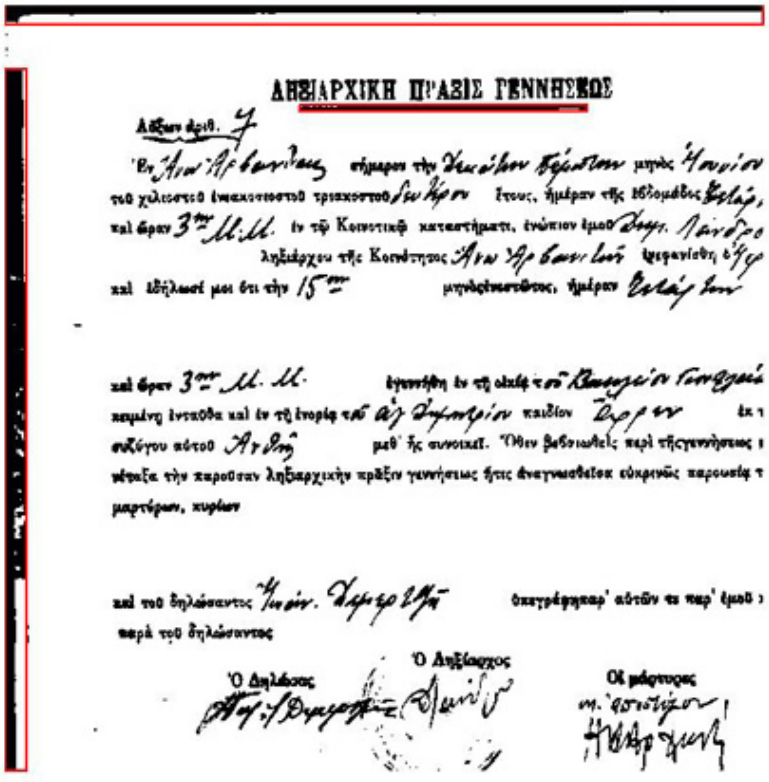
- Τα CC που αποτελούν γραμμή (θόρυβο ή τυπωμένη ή χειρόγραφη) απορρίπτονται με βάση τις παρακάτω συνθήκες :

$$\frac{CC_h}{CC_w} > 4.5$$

$$\frac{CC_w}{CC_h} > 4.5$$

- Τα CC που είναι πολύ μικρά απορρίπτονται. Πιο συγκεκριμένα τα CC που ικανοποιούν την παρακάτω συνθήκη :

  NoP < 15 όπου NoP είναι ο αριθμός των μαύρων pixel του CC.


ΔΗΜΑΡΧΙΚΗ ΠΡΑΞΙΣ ΓΕΝΝΗΣΕΩΣ

Σχήμα 21: Εντοπισμός θορύβου γραμμών

3.4.2 Επιλογή χαρακτηριστικών

Προκειμένου να οργανωθεί μια διαδικασία ομαδοποίησης πρέπει να γίνει η επιλογή χαρακτηριστικών (feature selection). Τα χαρακτηριστικά αυτά θα πρέπει να επιλεγούν κατάλληλα έτσι ώστε να κωδικοποιηθεί όσο τον δυνατόν περισσότερη πληροφορία σχετικά με το υπό εξέταση πρόβλημα. Για κάθε συνδεδεμένο στοιχείο θα δημιουργηθεί ένα διάνυσμα που θα αποτελείται από τα χαρακτηριστικά που θα εξαχθούν από αυτό.

Σε έρευνες για διάκριση χειρόγραφου – τυπωμένου έχουν χρησιμοποιήθηκαν διάφορα χαρακτηριστικά όπως χαρακτηριστικά συμμετρίας [12] των CC, ευθύτητας γραμμής ιδιοδιανύσματα [13]. Στην εργασία χρησιμοποιήθηκαν τα εξής χαρακτηριστικά :

- λόγος πλάτους και ύψους CC
- ύψος CC
- πυκνότητα μαύρων pixel
- προσανατολισμός (orientation)

τα χαρακτηριστικά αυτά θα αναλυθούν παρακάτω.

(α) (β)

Σχήμα 22: Παράδειγμα εξαγωγής χαρακτηριστικών: (α)τυπωμένου (β) χειρόγραφου

Πλάτος / ύψος CC

Ο λόγος πλάτους προς ύψους του CC είναι κατά βάση μικρότερος όταν περιέχεται σε αυτό τυπωμένο κείμενο και αυτό γιατί το χειρόγραφο έχει ιδιαιτερότητες (ενωμένοι χαρακτήρες, γραφή προς τα πάνω κ.α.).Για παράδειγμα το αντίστοιχο χαρακτηριστικό που εξάγεται για της εικόνες του Σχήματος 22 είναι για το τυπωμένο 0.63 και για το χειρόγραφο 1.

Ύψος

Το ύψος της γραφής των κειμένων χειρόγραφου και τυπωμένου διαφέρει. Έχει παρατηρηθεί ότι το ύψος στο χειρόγραφο κείμενο συνήθως είναι μεγαλύτερο από αυτό του τυπωμένου. Αυτό το χαρακτηριστικό κανονικοποιείται με βάση το ύψος της αχικής εικόνας εγγράφου. Για παράδειγμα το αντίστοιχο χαρακτηριστικό που εξάγεται για της εικόνες του Σχήματος 22 για το τυπωμένο είναι 1 και για το χειρόγραφο 0. 97.

Πυκνότητα μαύρων pixel

Αυτό το χαρακτηριστικό είναι το πλήθος των μαύρων pixel του αντικειμένου που βρίσκεται στο CC κανονικοποιημένο ως προς το συνολικό αριθμό των pixel που περικλείει το CC.

Γενικά στο τυπωμένο κείμενο παρουσιάζεται μεγαλύτερη πυκνότητα σε μαύρα pixel από ότι το χειρόγραφο, εκτός αν ο συγγραφέας γράφει πολύ έντονα ή με χοντρό στυλό. Η πυκνότητα που εξάγεται για της εικόνες του Σχήματος 22 είναι για το τυπωμένο 0.65 και για το χειρόγραφο 0.173.

Προσανατολισμός

Το χαρακτηριστικό του προσανατολισμού είναι η γωνία που σχηματίζεται μεταξύ του άξονα x και του κύριου άξονα της έλλειψης που έχει την ίδια ροπή δευτέρου βαθμού με το αντικείμενο. Για τις εικόνες του Σχήματος 22 ο προσανατολισμός είναι, για το τυπωμένο 0.007 και για το χειρόγραφο 0.78.

Στο παρακάτω Σχήμα 23 φαίνονται οι άξονες και ο προσανατολισμός της έλλειψης. Στο αριστερό μέρος της εικόνας έχει σχηματιστεί η έλλειψη που δημιουργεί το αντικείμενο της εικόνας, και αριστερά φαίνεται η ίδια έλλειψη με σχεδιασμένα κάποια χαρακτηριστικά : οι μπλε γραμμές είναι οι άξονες, οι κόκκινες τελείες είναι οι εστίες της έλλειψης και ο προσανατολισμός είναι η γωνία που σχηματίζεται από την οριζόντια γραμμή και τον κύριο άξονα.

(α) (β)

Σχήμα 23:Παράδειγμα προσανατολισμού

3.4.3 k-Medoids

Στη βιβλιογραφία για την διάκριση χειρόγραφου-τυπωμένου χρησιμοποιούνται ταξινομητές όπως ο SVM [14], k-means [15] κ.α.. Στο προτεινόμενο σύστημα δεν χρησιμοποιήθηκαν ταξινομητές αλλά ομαδοποιητής δηλαδή εκμάθηση χωρίς επίβλεψη. Από όσο μπορούμε να γνωρίζουμε δεν έχει γίνει clustering για διάκριση χειρόγραφου-τυπωμένου.

