# Improving the quality of the output

Tesseract documentation

View on GitHub

# Improving the quality of the output

There are a variety of reasons you might not get good quality output from Tesseract. It’s important to note that, unless you’re using a very unusual font or a new language, retraining Tesseract is unlikely to help.

## Image processing

Tesseract does various image processing operations internally (using the Leptonica library) before doing the actual OCR. It generally does a very good job of this, but there will inevitably be cases where it isn’t good enough, which can result in a significant reduction in accuracy.

You can see how Tesseract has processed the image by using the configuration variable `tessedit_write_images` to `true` (or using configfile `get.images` ) when running Tesseract. If the resulting `tessinput.tif` file looks problematic, try some of these image processing operations before passing the image to Tesseract.

## Inverting images

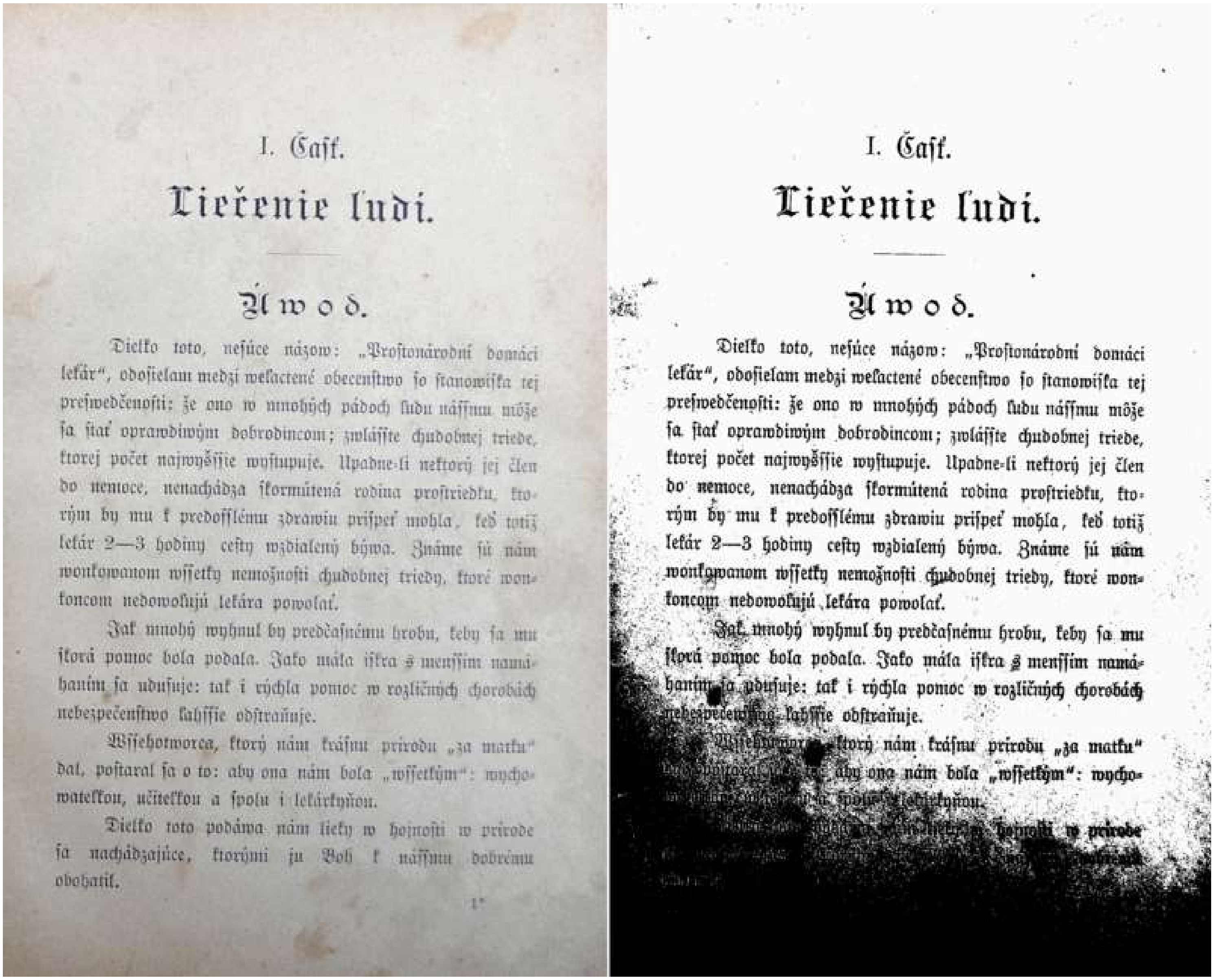
While tesseract version 3.05 (and older) handle inverted image (dark background and light text) without problem, for 4.x version use dark text on light background.

