Recognition of Greek Polytonic on Historical Degraded Texts using HMMs

Vassilis Katsouros[1], Vassilis Papavassiliou[1], Fotini Simistira[3,1], and Basilis Gatos[2]
[1]Institute for Language and Speech Processing (ILSP)
Athena Research and Innovation Center, Athens, Greece
Email: {vpapa, fotini, vsk}@ilsp.gr

[2]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos", Athens, Greece
bgat@iit.demokritos.gr

[3]DIVA Research Group
University of Fribourg, Switzerland
Email: foteini.simistira@unifr.ch

***Abstract*— Optical Character Recognition (OCR) of ancient Greek polytonic scripts is a challenging task due to the large number of character classes, resulting from variations of diacritical marks on the vowel letters. Classical OCR systems require a character segmentation phase, which in the case of Greek polytonic scripts is the main source of errors that finally affects the overall OCR performance. This paper suggests a character segmentation free HMM-based recognition system and compares its performance with other commercial, open source, and state-of-the art OCR systems. The evaluation has been carried out on a challenging novel dataset of Greek polytonic degraded texts and has shown that HMM-based OCR yields character and word level error rates of 8.61% and 25.30% respectively, which outperforms most of the available OCR systems and it is comparable with the performance of the state-of- the-art system based on LSTM Networks proposed recently.**

Keywords— Hidden Markov Models; Optical Character Recognition; Greek polytonic

I. Introduction

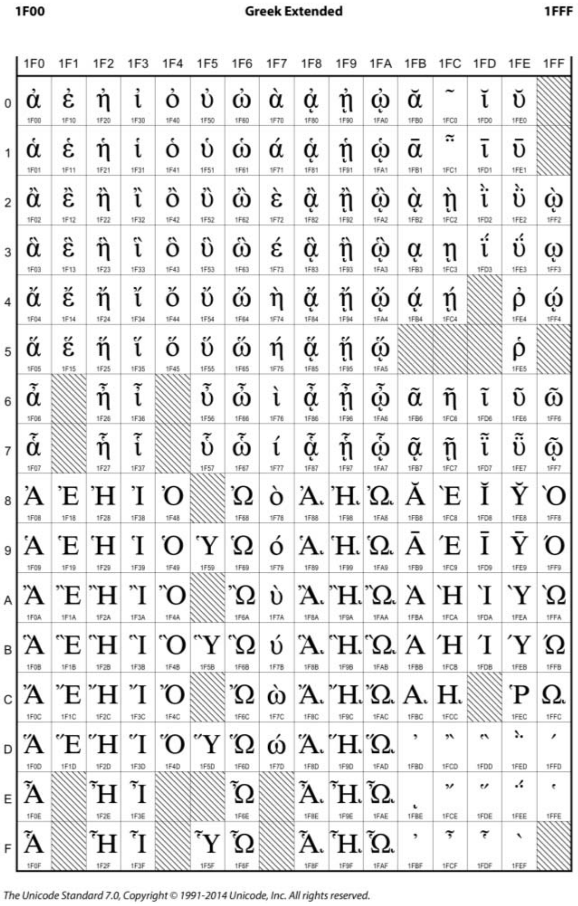
Most classical OCR systems developed for text recognition of scanned document images employ a character segmentation-recognition method. Therefore, the performance of such systems is significantly dependent on successful character segmentation hypothesis. However, in cases where there are several diacritical marks above or below character letters as well as in the presence of noise resulting from the scanning of the document and the binarization stages standard character segmentation approaches introduce more symbols either by over-segmenting characters or by considering connected components of noise as possible characters. This is a typical situation one may observe in historical document images of Greek polytonic scripts.

The Greek polytonic system was used from around 200 BC up to modern times until 1982. Greek polytonic uses many different diacritics in several categories, with each of these denoted a significant variation in the pronunciation. In particular, the first category involves three different accents placed above the vowels, the acute accent, e.g. ά, to mark high pitch on a short vowel, the grave accent, e.g. ὰ, to mark normal or low pitch, and the circumflex, e.g. ᾶ, to mark high and falling pitch within one syllable; the second category involves two breathings placed also above vowels, the rough breathing, e.g. ἁ, to indicate a voiceless glottal fricative (/h/) before a vowel and the smooth breathing, e.g. ἀ, to indicate the absence of /h/; the diaeresis appears on the letters ι and υ, e.g. ϊ and ϋ, to show that a pair of vowel letters is pronounced separately, rather than as a diphthong; and the iota subscript placed under the long vowels α, η, and ω, e.g. ᾳ, to indicate the ancient long diphthongs αι, ηι, and ωι, in which the ι is no longer pronounced.

In Modern Greek the pitch accent was replaced by a dynamic accent, and the /h/ was lost, so most of the diacritics of polytonic Greek have no phonetic significance, but merely reveal the underlying Ancient Greek etymology. Monotonic orthography, the standard system for Modern Greek, retains a single accent (´) to indicate stress, the diaeresis (¨) to indicate a diphthong, sometimes used in combination. The total number of Greek letters is 49, 25 lower case and 24 upper case, among of which 14 are vowels. When vowels in Greek polytonic are combined with the diacritic marks mentioned above they result in more than 270 character classes as shown in Fig 1.

One way of tackling the problem of diacritics variations of characters is to add a module in the character segmentation phase that identifies diacritic marks, reduces the number of classes and at a later stage combines the recognition results with contextual knowledge of the polytonic Greek orthography [3]. However, this method is based on sophisticated character segmentation taking into account the diacritic marks but still performs with a little success especially when degraded document images are considered, which is mostly the case in digital archives. Therefore, digitization of Greek polytonic document archives is still a challenging problem that leaves space for much improvement.