Οι τεχνικές ομαδοποίησης ουσιαστικά αποτελούνται από αλγορίθμους που εφαρμόζονται σε ένα σύνολο από ετερογενή δεδομένα, προκειμένου να δημιουργήσουν ομογενείς ομάδες στηριζόμενοι σε κάποιο δεδομένο μοντέλο ή κάποιο μέτρο ομοιότητας [16].

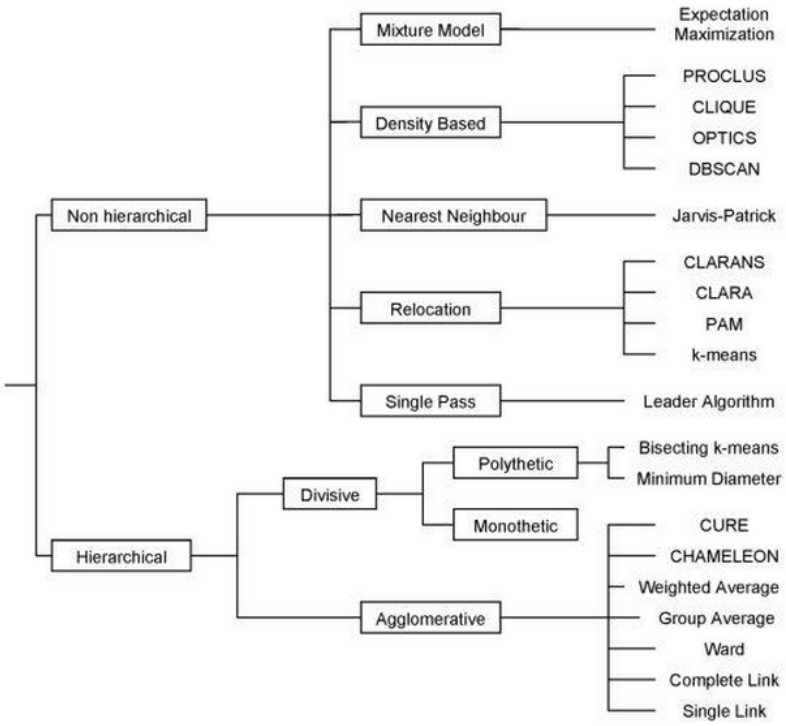
Χαρακτηριστικό των τεχνικών ομαδοποίησης αποτελεί το γεγονός ότι πρόκειται για μη επιβλεπόμενες διαδικασίες. Τα γενικά βήματα που περιλαμβάνει μια διαδικασία ομαδοποίησης είναι τα εξής :

- Εξαγωγή κατάλληλων χαρακτηριστικών που περιγράφουν κάθε στοιχείο των δεδομένων
- Επιλογή κατάλληλου μέτρου ομοιότητας
- Ομαδοποίηση των δεδομένων με χρήση κατάλληλα επιλεγμένης τεχνικής ομαδοποίησης

Οι τελικές ομάδες μπορεί να είναι επικαλυπτόμενες ή μη επικαλυπτόμενες. Όταν το σύνολο δεδομένων αναλύεται με επαναληπτικό τρόπο , τέτοιο ώστε σε κάθε επαναληπτικό βήμα δύο ομάδες να ενώνονται σε μία, ή παρατηρείται το αντίθετο,

τότε η τεχνική ομαδοποίησης λέγεται ιεραρχική (hierarchical). Αντίθετα όταν το σύνολο των δεδομένων αναλύεται κατά τέτοιο τρόπο ώστε να διαχωριστεί και τελικά να δώσει ένα αριθμό από τελικές ομάδες, τότε η τεχνική λέγεται μη ιεραρχική (non hierarchical).

Μια αναλυτική κατηγοριοποίηση των πιο γνωστών τεχνικών φαίνεται στο Διάγραμμα 3.



Διάγραμμα 3: Τεχνικές ομαδοποίησης

Μετά από σύγκριση των αποτελεσμάτων ακρίβειας μεταξύ αλγορίθμων (Πίνακας 1) καταλήξαμε στον k-medoids (PAM) γιατί είχε την υψηλότερη ακρίβεια. Οι αλγόριθμοι που δοκιμάστηκαν και απορρίφθηκαν είναι οι εξής :

- FCM (Fuzzy c-Means – Ασαφής Αλγόριθμος c-Μέσων)
- LLA (Leaky Learning Algorithm – Αλγόριθμος Διαρρέουσας Ανταγωνιστικής Μάθησης)
- PCM (Possibilistic c-Means – Ενδεχόμενων c-Μέσων)
- Spectral (Φασματική ομαδοποίηση)
- K-means (ISODATA), (k-Μέσων)

- BSAS (Basic Sequential Algorithmic Scheme)
- Ward – dendrogram

	Printed	Handwritten
PAM	88,5	98,1
FCM	90,5	85,3
LLA	87,2	89,4
PCM	85,4	89,3
Spectral	84,3	95,1
k-means	91,6	89,7
BSAS	89,6	86,1
Ward-dedrogram	80,3	94,9

Πίνακας 1: Ακρίβεια μεταξύ αλγορίθμων ομαδοποίησης

Ο k-medoids ανήκει στην κατηγορία αλγορίθμων αυστηρής ομαδοποίησης, όπου κάθε διάνυσμα ανήκει μόνο σε μια ομάδα. Επίσης κάθε ομάδα αντιπροσωπεύεται από ένα διάνυσμα το οποίο επιλέγεται από τα στοιχεία του συνόλου των διανυσμάτων (έστω X) και ονομάζεται εσωτερικός αντιπρόσωπος (medoid). Εκτός από τον εσωτερικό αντιπρόσωπο, κάθε ομάδα περιέχει όλα τα στοιχεία του X τα οποία :

- δεν χρησιμοποιούνται ως εσωτερικοί αντιπρόσωποι σε άλλες ομάδες και
- βρίσκονται εγγύτερα στο εσωτερικό αντιπρόσωπο όλων των ομάδων.

Η ποιότητα της ομαδοποίησης που σχετίζεται με ένα συγκεκριμένο σύνολο εσωτερικών αντιπροσώπων αξιολογείται μέσω της ακόλουθης συνάρτησης.

$$J(\Theta,U)=\sum_{i\in I_{X-\Theta}}\sum_{j\in I_{\Theta}}u_{ij}d(x_i,x_j)$$

$$u_{ij}=\begin{cases}1, \alpha\nu\ d(x_i,x_q)=\min_{q\in I_{\Theta}} d(x_i,x_q)\\ 0,\ \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\alpha\end{cases}$$

όπου ,

- Θ το σύνολο εσωτερικών αντιπροσώπων όλων των ομάδων
- I_{Θ} σύνολο δεικτών των σημείων του X
- $I_{X-\Theta}$ σύνολο των σημείων που δεν είναι εσωτερικοί αντιπρόσωποι.

Η επιλογή του συνόλου των εσωτερικών αντιπροσώπων είναι ισοδύναμη με την ελαχιστοποίηση της J (Θ, U).

Τα πλεονεκτήματα της χρήσης του k-Medoids αλγορίθμου είναι :

- Χρησιμοποιούνται για σύνολα δεδομένων που προέρχονται και από διακριτά πεδία τιμών
- Δεν είναι ευαίσθητος στην ύπαρξη ακραίων σημείων (outliers).