## Rescaling

Tesseract works best on images which have a DPI of at least 300 dpi, so it may be beneficial to resize images. For more information see the FAQ.

“Willus Dotkom” made interesting test for Optimal image resolution with suggestion for optimal Height of capital letter in pixels.

## Binarisation



I. Čaſt.

Liečenie ludi.

Úwod.

Dielko toto, neſúce názow: „Proſtonárodní domáci lekár“, odoſielam medzi weľactené obecenſtwo ſo ſtanowiſka tej preſwedčenoſti: že ono w mnohých pádoch ľudu náſſmu môže ſa ſtať opravdiwým dobrodincom; zwláſſte chudobnej triede, ktorej počet najwyššie wyſtupuje. Upadne-li nektorý jej člen do nemoce, nenachádza ſkormútená rodina proſtriedku, ktorým by mu k predoſſlému zdrawiu priſpeť mohla, keď totiž lekár 2—3 hodiny ceſty wzdialený býwa. Známe ſú nám wonkowanom wſſetky nemožnoſti chudobnej triedy, ktoré wonkoncom nedowoľujú lekára powolať.

Jak mnohý wyhnul by predčaſnému hrobu, keby ſa mu ſkorá pomoc bola podala. Jako mála iſkra ž menſſím namáhaním ſa uduſuje: tak i rýchla pomoc w rozličných chorobách nebezpečenſtwo ľahſſie odſtraňuje.

Wſſehotworca, ktorý nám kráſnu prírodu „za matku“ dal, poſtaral ſa o to: aby ona nám bola „wſſetkým“: wychowateľkou, učiteľkou a ſpolu i lekárkyňou.

Dielko toto podáwa nám lieky w hojnoſti w prirode ſa nachádzajúce, ktorými ju Boh k náſſmu dobrému obohatil.

1*

I. Čaſt.

Liečenie ludi.

Úwod.

Dielko toto, neſúce názow: „Proſtonárodní domáci lekár“, odoſielam medzi weľactené obecenſtwo ſo ſtanowiſka tej preſwedčenoſti: že ono w mnohých pádoch ľudu náſſmu môže ſa ſtať opravdiwým dobrodincom; zwláſſte chudobnej triede, ktorej počet najwyššie wyſtupuje. Upadne-li nektorý jej člen do nemoce, nenachádza ſkormútená rodina proſtriedku, ktorým by mu k predoſſlému zdrawiu priſpeť mohla, keď totiž lekár 2—3 hodiny ceſty wzdialený býwa. Známe ſú nám wonkowanom wſſetky nemožnoſti chudobnej triedy, ktoré wonkoncom nedowoľujú lekára powolať.