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.60



1F00 Greek Extended 1FFF

	1F0	1F1	1F2	1F3	1F4	1F5	1F6	1F7	1F8	1F9	1FA	1FB	1FC	1FD	1FE	1FF
0	ἀ 1F00	ἐ 1F10	ἠ 1F20	ἰ 1F30	ὀ 1F40	ὐ 1F50	ὠ 1F60	ὰ 1F70	ᾀ 1F80	ᾐ 1F90	ᾠ 1FA0	ᾰ 1FB0	῀ 1FC0	ῐ 1FD0	ῠ 1FE0	
1	ἁ 1F01	ἑ 1F11	ἡ 1F21	ἱ 1F31	ὁ 1F41	ὑ 1F51	ὡ 1F61	ά 1F71	ᾁ 1F81	ᾑ 1F91	ᾡ 1FA1	ᾱ 1FB1	῁ 1FC1	ῑ 1FD1	ῡ 1FE1	
2	ἂ 1F02	ἒ 1F12	ἢ 1F22	ἲ 1F32	ὂ 1F42	ὒ 1F52	ὢ 1F62	ὲ 1F72	ᾂ 1F82	ᾒ 1F92	ᾢ 1FA2	ᾲ 1FB2	ῂ 1FC2	ῒ 1FD2	ῢ 1FE2	ῲ 1FF2
3	ἃ 1F03	ἓ 1F13	ἣ 1F23	ἳ 1F33	ὃ 1F43	ὓ 1F53	ὣ 1F63	έ 1F73	ᾃ 1F83	ᾓ 1F93	ᾣ 1FA3	ᾳ 1FB3	ῃ 1FC3	ΐ 1FD3	ΰ 1FE3	ῳ 1FF3
4	ἄ 1F04	ἔ 1F14	ἤ 1F24	ἴ 1F34	ὄ 1F44	ὔ 1F54	ὤ 1F64	ὴ 1F74	ᾄ 1F84	ᾔ 1F94	ᾤ 1FA4	ᾴ 1FB4	ῄ 1FC4		ῤ 1FE4	ῴ 1FF4
5	ἅ 1F05	ἕ 1F15	ἥ 1F25	ἵ 1F35	ὅ 1F45	ὕ 1F55	ὥ 1F65	ή 1F75	ᾅ 1F85	ᾕ 1F95	ᾥ 1FA5				ῥ 1FE5	
6	ἆ 1F06		ἦ 1F26	ἶ 1F36		ὖ 1F56	ὦ 1F66	ὶ 1F76	ᾆ 1F86	ᾖ 1F96	ᾦ 1FA6	ᾶ 1FB6	ῆ 1FC6	ῖ 1FD6	ῦ 1FE6	ῶ 1FF6
7	ἇ 1F07		ἧ 1F27	ἷ 1F37		ὗ 1F57	ὧ 1F67	ί 1F77	ᾇ 1F87	ᾗ 1F97	ᾧ 1FA7	ᾷ 1FB7	ῇ 1FC7	ῗ 1FD7	ῧ 1FE7	ῷ 1FF7
8	Ἀ 1F08	Ἐ 1F18	Ἠ 1F28	Ἰ 1F38	Ὀ 1F48		Ὠ 1F68	ὸ 1F78	ᾈ 1F88	ᾘ 1F98	ᾨ 1FA8	Ᾰ 1FB8	Ὲ 1FC8	Ῐ 1FD8	Ῠ 1FE8	Ὸ 1FF8
9	Ἁ 1F09	Ἑ 1F19	Ἡ 1F29	Ἱ 1F39	Ὁ 1F49	Ὑ 1F59	Ὡ 1F69	ό 1F79	ᾉ 1F89	ᾙ 1F99	ᾩ 1FA9	Ᾱ 1FB9	Έ 1FC9	Ῑ 1FD9	Ῡ 1FE9	Ό 1FF9
A	Ἂ 1F0A	Ἒ 1F1A	Ἢ 1F2A	Ἲ 1F3A	Ὂ 1F4A		Ὢ 1F6A	ὺ 1F7A	ᾊ 1F8A	ᾚ 1F9A	ᾪ 1FAA	Ὰ 1FBA	Ὴ 1FCA	Ὶ 1FDA	Ὺ 1FEA	Ὼ 1FFA
B	Ἃ 1F0B	Ἓ 1F1B	Ἣ 1F2B	Ἳ 1F3B	Ὃ 1F4B	Ὓ 1F5B	Ὣ 1F6B	ύ 1F7B	ᾋ 1F8B	ᾛ 1F9B	ᾫ 1FAB	Ά 1FBB	Ή 1FCB	Ί 1FDB	Ύ 1FEB	Ώ 1FFB
C	Ἄ 1F0C	Ἔ 1F1C	Ἤ 1F2C	Ἴ 1F3C	Ὄ 1F4C		Ὤ 1F6C	ὼ 1F7C	ᾌ 1F8C	ᾜ 1F9C	ᾬ 1FAC	ᾼ 1FBC	ῌ 1FCC		Ῥ 1FEC	ῼ 1FFC
D	Ἅ 1F0D	Ἕ 1F1D	Ἥ 1F2D	Ἵ 1F3D	Ὅ 1F4D	Ὕ 1F5D	Ὥ 1F6D	ώ 1F7D	ᾍ 1F8D	ᾝ 1F9D	ᾭ 1FAD	᾽ 1FBD	῍ 1FCD	῝ 1FDD	῭ 1FED	´ 1FFD
E	Ἆ 1F0E		Ἦ 1F2E	Ἶ 1F3E			Ὦ 1F6E		ᾎ 1F8E	ᾞ 1F9E	ᾮ 1FAE	ι 1FBE	῎ 1FCE	῞ 1FDE	΅ 1FEE	῾ 1FFE
F	Ἇ 1F0F		Ἧ 1F2F	Ἷ 1F3F		Ὗ 1F5F	Ὧ 1F6F		ᾏ 1F8F	ᾟ 1F9F	ᾯ 1FAF	᾿ 1FBF	῏ 1FCF	῟ 1FDF	` 1FEF	

The Unicode Standard 7.0, Copyright © 1991-2014 Unicode, Inc. All rights reserved.

Fig. 1. Extended Greek characters: Diacritic marks modifying vowel characters (retrieved from http://www.unicode.org/charts/PDF/U1F00.pdf)

In this paper, we suggest and evaluate segmentation free HMM-based techniques for the recognition of text-lines of printed Greek polytonic scripts. We compare the performance of the HMM recognizers with current open-source and commercial OCR engines. We perform evaluation experiments on challenging datasets obtained from a novel database for Greek polytonic scripts. Evaluation results show a promising performance of the proposed HMM-based method using language models of bigrams at character level.

This research work is part of the OldDocPro project[1] which aims towards the recognition of Greek polytonic machine-printed and handwritten documents. In OldDocPro, we strive toward research and technology to advance the frontiers and facilitate current and future efforts of digitization of old Greek document archives of old Greek documents turning them into a digital collection with full-text access capabilities using novel OCR methods.

The remainder of the paper is organized as follows. In Section II, the description of the datasets of Greek polytonic and the HMM-based recognition systems are presented. In Section III, we present the evaluation experiments and discuss the performance of various configurations. Finally, conclusions and future work are discussed in Section IV.

TABLE I. OVERVIEW OF GREEK POLYTONIC DATASETS

Politician/Year	Pages	Number of Text Lines	Number of Words	Number of Characters
Greek Official Government Gazette				
FEK/1959	5	691	4,998	28,591
Greek Parliament Proceedings				
Saripolos/1864	6	642	5,797	30,533
Venizelos/1931	5	522	4,484	22,923
Markezinis/1953	18	1,665	13,033	72,750
Vlahou/ 1977	4	373	3,343	16,714
Total	**38**	**3,893**	**31,655**	**171,511**

II. DESCRIPTION OF RECOGNITION SYSTEM

A. Datasets

The datasets of printed polytonic Greek scripts used in the experiment of this work is part of the GRPOLY-DB database [4] and involves scanned document images of the Greek Parliament Proceedings corresponding to speeches of four Greek politicians (Vlahou in 1977, Markezinis in 1953, Saripolos in 1864 and Venizelos in 1931) and a few pages from the Greek Official Government Gazette (1959) covering different time periods. Quantitative details on these datasets are given in Table I.