Ο αλγόριθμος δέχεται το πλήθος των κλάσεων στο οποίο θα ομαδοποιηθούν τα CC (δηλαδή δυο αφού θα έχουμε χειρόγραφο ή τυπωμένο) και τα διανύσματα των χαρακτηριστικών του κάθε συνδεδεμένου στοιχείου. Δηλαδή δέχεται σαν είσοδο ένα πίνακα δεδομένων A όπου κάθε δεδομένο καταλαμβάνει μια στήλη στον πίνακα και οι τιμές των χαρακτηριστικών είναι οι γραμμές του.

$$A = \begin{bmatrix} \Lambda ογος\_πλατους - υψους1 & \Lambda ογος\_πλατους - υψους2 & \Lambda ογος\_πλατους - υψουςn \\ Πυκνοτητα\_μαυρων\_pixel1 & Πυκνοτητα\_μαυρων\_pixel2 & Πυκνοτητα\_μαυρων\_pixeln \\ Υψος1 & Υψος2 & Υψοςn \\ Προσανατολισμος1 & Προσανατολισμος2 & Προσανατολισμοςn \end{bmatrix}$$

Σχήμα 24: Παράδειγμα διανυσμάτων χαρακτηριστικών

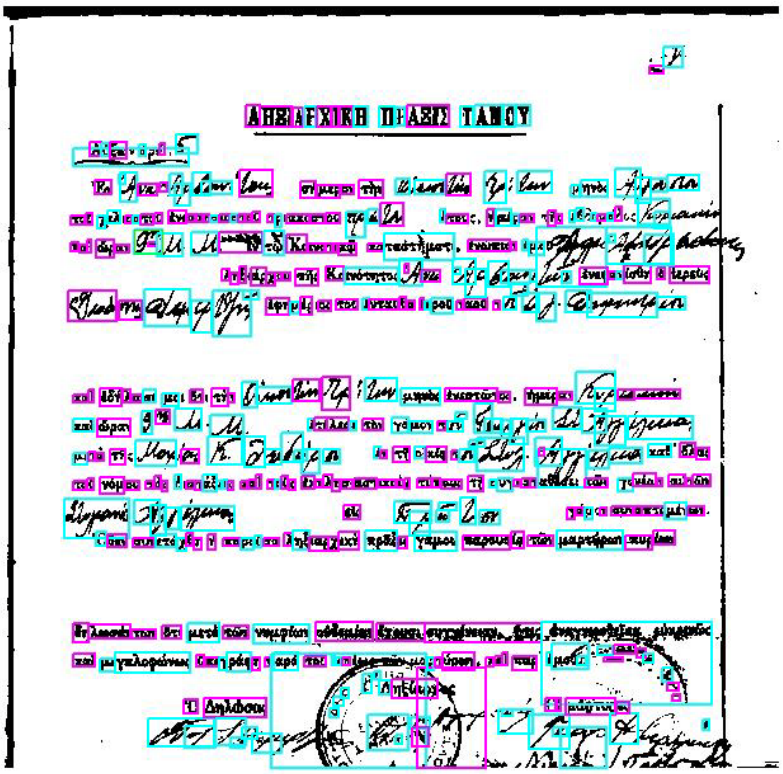
Ο αλγόριθμος που χρησιμοποιήθηκε εφαρμόζεται όπως παρακάτω :

[bell, cost, w, a] = k-medoids (x, m, sed)

όπου

- x είναι μητρώο διαστάσεων l x N που περιέχει ένα διάνυσμα δεδομένων ανά στήλη,
- m είναι ο αριθμός των ομάδων,
- sed είναι ο ακέραιος που χρησιμοποιείται για την αρχικοποίηση της MATLAB συνάρτησης rand,

- bel είναι το N-διάστατο διάστημα, το i-οστό στοιχείο του οποίου περιέχει τη ταυτότητα (αριθμό) της ομάδας στην οποία ανήκει το i-οστό διάνυσμα δεδομένων,
- cost είναι η τιμή της J(Θ) που αντιστοιχεί στην τελική ομαδοποίηση που έδωσε ο αλγόριθμος .
- w είναι lxm που περιέχει ένα medoid ανά στήλη και
- a είναι το m-διάστατο διάνυσμα που περιέχει τον αντίστοιχο αριθμό διανύσματος των medoid.



Σχήμα 25:Διάκριση τυπωμένου χειρόγραφου

3.5 Εντοπισμός γραμμών

Αφού γίνει η ομαδοποίηση των ενωμένων στοιχείων, ακολουθεί το βήμα εντοπισμού γραμμών. Το στάδιο εντοπισμού γραμμών ή κατάτμησης γραμμών χρησιμοποιείται σε πολλές τεχνικές για επεξεργασία εικόνων-εγγράφων. Η πιο γνωστή τεχνική είναι αυτή που χρησιμοποιεί οριζόντια προβολή. Από αυτή εντοπίζονται τα τοπικά ελάχιστα και ορίζονται ως σημεία κατάτμησης της γραμμής. Αυτή η τεχνική όμως δεν

αποδίδει σε κάποια είδη εγγράφων όπως τα χειρόγραφα ή μεικτά έγγραφα αλλά και στα έγγραφα που παρατηρείται θόρυβος ή κλίση. Έχουν προταθεί πολλές τεχνικές που ασχολούνται με την κατάτμηση γραμμής, υπάρχουν τεχνικές που στηρίζονται στη χρήση φίλτρων [17] , κάποιες συνδυάζουν ενωμένα στοιχεία και προβολές [18] και άλλες χρησιμοποιούν Gaussian παράθυρα [19], μετασχηματισμούς Hough [20] κ.α..

Στο προτεινόμενο σύστημα ο εντοπισμός γραμμών γίνεται μετά από τα στάδια της Διόρθωσης γωνίας εκτροπής και τη Διάκριση χειρόγραφου-τυπωμένου.
Εφόσον το προηγούμενο βήμα της διάκρισης παρέχει την πληροφορία για κάθε ενωμένο στοιχείο, στο σημείο αυτό του συστήματος το χειρόγραφο κείμενο θα απορριφθεί ώστε να διευκολυνθεί ο εντοπισμός των γραμμών.

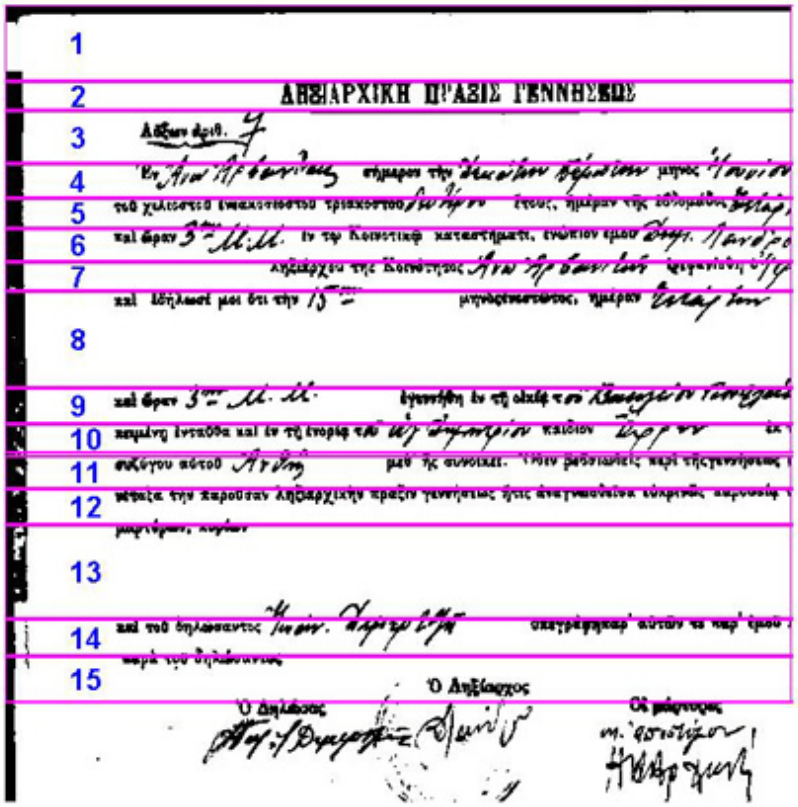
Αναλυτικότερα :

- Διαγράφονται τα ενωμένα στοιχεία που έχουν θεωρηθεί από τον ομαδοποιητή χειρόγραφα.
- Από την εικόνα που προκύπτει, εξάγεται η οριζόντια προβολή.
- Από αυτή επιλέγονται μόνο οι τιμές που είναι μεγαλύτερες από το κατώφλι.
- Ορίζονται τα όρια γραμμής.