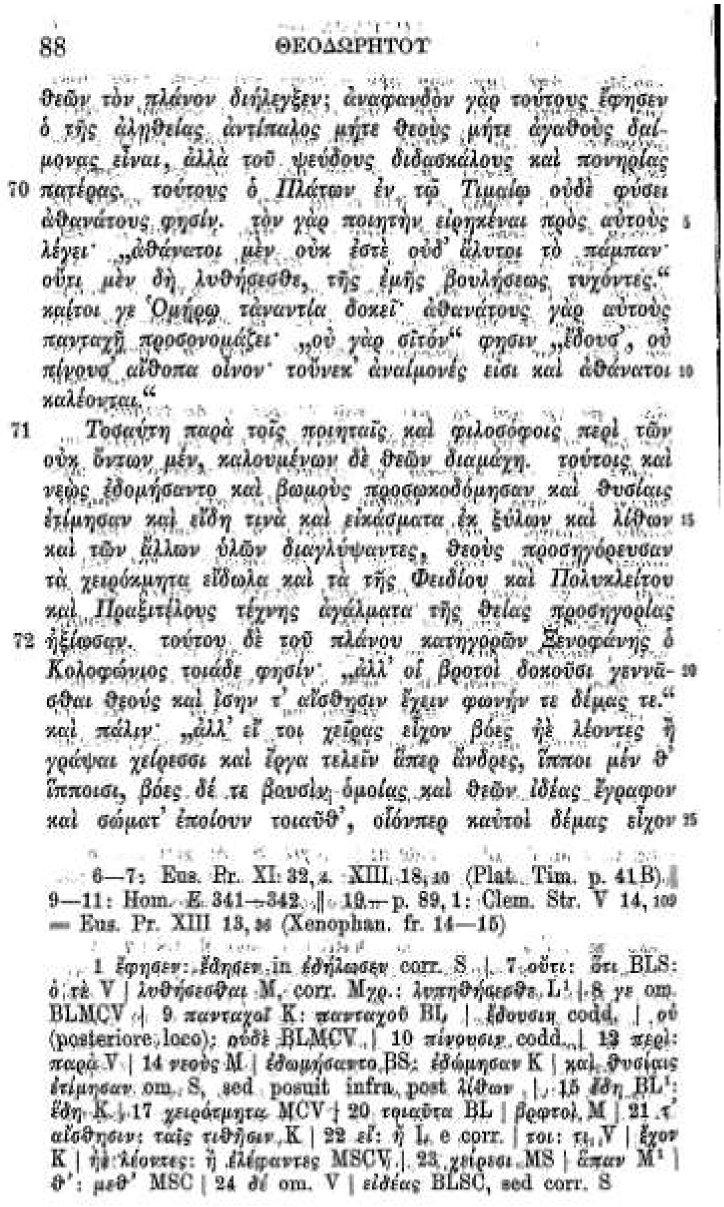
Jak mnohý wyhnul by predčaſnému hrobu, keby ſa mu ſkorá pomoc bola podala. Jako mála iſkra ž menſſím namáhaním ſa uduſuje: tak i rýchla pomoc w rozličných chorobách nebezpečenſtwo ľahſſie odſtraňuje.

This is converting an image to black and white. Tesseract does this internally (Otsu algorithm), but the result can be suboptimal, particularly if the page background is of uneven darkness.

Tesseract 5.0.0 added two new Leptonica based binarization methods: Adaptive Otsu and Sauvola. Use `tesseract --print-parameters | grep thresholding_` to see the relevant configurable parameters.

If you are not able to fix this by providing a better input image, you can try a different algorithm. See ImageJ Auto Threshold (java) or OpenCV Image Thresholding (python) or scikit-image Thresholding documentation (python).

## Noise Removal



88 ΘΕΟΔΩΡΗΤΟΥ

θεῶν τὸν πλάνον διήλεγξεν; ἀναφανδὸν γὰρ τούτους ἔφησεν ὁ τῆς ἀληθείας ἀντίπαλος μήτε θεοὺς μήτε ἀγαθοὺς δαίμονας εἶναι, ἀλλὰ τοῦ ψεύδους διδασκάλους καὶ πονηρίας
70 πατέρας. τούτους ὁ Πλάτων ἐν τῷ Τιμαίῳ οὐδὲ φύσει ἀθανάτους φησίν. τὸν γὰρ ποιητὴν εἰρηκέναι πρὸς αὐτοὺς λέγει· „ἀθάνατοι μὲν οὐκ ἐστὲ οὐδ᾽ ἄλυτοι τὸ πάμπαν· οὔτι μὲν δὴ λυθήσεσθε, τῆς ἐμῆς βουλήσεως τυχόντες.“ καίτοι γε Ὁμήρῳ τἀναντία δοκεῖ· ἀθανάτους γὰρ αὐτοὺς πανταχῆ προσονομάζει· „οὐ γὰρ σῖτον“ φησιν „ἔδουσ᾽, οὐ πίνουσ᾽ αἴθοπα οἶνον· τοὔνεκ᾽ ἀναίμονές εἰσι καὶ ἀθάνατοι καλέονται.“