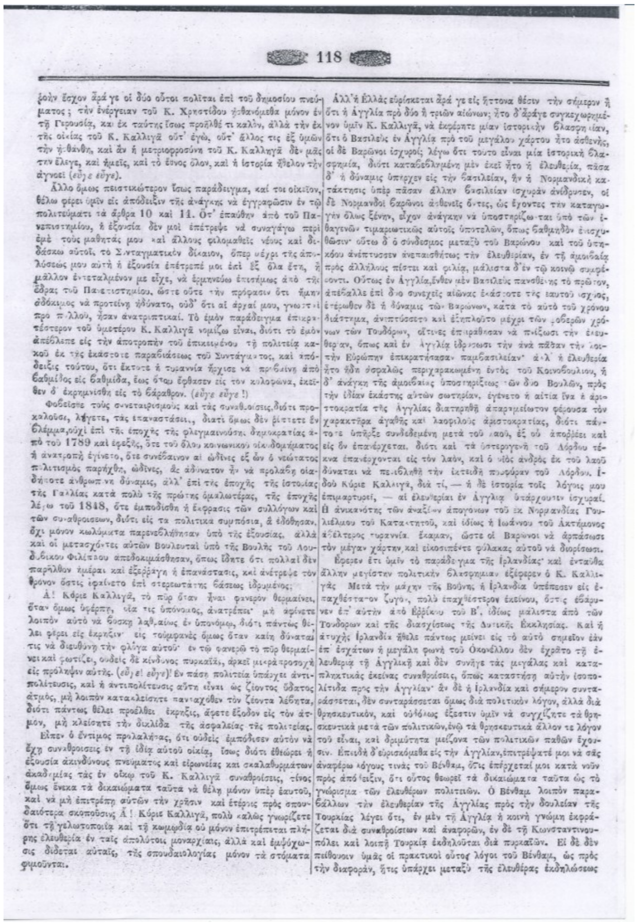
For the creation of the datasets, we used the original grayscale images of all pages together with the corresponding ground truth texts. We first binarized [5] the grayscale images (Fig. 2) and then applied layout analysis and segmentation processes [6] to extract the respective text-lines (Fig. 3a & b) and words. In order to assign the text information to the corresponding text-lines an automatic transcript mapping procedure was applied [7]. Finally, the text-line and word segmentation results as well as the respective transcripts' alignment (Fig. 3c) were verified and corrected manually using the Aletheia framework [8].

The number of unique character classes contained in the above mentioned datasets is 189, including Greek characters, extended Greek characters, numbers, special characters, hyphenation marks, etc.

B. Feature Extraction

To extract a sequence of features from a text line image we use a sliding window, which it is moved along the text line following the writing direction from left to right. The windows are called frames and from each frame we calculate the respective feature vector. In our approach we have adopted two types of feature sets; the first is a variation of the geometric features firstly suggested in [9] and the second set has been recently suggested and widely tested in handwriting

[1] https://www.iit.demokritos.gr/project/olddocpro



118

ξοὴν ἔσχον ἆρά γε οἱ δύο οὗτοι πολῖται ἐπὶ τοῦ δημοσίου πνεύματος; τὴν ἐνέργειαν τοῦ Κ. Χρηστίδου ᾐσθανόμεθα μόνον ἐν τῇ Γερουσίᾳ, καὶ ἐκ ταύτης ἴσως προῆλθέ τι καλὸν, ἀλλὰ τὴν ἐκ τῆς οἰκίας τοῦ Κ. Καλλιγᾶ οὔτ' ἐγὼ, οὔτ' ἄλλος τις ἐξ ὑμῶν τὴν ᾐσθάνθη, καὶ ἂν ἡ μετριοφροσύνη τοῦ Κ. Καλληγᾶ δὲν μᾶς την ἔλεγε, καὶ ἡμεῖς, καὶ τὸ ἔθνος ὅλον, καὶ ἡ ἱστορία ἤθελον τὴν ἀγνοεῖ (εὖγε εὖγε).

Ἄλλο ὅμως πειστικώτερον ἴσως παράδειγμα, καί τοι οἰκεῖον, θέλω φέρει ὑμῖν εἰς ἀπόδειξιν τῆς ἀνάγκης νὰ ἐγγραφῶσιν ἐν τῷ πολιτεύματι τὰ ἄρθρα 10 καὶ 11. Ὅτ' ἐπαύθην ἀπὸ τοῦ Πανεπιστημίου, ἡ ἐξουσία δὲν μοὶ ἐπέτρεψε νὰ συναγάγω περὶ ἐμὲ τοὺς μαθητάς μου καὶ ἄλλους φιλομαθεῖς νέους καὶ διδάσκω αὐτοῖς τὸ Συνταγματικὸν δίκαιον, ὅπερ μέχρι τῆς ἀπολύσεώς μου αὐτὴ ἡ ἐξουσία ἐπέτρεπέ μοι ἐπὶ ἓξ ὅλα ἔτη, ἢ μᾶλλον ἐντεταλμένον με εἶχε, νὰ ἑρμηνεύω ἐπιστήμως ἀπὸ τῆς ἕδρας τοῦ Πανεπιστημίου, ὥστε οὔτε τὴν πρόφασιν ὅτι ἤμην ἀδόκιμος νὰ προτείνῃ ἠδύνατο, οὐδ' ὅτι αἱ ἀρχαί μου, γνωσταὶ πρὸ πολλοῦ, ἦσαν ἀνατρεπτικαί. Τὸ ἐμὸν παράδειγμα ἐπικρατέστερον τοῦ ὑμετέρου Κ. Καλλιγᾶ νομίζω εἶναι, διότι τὸ ἐμὸν ἀπέβλεπε εἰς τὴν ἀποτροπὴν τοῦ ἐπικειμένου τῇ πολιτείᾳ κακοῦ ἐκ τῆς ἑκάστοτε παραβιάσεως τοῦ Συντάγματος, καὶ ἀπόδειξις τούτου, ὅτι ἔκτοτε ἡ τυραννία ἤρχισε νὰ προβαίνῃ ἀπὸ βαθμίδος εἰς βαθμίδα, ἕως ὅτου ἔφθασεν εἰς τὸν κολοφῶνα, ἐκεῖθεν δ' ἐκρημνίσθη εἰς τὸ βάραθρον. (εὖγε εὖγε !)