Σε αντίθεση με άλλα συστήματα που κόβουν και αποθηκεύουν τις γραμμές κειμένου, αποφασίστηκε να γίνει μια λίστα των γραμμών. Δηλαδή το βήμα εντοπισμού γραμμών θα μας δίνει την πληροφορία του αριθμού των γραμμών αλλά και ποια pixel της εικόνας περικλείονται στη κάθε γραμμή.

Για να μειωθεί ακόμα περισσότερο το υπολογιστικό κόστος και ο όγκος των δεδομένων αρκεί για κάθε γραμμή να αποθηκευτεί ένα διάνυσμα που περιέχει μόνο τα pixel που ορίζουν το ύψος της γραμμής αφού το πλάτος θα είναι πάντα ίσο με το πάτος της εικόνας. Για παράδειγμα έστω ότι μια γραμμή ισούται με ένα πίνακα 30x250 pixel, η πληροφορία που θα αποθηκευτεί θα είναι μια στήλη του παραπάνω πίνακα, δηλαδή ένα διάνυσμα 30 στοιχείων. (εικόνα επεξηγηματική)

Όπως βλέπουμε στο Σχήμα 26 για τη γραμμή δυο αποθηκεύτηκε ένα διάνυσμα μεγέθους 22 που περιέχει τις τιμές 59 μέχρι 80 με βήμα 1.



Σχήμα 26: Ορισμός Γραμμών

Έτσι έχοντας τα παραπάνω δεδομένα για κάθε γραμμή μια σύγκριση με τα αντίστοιχα pixel (κατά ύψος) των συνδεδεμένων στοιχείων μπορεί πάρει τη εξής πληροφορία: σε ποια γραμμή ανήκει το κάθε ενωμένο στοιχείο ή αντίστροφα, ποια ενωμένα στοιχεία περιέχει η κάθε γραμμή.

3.6 Εξαγωγή εικόνων – πληροφορίας

Τα προηγούμενα στάδια του συστήματος μπορούν να εφαρμοστούν σε οποιοδήποτε μεικτό έγγραφο ως προ-επεξεργασία. Σε κάθε είδος μεικτού εγγράφου υπάρχει συγκεκριμένη διάταξη τυπωμένου κειμένου, η οποία και ορίζει στον συγγραφέα τι πληροφορία πρέπει να καταγραφεί (και σε ποια περιοχή του εγγράφου). Χρησιμοποιώντας λοιπόν τις πληροφορίες που έχουν συλλεχθεί από το σύστημα, και αξιοποιώντας κατάλληλα την εκ των πρότερων γνώση της διάταξης του συγκεκριμένου μεικτού εγγράφου, γίνεται εφικτός ο εντοπισμός συγκεκριμένων

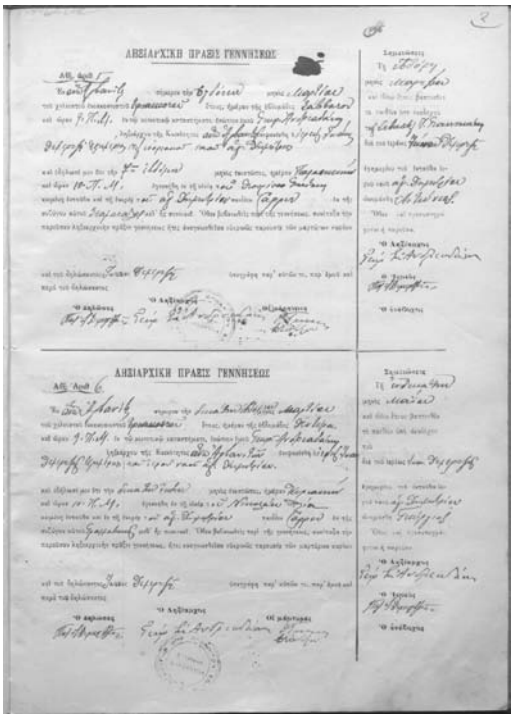
εικόνων πληροφορίας.

Στην ενότητα αυτή θα παρουσιαστούν και θα αναλυθούν τα βήματα για τον εντοπισμό συγκεκριμένης πληροφορίας από το ληξιαρχικό αρχείο Σάμου.

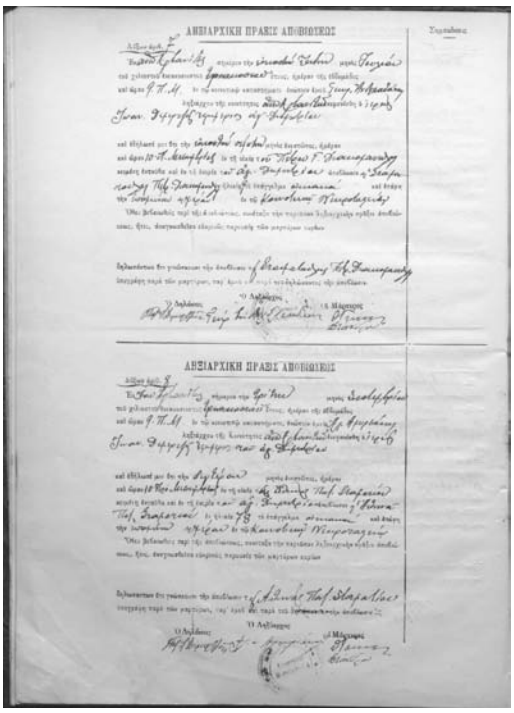
3.6.1 Περιγραφή διάταξης κειμένου

Όπως ήδη αναφέρθηκε κάθε μεικτό έγγραφο έχει μια συγκεκριμένη διάταξη ως προς τη θέση που είναι γραμμένα διάφορα στοιχεία, έχει παρατηρηθεί ότι τα ληξιαρχικά αρχεία κάθε περιοχής / Νομού έχουν μια συγκεκριμένη φόρμα. Τα ληξιαρχικά έγγραφα της Σάμου ακολουθούν και αυτά μια συγκεκριμένη δομή. Πιο συγκεκριμένα στα αρχεία γεννήσεως, στο πάνω μέρος του εγγράφου είναι τυπωμένο το γεγονός για το οποίο γίνεται η καταγραφή. Στην αρχή της πρώτης παραγράφου είναι καταγεγραμμένος ο τόπος που συνέβη το γεγονός και στο τέλος της ίδιας σειράς η ημερομηνία. Στην δεύτερη παράγραφο στο τέλος της πρώτης γραμμής υπάρχει το όνομα πατρός και λίγο πιο κάτω το όνομα μητρός. Το όνομα του ανθρώπου στον οποίο αναφέρεται η εγγραφή είναι στη μέση του δεξιού πλαισίου της εικόνας.

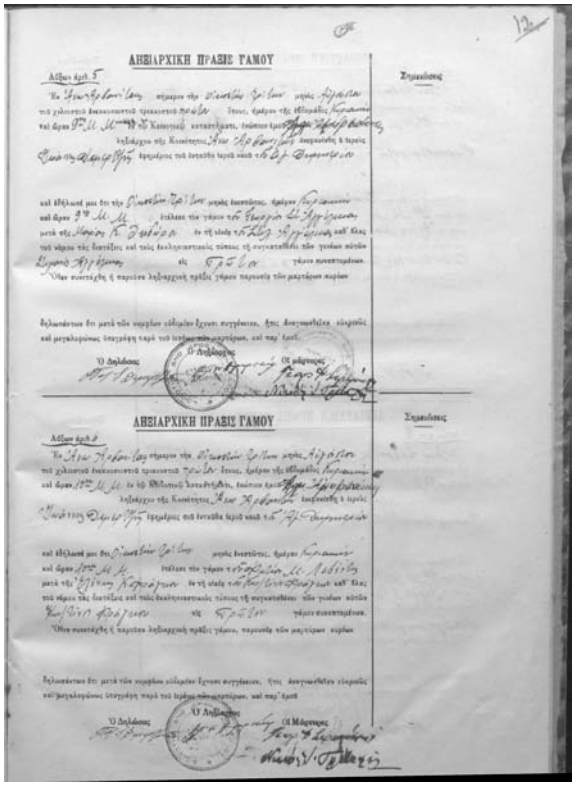
Αυτή είναι η εκ των προτέρων γνώση της θέσης των εικόνων πληροφορίας, που πρέπει το σύστημα να εξάγει.