Φοβεῖσθε τοὺς συνεταιρισμοὺς καὶ τὰς συναθροίσεις, διότι προκαλοῦσι, λέγετε, τὰς ἐπαναστάσεις, διατί ὅμως δὲν ῥίπτετε ἓν βλέμμα, οὐχὶ ἐπὶ τῆς ἐποχῆς τῆς φλεγμαινούσης δημοκρατίας ἀπὸ τοῦ 1789 καὶ ἐφεξῆς, ὅτε τοῦ ὅλου κοινωνικοῦ οἰκοδομήματος ἡ ἀνατροπὴ ἐγίνετο, ὅτε συνέβαινον αἱ ὠδῖνες ἐξ ὧν ὁ νεώτατος πολιτισμὸς παρήχθη, ὠδῖνες, ἃς ἀδύνατον ἦν νὰ προλάβῃ οἱαδήποτε ἀνθρωπίνη δύναμις, ἀλλ' ἐπὶ τῆς ἐποχῆς τῆς ἱστορίας τῆς Γαλλίας κατὰ πολὺ τῆς πρώτης ὁμαλωτέρας, τῆς ἐποχῆς λέγω τοῦ 1848, ὅτε ἐμποδίσθη ἡ ἔκφρασις τῶν συλλόγων καὶ τῶν συναθροίσεων, διότι εἰς τὰ πολιτικὰ συμπόσια, ἃ ἐδόθησαν, ὄχι μόνον κωλύματα παρενεβλήθησαν ὑπὸ τῆς ἐξουσίας, ἀλλὰ καὶ οἱ μετασχόντες αὐτῶν Βουλευταὶ ὑπὸ τῆς Βουλῆς τοῦ Λουδοβίκου Φιλίππου ἀπεδοκιμάσθησαν, ὅπως ἴδητε ὅτι πολλαὶ δὲν παρῆλθον ἡμέραι καὶ ἐξεῤῥάγη ἡ ἐπανάστασις, καὶ ἀνέτρεψε τὸν θρόνον ὅστις ἐφαίνετο ἐπὶ στερεωτάτης βάσεως ἱδρυμένος;

Ἀ! Κύριε Καλλιγᾶ, τὸ πῦρ ὅταν ἦναι φανερὸν θερμαίνει, ὅταν ὅμως ὑφέρπῃ, οἷα τις ὑπόνομος, ἀνατρέπει· μὴ ἀφίνετε λοιπὸν αὐτὸ νὰ βόσκῃ λαθραίως ἐν ὑπονόμῳ, διότι πάντως θέλει φέρει εἰς ἔκρηξιν· εἰς τοὐμφανὲς ὅμως ὅταν καίῃ δύναταί τις νὰ διευθύνῃ τὴν φλόγα αὐτοῦ· ἐν τῷ φανερῷ τὸ πῦρ θερμαίνει καὶ φωτίζει, οὐδεὶς δὲ κίνδυνος πυρκαϊᾶς, ἀρκεῖ μικρὰ προσοχὴ εἰς πρόληψιν αὐτῆς. (εὖγε! εὖγε)! Ἐν πάσῃ πολιτείᾳ ὑπάρχει ἀντιπολίτευσις, καὶ ἡ ἀντιπολίτευσις αὕτη εἶναι ὡς ζέοντος ὕδατος ἀτμός, μὴ λοιπὸν κατακλείσητε πανταχόθεν τὸν ζέοντα λέβητα, διότι πάντως θέλει προέλθει ἔκρηξις, ἄφετε ἔξοδον εἰς τὸν ἀτμόν, μὴ κλείσητε τὴν δικλίδα τῆς ἀσφαλείας τῆς πολιτείας.

Εἶπεν ὁ ἔντιμος προλαλήσας, ὅτι οὐδεὶς ἐμπόδισεν αὐτὸν νὰ ἔχῃ συναθροίσεις ἐν τῇ ἰδίᾳ αὐτοῦ οἰκίᾳ, ἴσως διότι ἐθεώρει ἡ ἐξουσία ἀκινδύνους πνεύματος καὶ εἰρωνείας καὶ σκαλαθυρμάτων ἀκαδημίας τὰς ἐν οἴκῳ τοῦ Κ. Καλλιγᾶ συναθροίσεις, τίνος ὅμως ἕνεκα τὰ δικαιώματα ταῦτα νὰ θέλῃ μόνον ὑπὲρ ἑαυτοῦ, καὶ νὰ μὴ ἐπιτρέπῃ αὐτῶν τὴν χρῆσιν καὶ ἑτέροις πρὸς σπουδαιότερα σκοποῦσιν; Ἀ! Κύριε Καλλιγᾶ, πολὺ καλῶς γνωρίζετε ὅτι τῇ γελωτοποιΐᾳ καὶ τῇ κωμῳδίᾳ οὐ μόνον ἐπιτρέπεται πλήρης ἐλευθερία ἐν ταῖς ἀπολύτοις μοναρχίαις, ἀλλὰ καὶ ἐμψύχωσις δίδεται αὐταῖς, τῆς σπουδαιολογίας μόνον τὰ στόματα φιμοῦνται.