(α)



(β)


ΛΗΞΙΑΡΧΙΚΗ ΠΡΑΞΙΣ ΓΑΜΟΥ

Σημείωσις

ΛΗΞΙΑΡΧΙΚΗ ΠΡΑΞΙΣ ΓΑΜΟΥ

Σημείωσις

(γ)

Σχήμα 27: Κατηγορίες πράξεων ληξιαρχικού αρχείου: (α) γέννησης (β) θανάτου (γ) γάμου

3.6.2 Διάκριση ληξιαρχικής πράξης

Στα ληξιαρχικά αρχεία Σάμου καταγράφονται πράξης γεννήσεως, θανάτου και γάμου. Για κάθε κατηγορία παρατηρείται διαφορετική διάταξη κειμένου αλλά και διαφορετική ζητούμενη πληροφορία. Άρα, προκύπτει η ανάγκη για διάκριση των ληξιαρχικών πράξεων από το σύστημα ώστε να είναι εφικτή η ξεχωριστή μετέπειτα αντιμετώπιση τους.

Για να γίνει αυτή η διάκριση το σύστημα εκμεταλλεύεται κάποιες διαφορές στη δομή αλλά και στο περιεχόμενο των εικόνων εγγράφου. Αναλυτικότερα, η εγγραφή γεννήσεως περιέχει κείμενο και στο δεξί πλαίσιο της ενώ οι άλλες δυο όχι. Έτσι γίνεται η διάκριση της πράξης γεννήσεως από τις άλλες δυο. Για την διάκριση των αρχείων γάμου και θανάτου εξετάζεται η επικεφαλίδα που βρίσκεται στο πάνω μέρος κάθε εγγράφου. Πιο συγκεκριμένα, εξετάζεται το μέγεθος της τελευταίας λέξης της

επικεφαλίδας στο οποίο αναφέρεται το είδος της πράξης.

(α)

(β)

Σχήμα 28: Επικεφαλίδες ληξιαρχικών αρχείων: (α) θανάτου (β)γάμου

3.6.3 Εντοπισμός εικόνας πληροφορίας

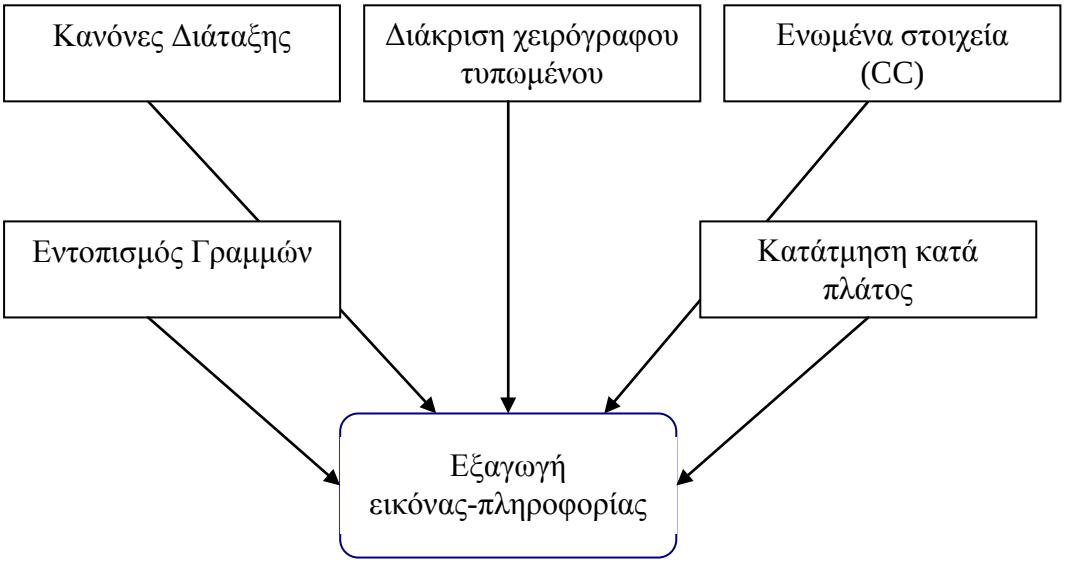
Για να μπορεί το σύστημα να εξάγει την πληροφορία που έχει ζητηθεί, πρέπει να ξέρει το σύνολο των pixel που την αποτελούν και την ακριβή θέση τους.
Το σύστημα γνωρίζει τα εξής :

- τα ενωμένα στοιχεία,
- τη γραμμή που ανήκουν τα CC ή αντίστροφα,
- αν κάθε ενωμένο στοιχείο είναι χειρόγραφο ή τυπωμένο,
- τη θέση των ενωμένων στοιχείων ως προς την εικόνα εγγράφου.

Αν συνδυαστεί η παραπάνω γνώση έχουμε ένα σύνολο εικονοστοιχείων για κάθε πληροφορία. Δηλαδή, εφόσον ξέρουμε ότι το όνομα πατρός βρίσκεται στην ένατη σειρά και είναι χειρόγραφο, μπορούμε να έχουμε το σύνολο των CC που ανήκουν στην ένατη σειρά εκτός από αυτά που έχουν θεωρηθεί τυπωμένα. Για να περιοριστεί το σύνολο αυτό θα πρέπει να εξαιρέσουμε και άλλα CC.

Το μέχρι τώρα σύστημα μπορεί να εντοπίσει κάποια πληροφορία περιορίζοντας τον αριθμό των CC ως προς την κατά ύψος θέση του στην εικόνα. Απαραίτητο λοιπόν είναι και ο περιορισμός τους ως προς το πλάτος της εικόνας για να γίνει πιο ακριβής ο εντοπισμός του. Η λύση που δόθηκε είναι η ισομερή κατά πλάτος τμηματοποίηση της εικόνας-εγγράφου και ο έλεγχος της ύπαρξης των CC μέσα στο κάθε τμήμα της.

Συμπερασματικά, μπορεί να δημιουργηθούν κανόνες που θα εντοπίζουν τις διάφορες εικόνες – πληροφορίας.



Διάγραμμα 4: Συνδυασμός πληροφορίας για τον εντοπισμό της ζητούμενης εικόνας

Και στις τρεις πράξεις οι εικόνες της ημερομηνίας και του τόπου εντοπίζονται στο ίδιο σημείο της εικόνας εγγράφου και οι κανόνες που χρησιμοποιούνται από το σύστημα είναι :

- Η εικόνα που περιγράφει την Ημερομηνία αποτελείται από τα CC που ανήκουν :
    - στην κατηγορία χειρόγραφων,
    - στην τέταρτη γραμμή,
    - στο δεξί μέρος της σελίδας εγγράφου.
- Η εικόνα που περιγράφει τον Τόπο αποτελείται από τα CC που ανήκουν :
    - στην κατηγορία χειρόγραφων,
    - στην τέταρτη γραμμή,
    - στο αριστερό μέρος της σελίδας εγγράφου.

Δηλαδή όσον αφορά τις εικόνες – εγγράφων που περιέχουν εγγραφή γέννησης, οι κανόνες είναι οι παρακάτω :

- Η εικόνα που περιγράφει το Όνομα Πατρός αποτελείται από τα CC που ανήκουν :

    - στην κατηγορία χειρόγραφου .
    - στην ένατη γραμμή,
    - στο δεξιό μέρος της σελίδας.