Ἀλλ' ἡ Ἑλλὰς εὑρίσκεται ἆρά γε εἰς ἥττονα θέσιν τὴν σήμερον ἢ ὅτι ἡ Ἀγγλία πρὸ δύο ἢ τριῶν αἰώνων; ἦτο δ' ἆράγε συγκεχωρημένον ὑμῖν Κ. Καλλιγᾶ, νὰ ἐκφέρητε μίαν ἱστορικὴν βλασφημίαν, ὅτι ὁ Βασιλεὺς ἐν Ἀγγλίᾳ πρὸ τοῦ μεγάλου χάρτου ἦτο ἀσθενής, οἱ δὲ Βαρῶνοι ἰσχυροί; λέγω ὅτι τοῦτο εἶναι μία ἱστορικὴ βλασφημία, διότι καταβεβλημένη μὲν ἐκεῖ ἦτο ἡ ἐλευθερία, πᾶσα δ' ἡ δύναμις ὑπῆρχεν εἰς τὴν βασιλείαν, ἣν ἡ Νορμανδικὴ κατάκτησις ὑπὲρ πᾶσαν ἄλλην βασιλείαν ἰσχυρὰν ἀνίδρυσεν, οἱ δὲ Νορμανδοὶ βαρῶνοι ἀσθενεῖς ὄντες, ὡς ἔχοντες τὴν καταγωγὴν ὅλως ξένην, εἶχον ἀνάγκην νὰ ὑποστηρίζωνται ὑπὸ τῶν ἰθαγενῶν τιμαριωτικῶς αὐτοῖς ὑποτελῶν, ὅπως βαθμηδὸν ἐνισχυθῶσιν· οὕτω δ' ὁ σύνδεσμος μεταξὺ τοῦ Βαρώνου καὶ τοῦ ὑπηκόου ἀνέπτυσσεν ἀνεπαισθήτως τὴν ἐλευθερίαν, ἐν τῇ ἀμοιβαίᾳ πρὸς ἀλλήλους πίστει καὶ φιλίᾳ, μάλιστα δ' ἐν τῷ κοινῷ συμφέροντι. Οὕτως ἐν Ἀγγλίᾳ, ἔνθεν μὲν Βασιλεὺς πανσθενὴς τὸ πρῶτον, ἀπέβαλλε ἐπὶ δύο συνεχεῖς αἰῶνας ἑκάστοτε τῆς ἑαυτοῦ ἰσχύος, ἑτέρωθεν δὲ ἡ δύναμις τῶν Βαρώνων, κατὰ τὸ αὐτὸ τοῦ χρόνου διάστημα, ἀνεπτύσσετο καὶ ἐξηπλοῦτο μέχρι τῶν φοβερῶν χρόνων τῶν Τουδόρων, οἵτινες ἐπειράθησαν νὰ πνίξωσι τὴν ἐλευθερίαν, ὅπως καὶ ἐν Ἀγγλίᾳ ἱδρύσωσι τὴν ἀνὰ πᾶσαν τὴν λοιπὴν Εὐρώπην ἐπικρατήσασαν παμβασιλείαν· ἀλλ' ἡ ἐλευθερία ἦτο ἤδη ἀσφαλῶς περιχαρακωμένη ἐντὸς τοῦ Κοινοβουλίου, ἡ δ' ἀνάγκη τῆς ἀμοιβαίας ὑποστηρίξεως τῶν δύο Βουλῶν, πρὸς τὴν ἰδίαν ἑκάστης αὐτῶν σωτηρίαν, ἐγένετο ἡ αἰτία ἵνα ἡ ἀριστοκρατία τῆς Ἀγγλίας διατηρηθῇ ἀπαραμείωτον φέρουσα τὸν χαρακτῆρα ἀγαθῆς καὶ λαοφιλοῦς ἀριστοκρατίας, διότι πάντοτε ὑπῆρξε συνδεδεμένη μετὰ τοῦ λαοῦ, ἐξ οὗ ἀποῤῥέει καὶ εἰς ὃν ἐπανέρχεται. διότι καὶ τὰ ὑστερογενῆ τοῦ Λόρδου τέκνα ἐπανέρχονται εἰς τὸν λαὸν, καὶ ὁ υἱὸς ἀνδρὸς ἐκ τοῦ λαοῦ δύναται νὰ περιβληθῇ τὴν ἰκτειδῆ πορφύραν τοῦ Λόρδου. Ἰδοὺ Κύριε Καλλιγᾶ, διὰ τί, — ἡ δὲ ἱστορία τοῖς λόγοις μου ἐπιμαρτυρεῖ, — αἱ ἐλευθερίαι ἐν Ἀγγλίᾳ ὑπάρχουσιν ἰσχυραί. Ἡ ἀνικανότης τῶν ἀναξίων ἀπογόνων τοῦ ἐκ Νορμανδίας Γουλιέλμου τοῦ Κατακτητοῦ, καὶ ἰδίως ἡ Ἰωάννου τοῦ Ἀκτήμονος ἀβέλτερος τυραννία ἔκαμαν, ὥστε οἱ Βαρῶνοι νὰ ἁρπάσωσι τὸν μέγαν χάρτην, καὶ εἰκοσιπέντε φύλακας αὐτοῦ νὰ διορίσωσι.

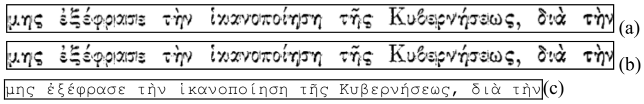
Ἔφερεν ἔτι ὑμῖν τὸ παράδειγμα τῆς Ἰρλανδίας· καὶ ἐνταῦθα ἄλλην μεγίστην πολιτικὴν βλασφημίαν ἐξέφερεν ὁ Κ. Καλλιγᾶς. Μετὰ τὴν μάχην τῆς Βοΰνης ἡ Ἰρλανδία ὑπέπεσεν εἰς ἐπαχθέστατον ζυγὸν, πολὺ ἐπαχθέστερον ἐκείνου, ὅστις ἐβάρυνεν ἐπ' αὐτὴν ἀπὸ Ἐῤῥίκου τοῦ Β', ἰδίως μάλιστα ἀπὸ τῶν Τουδόρων καὶ τῆς διασχίσεως τῆς Δυτικῆς Ἐκκλησίας. Καὶ ἡ ἀτυχὴς Ἰρλανδία ἤθελε πάντως μείνει εἰς τὸ αὐτὸ σημεῖον ἐὰν ἐπ' ἐσχάτων ἡ μεγάλη φωνὴ τοῦ Ὀκονέλλου δὲν ἐχρᾶτο τῇ ἐλευθερίᾳ τῇ Ἀγγλικῇ καὶ δὲν συνῆγε τὰς μεγάλας καὶ καταπληκτικὰς ἐκείνας συναθροίσεις, ὅπως καταστήσῃ αὐτὴν ἰσοπολίτιδα πρὸς τὴν Ἀγγλίαν· ἂν δὲ ἡ Ἰρλανδία καὶ σήμερον συνταράσσεται, δὲν συνταράσσεται ὅμως διὰ πολιτικὸν λόγον, ἀλλὰ διὰ θρησκευτικόν, καὶ οὐδόλως ἔξεστιν ὑμῖν νὰ συγχίζητε τὰ θρησκευτικὰ μετὰ τῶν πολιτικῶν, ἐνῷ τὰ θρησκευτικὰ ἄλλον τε λόγον τοῦ εἶναι, καὶ δριμύτητα μείζονα τῶν πολιτικῶν παθῶν ἔχουσιν. Ἐπειδὴ δ' εὑρισκόμεθα εἰς τὴν Ἀγγλίαν, ἐπιτρέψατέ μοι νὰ σᾶς ἀναφέρω λόγους τινὰς τοῦ Βένθαμ, ὅστις ἐπέρχεταί μοι κατὰ νοῦν πρὸς ἀπόδειξιν, ὅτι οὗτος θεωρεῖ τὰ δικαιώματα ταῦτα ὡς τὸ γνώρισμα τῶν ἐλευθέρων πολιτειῶν. Ὁ Βένθαμ λοιπὸν παραβάλλων τὴν ἐλευθερίαν τῆς Ἀγγλίας πρὸς τὴν δουλείαν τῆς Τουρκίας λέγει ὅτι, ἐν μὲν τῇ Ἀγγλίᾳ ἡ κοινὴ γνώμη ἐκφράζεται διὰ συναθροίσεων καὶ ἀναφορῶν, ἐν δὲ τῇ Κωνσταντινουπόλει καὶ λοιπῇ Τουρκίᾳ ἐκδηλοῦται διὰ πυρκαϊῶν. Εἰ δὲ δὲν πείθουσιν ὑμᾶς οἱ πρακτικοὶ οὗτοι λόγοι τοῦ Βένθαμ, ὡς πρὸς τὴν διαφορὰν, ἥτις ὑπάρχει μεταξὺ τῆς ἐλευθέρας ἐκδηλώσεως

Fig. 2. Sample of a two-column document page.

recognition [10]. For convenience, we briefly describe in the sequel the two methods.