- Αντίστοιχα η εικόνα που περιγράφει το Όνομα Μητρός αποτελείται από τα CC που ανήκουν :

    - στη κατηγορία χειρόγραφου,
    - στην ενδέκατη γραμμή,
    - στο αριστερό μέρος της σελίδας.

- Τέλος η εικόνα που περιγράφει το Όνομα αποτελείται από τα CC που ανήκουν :

    - στην κατηγορία χειρόγραφου,
    - στην δέκατη γραμμή του δεξιού πλαισίου.

Στο δεξιό πλαίσιο είναι περιττό να γίνει κατά πλάτος τμηματοποίηση αφού σε κάθε γραμμή περιέχονται λίγες λέξεις τόσες ώστε η διάκριση τυπωμένου – χειρόγραφου αρκεί για την εύρεση θέσης ως προς το πλάτος της εικόνας.

Για τις εικόνες – εγγράφων που περιέχουν εγγραφή θανάτου, οι κανόνες είναι οι παρακάτω :

- Η εικόνα που περιγράφει το Όνομα Πατρός ή Συζύγου (δεν διευκρινίζεται) αποτελείται από τα CC που ανήκουν :

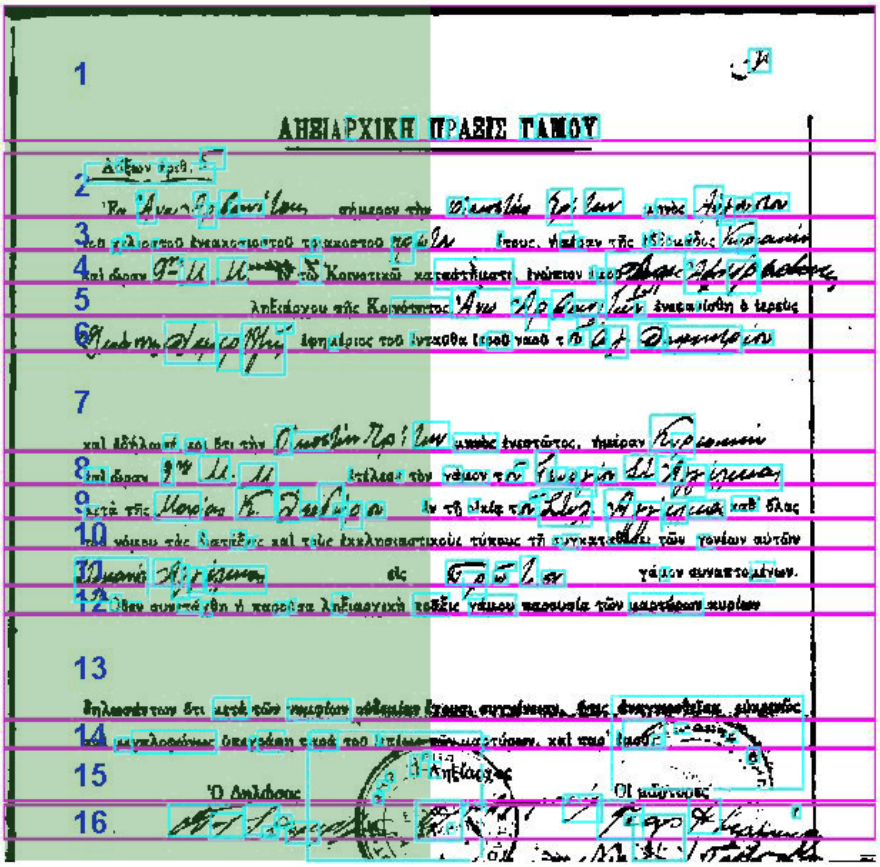
    - στην κατηγορία χειρόγραφου,
    - στην ένατη γραμμή,
    - στο δεξιό μέρος της σελίδας.

- Τέλος η εικόνα που περιγράφει το Όνομα αποτελείται από τα CC που ανήκουν :

- στην κατηγορία χειρόγραφου .
- στην δέκατη και ενδέκατη γραμμή,
- στο δεξιό μέρος (δέκατη γραμμή) και στο αριστερό μέρος (ενδέκατη γραμμή) της σελίδας.

Για τις εικόνες – εγγράφων που περιέχουν εγγραφή γάμου, οι κανόνες είναι οι παρακάτω :

- Η εικόνα που περιγράφει το Όνομα του Συζύγου αποτελείται από τα CC που ανήκουν :
  - στην κατηγορία χειρόγραφου .
  - στην δέκατη γραμμή,
  - στο δεξιό μέρος της σελίδας.
- Η εικόνα που περιγράφει το Όνομα της Συζύγου αποτελείται από τα CC που ανήκουν :
  - στην κατηγορία χειρόγραφου .
  - στην ενδέκατη γραμμή,
  - στο αριστερό μέρος της σελίδας.



Σχήμα 29: Απεικόνιση συνδυασμού πληροφοριών: η ροζ διαγράμμιση δείχνει τις γραμμές, τα γαλάζια πλαίσια τα χειρόγραφα ενωμένα στοιχεία και το γαλάζιο ορθογώνιο δηλώνει τη δεξιά πλευρά τις εκόνας.

Κεφάλαιο 4

Αποτελέσματα

Στο κεφάλαιο αυτό θα παρουσιαστούν τα πειράματα που έγιναν με σκοπό την αξιολόγηση του συστήματος επεξεργασίας εικόνων ληξιαρχικού αρχείου και εξαγωγής πληροφορίας. Πραγματοποιήθηκαν πειράματα με εικόνες εγγράφων από διάφορα αρχεία.

Για την αξιολόγηση ολόκληρου του συστήματος έγιναν πειράματα στα ληξιαρχικά αρχεία των Άνω Αρβανιτών ενώ για την αξιολόγηση του συστήματος χωρίς το στάδιο της εξαγωγής εικόνων πληροφορίας χρησιμοποιήθηκαν τα αρχεία της Κεφαλλονιάς και των Άνω Αρβανιτών.

4.1 Πειραματικά δεδομένα

Για την αξιολόγηση του συστήματος χρησιμοποιήθηκαν 90 εικόνες εγγράφων του ληξιαρχικού αρχείου Σάμου και 30 από αυτό της Κεφαλλονιάς.

4.2 Μέτρο αξιολόγησης

Ως μέτρο αξιολόγησης των πειραμάτων για ολόκληρο το σύστημα θεωρήθηκε το

ποσοστό ακρίβειας που δείχνει ουσιαστικά το ποσοστό της πληροφορίας που εξάχθηκε σωστά.

Για την αξιολόγηση του συστήματος χωρίς το στάδιο εξαγωγής εικόνων πληροφορίας σαν μέτρα επιλέχθηκαν το ποσοστό ακρίβειας της διάκρισης τυπωμένου-χειρόγραφου και το ποσοστό ακρίβειας εντοπισμού γραμμών.

4.3 Αξιολόγηση προτεινόμενου συστήματος

Όπως έχει ήδη αναφερθεί η αξιολόγηση του συστήματος έγινε στο ληξιαρχικό αρχείο Σάμου. Θα παρουσιαστεί η ακρίβεια της εξαγωγής συγκεκριμένης πληροφορίας. Δηλαδή κάθε ζητούμενη εικόνα πληροφορίας θεωρείται σωστή αν περιέχει το κείμενο της πληροφορίας πλήρως χωρίς απώλεια πληροφορίας.

Πρέπει να υπογραμμιστεί ότι μετρώντας την ακρίβεια του τελευταίου σταδίου ενός συστήματος επεξεργασίας εικόνας μετριέται ουσιαστικά η επιτυχία όλων των σταδίων του. Για παράδειγμα ένα κακό binarization θα επηρεάσει αρνητικά επόμενα στάδια του συστήματος όπως ο εντοπισμός CC, η εξαγωγή χαρακτηριστικών κ.ο.κ. και επομένως αυτό θα καταγραφεί στην ακρίβεια του τελευταίου σταδίου.

Στον επόμενο Πίνακα φαίνονται τα αποτελέσματα για την ληξιαρχική πράξη της γέννησης και τις ζητούμενες εικόνες-πληροφορίας: Ημερομηνία, Τόπο, Όνομα Πατρός, Όνομα Μητρός, Όνομα.

Τόπος	Ημερομηνία	Πατέρας	Μητέρα	Όνομα
63,6	66,6	33,3	33,3	53,3

Πίνακας 2: Αποτελέσματα ληξιαρχικής πράξη της γέννησης

Όπως παρατηρεί κανείς τα ποσοστά ακρίβειας είναι χαμηλότερα στο όνομα πατέρα, μητέρας και ενδιαφερόμενου. Αυτό οφείλεται στο γεγονός κυρίως του λάθους εντοπισμού των περιοχών ως προ το ύψος. Επίσης παρατηρήθηκε κάποια απώλεια πληροφορίας λόγω απόρριψης κάποιων περιοχών εικόνας, οι οποίες θεωρήθηκαν εσφαλμένα τυπωμένες.

(α)

(β)

Σχήμα 30: Παράδειγμα εικόνας πληροφορίας, (α) δεκτή ως σωστή, (β) μη αποδεκτή

Στον παρακάτω Πίνακα φαίνονται τα αποτελέσματα για την ληξιαρχική πράξη του θανάτου και τις ζητούμενες εικόνες-πληροφορίας: Ημερομηνία, Τόπο, Όνομα Πατρός ή Συζύγου, Όνομα.

Τόπος	Ημερομηνία	Πατέρας/Σύζυγος	Όνομα
63,3	66,6	33,3	26,6

Πίνακας 3: Αποτελέσματα ληξιαρχικής πράξης θανάτου

Όπως και στην περίπτωση της πράξης του γάμου, έτσι και εδώ παρατηρήθηκαν μη ικανοποιητικά ποσοστά. Συγκεκριμένα Στο στοιχείο του ονόματος υπήρξε η μεγαλύτερη αστοχία και αυτό οφείλεται στην πολυπλοκότητα της θέσης του (βρίσκεται σε δύο γραμμές του κειμένου)

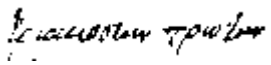
Στον Πίνακα φαίνονται τα αποτελέσματα για την ληξιαρχική πράξη του γάμου και τις ζητούμενες εικόνες-πληροφορίας: Ημερομηνία, Τόπο, Όνομα του Συζύγου, Όνομα της Συζύγου.

Τόπος	Ημερομηνία	Όνομα του Συζύγου	Όνομα της Συζύγου
70	63,3	36,6	33,3

Πίνακας 4: Αποτελέσματα ληξιαρχικής πράξης γάμου

Τέλος τα αποτελέσματα για την πράξη γάμου κυμαίνονται στα ίδια επίπεδα με τα αντίστοιχα των πράξεων θανάτου και γέννησης.

Ένας επιπλέον παράγοντας που επηρέασε την ακρίβεια του συστήματος είναι το λάθος Binarization σε πολλές περιπτώσεις των εικόνων εγγράφων. Αυτό οφείλεται στην πολύ χαμηλή ποιότητα εικόνας αλλά και στην ποικιλία φωτισμού.



Σχήμα 31: Παράδειγμα κακού Binarization

4.5 Αξιολόγηση διάκρισης τυπωμένου χειρόγραφου

Η αξιολόγηση του συστήματος μέχρι το στάδιο της διάκρισης τυπωμένου χειρόγραφου, έγινε σε δύο ληξιαρχικά αρχεία, της Σάμου και της Κεφαλλονιάς. Στο παρόν κεφάλαιο θα παρουσιαστεί η ακρίβεια της διάκρισης των ενωμένων στοιχεία σε τυπωμένο ή χειρόγραφο. Δηλαδή κάθε ενωμένο στοιχείο θεωρείται σωστά ομαδοποιημένο, αν στο κείμενο ή στο γράμμα που περιέχεται σε αυτό υπάρχει αντιστοίχηση της ομάδας με τον τρόπο γραφής.

Πιο συγκεκριμένα για την ακρίβεια του χειρόγραφου μετριέται το ποσοστό σωστά χαρακτηρισμένων ενωμένων στοιχείων ως προ το σύνολο των χειρόγραφων ενωμένων στοιχείων. Η ακρίβεια του τυπωμένου μετριέται αντίστοιχα.

	Τυπωμένο	Χειρόγραφο
Σάμου	85,6	94,3
Κεφαλλονιάς	88,7	98,2

Πίνακας 5: Ακρίβεια διάκρισης τυπωμένου- χειρόγραφου για τα δύο αρχεία

Από τον Πίνακα παρατηρείται καλή ακρίβεια όσον αφορά τη διάκριση τυπωμένου χειρόγραφου. Υψηλότερη ακρίβεια υπάρχει στη διάκριση χειρόγραφου, ενώ φαίνεται ότι στα αρχεία της Κεφαλλονιάς η ακρίβεια και του τυπωμένου και του χειρόγραφου είναι η μεγαλύτερη

Το γεγονός αυτό μπορεί να εξηγηθεί, εφόσον τα αρχεία της Σάμου σε σχέση με αυτά της Κεφαλλονιάς είναι πολύ κατώτερης ποιότητας. Τα χαρακτηριστικά που εξάγονται από τα αρχεία της Κεφαλλονιάς είναι πιο πλήρη και πιο αντιπροσωπευτικά από τα άλλα. Συνεπώς, ο ομαδοποιητής πετυχαίνει σε αυτά υψηλότερα σκορ. Επίσης, η

εύρεση ενωμένων στοιχείων είναι πιο σωστή στης Κεφαλλονιάς τα αρχεία. Χαρακτηριστικό παράδειγμα είναι τα ενωμένα στοιχεία που εξάγονται από το τυπωμένο κείμενο, στα μεν κάθε ενωμένο στοιχείο αποτελείται από ένα χαρακτήρα, ενώ στα άλλα όχι.

Κεφάλαιο 5

Συμπεράσματα

Στα προηγούμενα κεφάλαια παρουσιάστηκε ένα σύστημα επεξεργασίας εικόνων ληξιαρχικού αρχείου και εισαγωγής πληροφορίας. Το σύστημα αυτό μπορεί να αποτελέσει προεργασία για πολλές εφαρμογές. Η πιο συνήθης είναι η οπτική αναγνώριση προτύπων αλλά και εφαρμογές ανάκτησης εικόνων εγγράφων και δημιουργίας ψηφιακών εικόνων (Content Based Image Retrieval).

Το προτεινόμενο σύστημα αποτελείται από το στάδιο της κατάτμησης, του binarization, της διόρθωσης γωνίας εκτροπής, την διάκριση τυπωμένου-χειρόγραφου, τον ορισμό γραμμών, και εντοπισμό συγκεκριμένης εικόνας πληροφορίας.

Το στάδιο της κατάτμησης είναι απαραίτητο μόνο για το αρχείο της Σάμου, όπως και οι κανόνες για τον εντοπισμό συγκεκριμένης πληροφορίας. Το σύστημα θα μπορούσε να υιοθετηθεί για την επεξεργασία οποιουδήποτε μεικτού εγγράφου αρκεί να προσαρμοστούν σε αυτό η διαδικασία κατάτμησης της εικόνας (αν είναι απαραίτητη) και οι κανόνες που ορίζουν την εκ των προτέρων γνώση για την διάταξη του εγγράφου.