The first feature set consists of the geometric features calculated on a single pixel window (vertical line of pixels). In particular, the first three features are the number of black pixels with respect to the height, followed by the respective first and second moments. Features four and five give the position of the upper and lower contour respectively. The next two features give information on the orientation of the upper and lower contour by calculating their gradient with respect to the horizontal direction. The eighth feature is the number of black-white transitions in the vertical direction and, finally, the last feature is the number of black pixels between the upper and lower contour. All the features mentioned above are normalized with respect to the text line image height.

The second type of features is obtained by stacking row-wise the values of the pixels of each frame in a single vector and subsequently applying Principal Component Analysis (PCA) to reduce their dimensionality. These features are augmented by adding 4 geometric features representing the x and y centroid of the black pixels of the frame as well as their deviations from the centroid in the x and y directions.



μης ἐξέφρασε τὴν ἱκανοποίηση τῆς Κυβερνήσεως, διὰ τὴν (c)

Fig. 3 (a) Grayscale textline image extracted from a page document image, (b) Corresponding binarized textline image, and (c) Ground truth text.

C. Hidden Markov Models

Hidden Markov Models (HMMs) have been successfully used in problems with sequential learning, [11] such as continuous speech, handwritten and cursive text recognition, [9], [12], [13]. In handwritten and cursive text recognition text lines are converted into a sequence of features extracted from a sliding window to the writing direction of the script. As a result, HMM-based recognizers do not require segmentation into words or characters. This particular advantage we exploit in HMM modeling and recognition of Greek polytonic scripts.

HMMs are statistical generative models that used in systems demonstrating a Markov property on a latent or hidden state space. In HMMs the states are not visible but rather hidden and each state is associated with a probability distribution for the possible observations, called the emission probabilities. Additionally, states are related to each other with state transition probabilities and there is also a probability distribution which is associated to the initial state. The generation of a sequence of observations using an HMM is performed in the following fashion. Initially, a random state is selected and an observation is generated from the respective emission probability distribution. At the next time instant or frame, the next state is selected from the transition state probabilities and an observation is generated from the emission probability distribution. The process goes on until one reaches the final state, depending on the number of observations one wants to generate.

In our approach, we model each character with a multistate, left to right HMM, with Gaussian Mixture Models (GMMs) as emission probability density function over the feature space. A text-line is modeled by concatenating the models for each character in the text-line. Note that space is also modeled as a

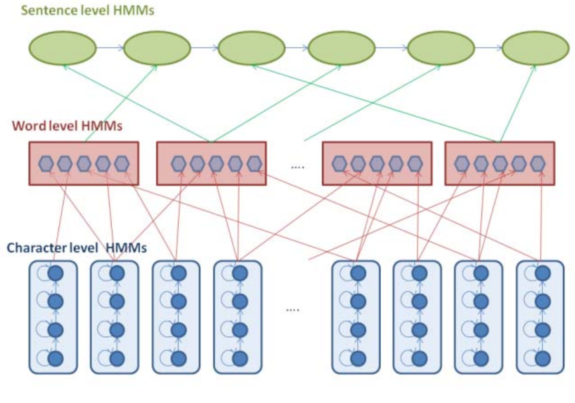


Fig.4. Network of HMMs to perform best path searching in the recognition phase.

separate character. GMMs are parameterized by the weights, the means and the variances for each Gaussian mixture. The number of mixtures for the GMMs, number of states of the HMMs and the transitions allowed among the states are also parameters of the model. Since GMMs model the observations/features, the required number of mixtures depends on the features' variability along the vertical axis and the dimension of the feature space. On the other hand, the number of states for the characters' HMMs depends on the horizontal variability of each character. These parameters, i.e. the number of states per character model and the number of Gaussian mixtures need to be tuned empirically. In our experiments we have used HMMs with states from 3 to 6 and Gaussian mixture densities from 64 up to 512. It must be noted that all character HMMs are modeled with the same number states. After validation we set HMMs to four states, with a left to right topology with self loops and transitions to adjacent only states with no skips. For each HMM two more non-emitting states are attached in the "beginning" and in the "end" which are used to provide transitions from one character model to other character models. Text lines are modeled either by concatenating character models in ergodic structure as shown in Fig. 4 or by concatenating individual word models which result as a concatenation of character HMMs (Fig. 4). Training for estimating the HMMs parameters is performed using the Baum-Welch re-estimation algorithm [11], by aligning iteratively the feature vectors with the character models in a maximum likelihood sense for each text-line.

Having estimated the HMM parameters for each target character, at the testing phase, after feature extraction the recognition process searches for a sequence of character models that has the highest probability to generate the given sequence of feature vectors that corresponds to the input text-line. The search process, except of the trained HMM character models, it uses a word or character lexicon and an appropriate statistical language model. The recognition is performed using the Viterbi algorithm [12] to perform best path search in combinations of different character level models in a network of concatenated HMMs shown in Fig. 4.

The choice of lexicon and language model is optional. In the presented system, we employ an open-vocabulary, by building a language model of back-off bigrams at character level. Higher order language models at character or word level are also possible and they generally result to lower error rates.

All the experiments for building HMMs are done using the Hidden Markov Toolkit (HTK) [12]. HTK is a portable toolkit that is primarily developed for building automatic speech recognition systems, but it can also be used for handwritten and optical character recognition tasks. The development of the language models have been carried out on a text corpus extracted from the Thesaurus Linguae Graecae (TLG) [15] using the SRI Language Modeling Toolkit [16][17].

III. Experimental Results and Evaluation

We evaluate state-of-the-art open-source and commercial OCR engines for Greek polytonic and compare their performance with the proposed HMM based method. The participating OCR engines are ABBYY FineReader v.11 [18] and Tesseract OCR engine [19]. In addition, we compare our results with the state-of-the-art systems suggested in [3] and [22].

The performance evaluation is carried out by comparing the recognized text with the ground truth and computing the character error rate (CER) with the help of the following formula

$$CER = \frac{ND + NI + NS}{N} \times 100$$

where N is the total number of characters, ND the number of deletions, NI the number of insertions and NS the number of substitutions. In a similar fashion, we calculate the word error rate (WER) by counting the number of matching words between recognized and ground truth texts and normalizing it with respect to the total number of words in the ground truth. A word is considered as an error if it has at least one character error. More details on the accuracy measures at character and word level can be found in [20].

As mentioned above, in our experiments we apply to sets of features. The first is denoted with *fs1* and is referred to the feature vector of 9 geometric features calculated for each pixel column of text line images. The second is denoted with *fs2* and we have selected the dimension of 20 for the PCA components,

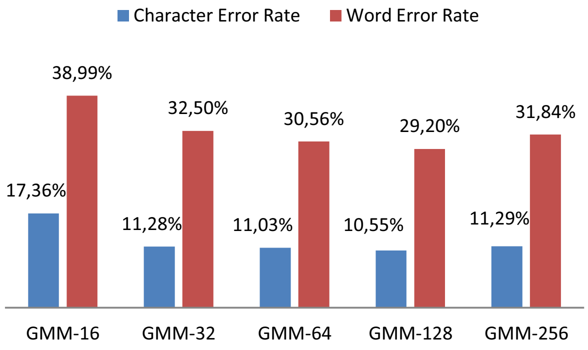


Fig. 5. Character and word error rates of HMMs for feature set *fs1* varying the number of Gaussian Mixtures Models on fold-1.

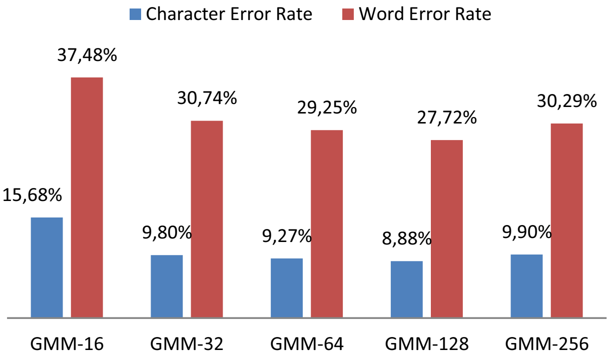
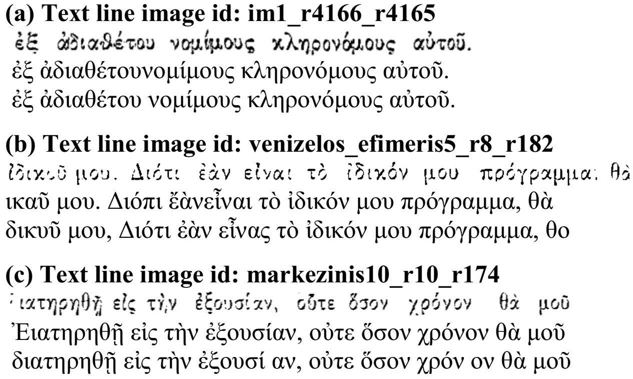


Fig. 6. Character and word error rates of HMMs for feature set *fs2* varying the number of Gaussian Mixtures Models on fold-1.

resulting to feature vectors of dimension 24. It must be noted that before feature extraction, each text line image was normalized to 60 pixels height. The size of the frames is set to 11x60 pixels and they are shifted by 1 pixel from left to right.

In our experiments we have used 12-folds keeping the ratios of Table I for the individual subsets of each fold. Two folds have been used for validating the number of states of the HMMs and the number of the mixtures for the GMMs, and the remainder 10 for running the evaluation experiments. We have varied the number of states of the HMMs from 3 to 6 states and have obtained the best results with 4-state HMMs for *fs1* and 3-state HMMs for *fs2*. We have run validation experiments for various numbers of mixtures for the emission probabilities in order to select the most suitable number. In Figs. 5 and 6 we illustrate the CER and WER for various numbers of mixtures on the validation fold. It can be seen that the lowest CER and WER for both feature sets *fs1* and *fs2* are obtained for the value of 128 mixtures. In the evaluation experiments that follow we adopt character-bigrams language models using 128 GMMs for both the *fs1* and the *fs2* HMMs.

For Tesseract no training was necessary, as we used the model for Greek polytonic built by Nick White [21]. For the ABBYY FineReader Engine we selected as training set 367 text-lines from the total of 3,202 of the Greek Parliament Proceedings, in a way so that each target character-class appears at least 5 times. We then semi -automatically segmented the selected text-line images into character images and used the training utility of the ABBYY FineReader engine SDK to create the respective characters' models. In addition, we have built a dictionary for Katharevousa (a form of the Greek language in the early 19th century) with the use of texts from the TLG corpus [15]. The remaining text-line images of the dataset were used for testing the performances of Tesseract and ABBYY FineReader.


(a) Text line image id: im1_r4166_r4165

ἐξ ἀδιαθέτου νομίμους κληρονόμους αὐτοῦ.

ἐξ ἀδιαθέτουνομίμους κληρονόμους αὐτοῦ.

ἐξ ἀδιαθέτου νομίμους κληρονόμους αὐτοῦ.

(b) Text line image id: venizelos_efimeris5_r8_r182

ἰδικοῦ μου. Διότι ἐὰν εἶναι τὸ ἰδικόν μου πρόγραμμα, θὰ

ικαῦ μου. Διόπι ἔὰνεἶναι τὸ ἰδικόν μου πρόγραμμα, θὰ

δικυῦ μου, Διότι ἐὰν εἶνας τὸ ἰδικόν μου πρόγραμμα, θο

(c) Text line image id: markezinis10_r10_r174

ιατηρηθῇ εἰς τὴν ἐξουσίαν, οὔτε ὅσον χρόνον θὰ μοῦ

Ἐιατηρηθῇ εἰ̓ς τὴν ἐξουσίαν, οὐ̓τε ὅσον χρόνον θὰ μοῦ

διατηρηθῇ εἰ̓ς τὴν ἐξουσί αν, οὐ̓τε ὅσον χρόν ον θὰ μοῦ

Fig. 7. Samples of recognized text-lines. The first result bellow the text-line image corresponds to the HMM recognizer for the feature set *fs1* and the second to the feature set *fs2*.