Τονίστηκε η αναγκαιότητα για ακρίβεια σε όλα τα βήματα του συστήματος, αφού κάθε

αστοχία που παρατηρείται σε ένα στάδιο μεταδίδεται και στο επόμενο.

Επισημάνθηκε ότι για την διαδικασία του binarization πρέπει να δίνεται ιδιαίτερη προσοχή στην επιλογή των μεταβλητών του αλγορίθμου, για κάθε είδος ληξιαρχικού εγγράφου εικόνας.

Ως μέθοδος εντοπισμού κειμένου-εικόνας χρησιμοποιήθηκε η διάκριση χειρόγραφου – τυπωμένου. Επίσης, χρησιμοποιήθηκε αποτελεσματικά στον ορισμό των γραμμών, αφού ξεπερνάει τα κλασικά προβλήματα που αντιμετωπίζει κανείς στο χειρόγραφο κείμενο.

Για πρώτη φορά υιοθετήθηκε ο αλγόριθμος ομαδοποίησης k-medoids για τη διάκριση τυπωμένου – χειρόγραφου, με ικανοποιητικά αποτελέσματα.

Έγινε σαφής η ανάγκη για ποιοτική ψηφιοποίηση των εικόνων – εγγράφων. Δηλαδή η ψηφιοποίηση σε υψηλή ανάλυση, ποιότητα και η αποθήκευση του με κατάλληλο τύπο αρχείου. Οι εικόνες εγγράφων της Σάμου δεν είχαν αρκετή πληροφορία και επηρέασαν την ακρίβεια του συστήματος ενώ στα αρχεία της Κεφαλλονιάς, σημείωσαν υψηλότερη ακρίβεια σε αντίστοιχα στάδια.

Στο στάδιο διάκρισης χειρόγραφου-τυπωμένου στις περιπτώσεις σφάλματος χαρακτηρίζονται συχνότερα χειρόγραφες περιοχές που δεν είναι, παρά μη χειρόγραφες περιοχές που είναι. Το γεγονός αυτό δεν επηρεάζει το τελικό αποτέλεσμα του συστήματος, αφού με αυτό το είδος σφάλματος δεν έχουμε απώλεια πληροφορίας.

Κεφάλαιο 6

Μελλοντική εργασία

Το προτεινόμενο σύστημα παρουσιάζει ικανοποιητικά αποτελέσματα αλλά θα μπορούσε να γίνει σε ορισμένες περιπτώσεις κάποια βελτίωση. Στον εντοπισμό συγκεκριμένων περιοχών συνήθως παρουσιάζεται σφάλμα ως προς το ύψος της θέσης τους. Θα μπορούσε να βελτιωθεί το στάδιο του ορισμού γραμμών, εξαλείφοντας αποτελεσματικότερα κάποιο είδος θορύβου. Επίσης θα μπορούσε να βρεθεί ένας εναλλακτικός τρόπος εύρεσης θέσης.

Βελτίωση στο σύστημα θα σημειωνόταν αν σχεδιαζόταν εφαρμογή διεπαφής ώστε ο χρήστης να μπορεί να εντοπίζει και να διορθώνει λάθη που προκύπτουν από το σύστημα. Ακόμη η διεπαφή αυτή θα μπορούσε να δημιουργηθεί για να μπορεί ο ίδιος ο χρήστης εύκολα να δίνει στο σύστημα την διάταξη του μεικτού εγγράφου, ώστε να αποφεύγεται η εκ νέου δημιουργία κανόνων για διαφορετικά είδη μεικτών εγγράφων.

Κάποια άλλη βελτίωση θα μπορούσε να προέλθει από μια διαφορετική προσέγγιση που αφορά τις περιοχές (CC) από όπου γίνεται η εξαγωγή χαρακτηριστικών. Δηλαδή να αντικατασταθούν τα ενωμένα στοιχεία με μια διαδικασία κατάτμησης λέξεων, και να εξάγονται χαρακτηριστικά από αυτές. Σε αυτή τη περίπτωση θα προστίθονταν στα ήδη υπάρχοντα χαρακτηριστικά που διαφοροποιούν χειρόγραφες από τυπωμένες λέξεις.

.

Κεφάλαιο 7

Βιβλιογραφία

[1] A. Jain and B. Yu, Document representation and its application to page decomposition, IEEE ¹rans. Pat-tern. Anal. Machine Intell. 20, 294Ð308 (1998)

[2] C. Strouthpoulos, N. Papamarkos, A.E. Atsalakis, Text extraction in complex color document, Pattern Recognition 35 (8) (2002) 1743–1758

[3] Y. Hamamoto *et al.*, “Recognition of handprinted Chinese characters using Gabor features,” in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, 1995, pp. 819–823.

[4] Otsu, N. “A threshold selection method from gray-level histograms”. IEEE Trans. Systems Man Cybernet. pp. 62-66, 9 (1), 1979.

[5] Niblack, W. “An Introduction to Digital image processing", pp 115-116, Prentice Hall, 1986.

[6] E. Kavallieratou and E. Stamatatos, Improving the quality of degraded document images, in Proc. Int’l Conf. Document Image Analysis for Libraries (DIAL), (Lyon, France), 2006.

[7] S. Vavilis, E. Kavallieratou .A tool for Tuning Binarization Techniques, ICDAR 2011.

[8] W. Postl, Detection of linear oblique structures and skew scan in digitized ocuments, Proceedings of the Eighth International Conference on Pattern Recognition, IEEE CS Press, Los Alamitos, CA,1986, pp. 687–689.

[9] T. Pavlidis, J. Zhou, Page segmentation by white streams, Proc. First Int. Conf. Doc. Anal. Recogn. (ICDAR), Int. Assoc. Pattern Recogn. (1991) 945–953

[10] D.S. Le, G.R. Thoma, H. Wechsler, Automated page orientation and skew angle

detection for binary document image, Pattern Recogn. 27 (10) (1994) 1325–1344.

[11] E. Kavallieratou, N. Fakotakis, and G. Kokkinakis. Skew angle estimation for printed and handwritten documents using the wigner-ville distribution. Image and Vision Computing, 20:813–824, 2002.

[12] K. Kuhnke, L. Simoncini, and Z. Kovacs-V, “A system for machinewritten and hand-written character distinction,” 3rd Intl. Conf on Document Analysis and Recognition (Vol. 2), pp. 811–814, 1995.

[13] S.Pinson , W.Barrett "Connected Component Level Discrimination of Handwritten and Machine-Printed Text Using Eigenfaces", 11th Intl. Conf on Document Analysis and Recognition , pp. 1394-1398, 2011.

[14] M.S. Shirdhonkar and Manesh B. Kokare, “Discrimination between Printed and Handwritten Text in Documents,” in IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, pp. 131-134, 2010.

[15] S. Chanda, K. Franke and U. Pal, “Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments,” in SAC’10, pp. 18-122, March 22-26, 2010.

[16] Downs, G. M. and Barnard, J. M. (2002) "Clustering methods and their uses in computational chemistry", in: K. B. Lipkowitz and D. B. Boyd (Eds) Reviews in Computational Chemistry, Vol. 18, pp. 1 –40

[17] A. Lemaitre, J. Camillerapp, “Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image”, Second International Conference on Document Image Analysis for Libraries, 38-45,2006

[18] E. Bruzzone, M.C. Coffetti, “An algorithm for extracting cursive text lines”, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 749-752,1999.

[19] Y. Li, Y. Zheng, D. Doermann, “Detecting Text Lines in Handwritten Documents”, 18th International Conference on Pattern Recognition, 2, 1030-1033, 2006.

[20] G. Louloudis, B. Gatos and C. Halatsis, “Text Line Detection in Unconstrained Handwritten Documents Using a Block-Based Hough Transform Approach”, 9th International Conference on Document Analysis and Recognition (ICDAR'07), pp. 599-603, Curitiba, Brazil, September 2007.

Κατάλογος Πινάκων

Κατάλογος Σχημάτων και Διαγραμμάτων