TABLE II. CHARACTER ERROR RATES (CER) IN % OF GREEK POLYTONIC

Dataset	Tesseract	ABBYY FineReader	HMM-GMM128 (fs1)	HMM-GMM128 (fs2)
Saripolos/ 1864	28.41	23.60	22.00	21.38
Venizelos / 1931	22.29	15.28	15.76	14.78
Markezinis/ 1953	31.13	19.70	7.66	6.21
Vlahou/ 1977	42.30	14.34	3.08	1.35
FEK/ 1959	-	-	3.99	2.33
Total	**30.37**	**19.20**	**9.93**	**8.61**

TABLE III. WORD ERROR RATES (WER) IN % OF GREEK POLYTONIC

Dataset	Tesseract	ABBYY FineReader	HMM-GMM128 (fs1)	HMM-GMM128 (fs2)
Saripolos/ 1864	71.71	46.69	41.09	33.37
Venizelos / 1931	66.71	55.54	36.84	29.20
Markezinis/ 1953	71.36	48.61	31.35	23.78
Vlahou/ 1977	77.61	42.51	28.23	20.72
FEK/ 1959	-	-	28.86	21.35
Total	**71.43**	**48.60**	**32.89**	**25.30**

Tables II and III present the evaluation results recorded concerning error rates on character and word level for the Tesseract, ABBYY FineReader and the HMM-based recognition systems. It can be seen that HMM-based recognition systems overcome the problem of character segmentation and outperform the open source and commercial engines of Tesseract and ABBYY FineReader respectively for both accuracy measures of CER and WER in every subset of the dataset. The HMM system based on the feature set *fs2* shows a better recognition performance than the HMM system based on the feature set *fs1*.

It must be noted that the state-of-the-art system that is based on sophisticated character segmentation for Greek polytonic proposed in [3], the authors report a CER of 9.91% and a WER of 37.32% on the *FEK/1959* subset, while for the system proposed in [22] that is based on LSTM networks the authors report a CER of 5.51% and a WER of 24.13%. We can see that the proposed character segmentation free HMM systems outperform the first state-of-the art system, while its performance is comparable with the second.

In Fig. 7 we present examples of recognized text-lines with the HMM recognizer. The HMM-based recognizer performs well for connected consecutive characters, e.g. the case of όμ in of the fourth word κληρονόμους of the text-line in Fig. 7a. Word segmentation is also successful in most cases, but it is better for the HMM of *fs2*. However, one may overcome these errors by employing explicit word segmentation and use character HMMs for word recognition. Insertion errors occur in the case of faded characters as can be seen in Fig. 7c which consists of several segments of connected components, e.g. in the fourth word ἐξουσίαν and the seventh word χρόνον of

Fig. 7c is split into two components ἐξουσ ίαν and χρόν ον respectively. One drawback of training character HMMs in the presence of noise is revealed by the misclassification error by combining the characters ι and κ that appear as in an α as can be seen in the first word of text line in Fig. 7b. Furthermore, there are a great number of errors where a character is misclassified with another character of the same letter, but with different accent. As discussed below, this problem one may overcome with a finer modeling for the HMMs of these characters.

IV. Conclusions and Future Work

In this paper we presented and evaluated HMM-based text-line recognizers for the Greek polytonic scripts. We demonstrated that HMM recognizers overcome the problem of successful character segmentation when there are several combinations of letters with diacritics marks and can outperform state-of-the-art character segmentation, open source and commercial OCR engines, while illustrating comparable results to systems based on LSTM networks. The performance of HMM recognizers may be further improved if one employs more complex language models such as word-bigrams or trigrams at the character level. We also plan to experiment with GMMs with common variance for character classes that are variations of the same letter, but with different diacritics as well as to make the number of states letter dependent. In this way we expect the means of the GMMs to play a more significant role in the discrimination among these characters.

Acknowledgment

The present work has been supported by the OldDocPro ARISTEIA II project funded by the Greek General Secretariat for Research and Technology. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] http://unicode.org/charts/PDF/U1F00.pdf

[2] http://www.unicode.org/charts/PDF/U0370.pdf

[3] B. Gatos, G. Louloudis and N. Stamatopoulos, "Greek Polytonic OCR based on Efficient Character Class Number Reduction," in Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, 2011, pp. 1155-159.

[4] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, V. Papavassiliou, F. Simistira, and V. Katsouros, "GRPOLY-DB: An Old Greek Polytonic Document Image Database," in Proc. of the 13th International Conference on Document Analysis and Recognition (ICDAR2015), Nancy, France, 2015, pp. 646–650.

[5] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive Degraded Document Image Binarization," Pattern Recognition, vol. 39, 2006, pp. 317–327.

[6] B. Gatos, G. Louloudis, and N. Stamatopoulos, "Segmentation of Historical Handwritten Documents into Text Zones and Text Lines," in Proc. of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014), Creta, Greece, 2014, pp. 464–469.

[7] N. Stamatopoulos, G. Louloudis, and B. Gatos, "Efficient Transcript Mapping to Ease the Creation of Document Image Segmentation Ground Truth with Text–Image Alignment," in Proc. of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR2010), Kolkata, India, 2010, pp. 226–231.

[8] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia – An Advanced Document Layout and Text Ground–Truthing System for Production Environments," in the Proc. of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, 2011, pp. 48–52.

[9] U.-V. Marti and H. Bunke, "Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition Systems," in Hidden Markov models: applications in computer vision, pp. 65-90, World Scientific Publishing Co., Inc. River Edge, NJ, USA, 2002.

[10] M. Kozielski, J. Forster, and H. Ney, "Moment-based image normalization for handwritten text recognition," in the Proc. of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR2012), Bari, Italy, 2012, pp. 256-261.

[11] Y. Bengio, "Markovian Models for Sequential Learning," Neural Computing Surveys, vol. 2, 1999, pp. 129-162.

[12] F. Jelinek, Statistical methods for speech recognition. Cambridge, MA, USA: MIT Press, 1997.

[13] A.J. Elms and J. Illingworth, "Modelling Polyfont Printed Characters With HMMs and a Shift Invariant Hamming Distance," in the Proc. of the 3rd International Conference on Document Analysis and Recognition (ICDAR1995), Montreal, Canada, 1995, pp. 504-507.

[14] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book Version 3.4. Cambridge University Press, 2006.

[15] http://www.tlg.uci.edu/

[16] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with N-gram features: An SRILM extension," in the Proc. of Interspeech, Makuhari, Japan, Sep. 2010, pp. 1155-159.

[17] http://www.speech.sri.com/projects/srilm/

[18] http://www.abbyy.com.gr/ocr_sdk/

[19] https://code.google.com/p/tesseract-ocr/

[20] S. Rice, Measuring the Accuracy of Page-Reading Systems, PhD Thesis, University of Nevada, Las Vegas, 1996.

[21] Nick White, "Training Tesseract for Ancient Greek OCR," Εὔτυπον, No 28-29 - October 2012 (retrieved from http://eutypon.gr/eutypon/pdf/e2012-29/e29-a01.pdf).

[22] F. Simistira, A. Ul-Hassany, V. Papavassiliou, B. Gatos, V. Katsouros, and M. Liwicki, "Recognition of Historical Greek Polytonic Scripts Using LSTM Networks," in Proc. of the 13th International Conference on Document Analysis and Recognition (ICDAR2015), Nancy, France, 2015, pp. 766–770.