71 Τοσαύτη παρὰ τοῖς ποιηταῖς καὶ φιλοσόφοις περὶ τῶν οὐκ ὄντων μέν, καλουμένων δὲ θεῶν διαμάχη. τούτοις καὶ νεὼς ἐδομήσαντο καὶ βωμοὺς προσῳκοδόμησαν καὶ θυσίαις ἐτίμησαν καὶ εἴδη τινὰ καὶ εἰκάσματα ἐκ ξύλων καὶ λίθων καὶ τῶν ἄλλων ὑλῶν διαγλύψαντες, θεοὺς προσηγόρευσαν τὰ χειρόκμητα εἴδωλα καὶ τὰ τῆς Φειδίου καὶ Πολυκλείτου καὶ Πραξιτέλους τέχνης ἀγάλματα τῆς θείας προσηγορίας
72 ἠξίωσαν. τούτου δὲ τοῦ πλάνου κατηγορῶν Ξενοφάνης ὁ Κολοφώνιος τοιάδε φησίν· „ἀλλ᾽ οἱ βροτοὶ δοκοῦσι γεννᾶσθαι θεούς καὶ ἴσην τ᾽ αἴσθησιν ἔχειν φωνήν τε δέμας τε.“ καὶ πάλιν· „ἀλλ᾽ εἴ τοι χεῖρας εἶχον βόες ἠὲ λέοντες ἢ γράψαι χείρεσσι καὶ ἔργα τελεῖν ἅπερ ἄνδρες, ἵπποι μέν θ᾽ ἵπποισι, βόες δέ τε βουσὶν ὁμοίας καὶ θεῶν ἰδέας ἔγραφον καὶ σώματ᾽ ἐποίουν τοιαῦθ᾽, οἷόνπερ καὐτοὶ δέμας εἶχον

6—7: Eus. Pr. XI 32, 4. XIII 18, 10 (Plat. Tim. p. 41 B) ‖ 9—11: Hom. E 341—342 ‖ 19 — p. 89, 1: Clem. Str. V 14, 109 = Eus. Pr. XIII 13, 36 (Xenophan. fr. 14—15)

1 ἔφησεν: ἔδησεν in ἐδήλωσεν corr. S | 7 οὔτι: ὅτι BLS: ὅ τε V | λυθήσεσθαι M, corr. Mγρ.: λυπηθήσεσθε $L^1$ | 8 γε om. BLMCV | 9 πανταχοῖ K: πανταχοῦ BL | ἔδουσιν codd. | οὐ (posteriore loco): οὐδὲ BLMCV | 10 πίνουσιν codd. | 12 περὶ: παρὰ V | 14 νεοὺς M | ἐδωμήσαντο BS: ἐδόμησαν K | καὶ θυσίαις ἐτίμησαν om. S, sed posuit infra post λίθων | 15 ἴδη $BL^1$: ἔδη K | 17 χειρότμητα MCV | 20 τοιαῦτα BL | βροτοὶ M | 21 τ᾽ αἴσθησιν: ταῖς τιθῆσιν K | 22 εἴ: ἢ L e corr. | τοι: τι V | ἔχον K | ἠὲ λέοντες: ἢ ἐλέφαντες MSCV | 23 χείρεσι MS | ἅπαν $M^1$ | θ᾽: μεθ᾽ MSC | 24 δέ om. V | εἰδέας BLSC, sed corr. S

Noise is random variation of brightness or colour in an image, that can make the text of the image more difficult to read. Certain types of noise cannot be removed by Tesseract in the binarisation step, which can cause accuracy rates to drop.

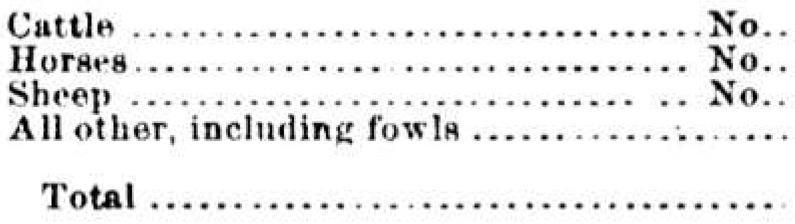
## Dilation and Erosion

Bold characters or Thin characters (especially those with Serifs) may impact the recognition of details and reduce recognition accuracy. Many image processing programs allow Dilation and Erosion of edges of characters against a common background to dilate or grow in size (Dilation) or shrink (Erosion).

Heavy ink bleeding from historical documents can be compensated for by using an Erosion technique. Erosion can be used to shrink characters back to their normal glyph structure.

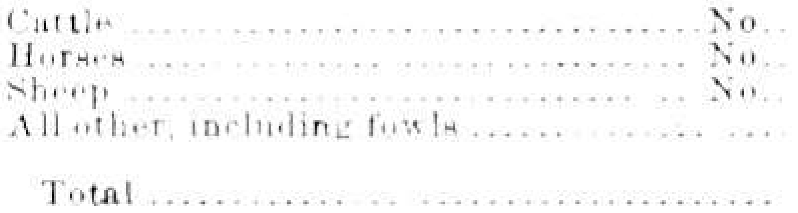
For example, GIMP's Value Propagate filter can create Erosion of extra bold historical fonts by reducing the Lower threshold value.

Original:



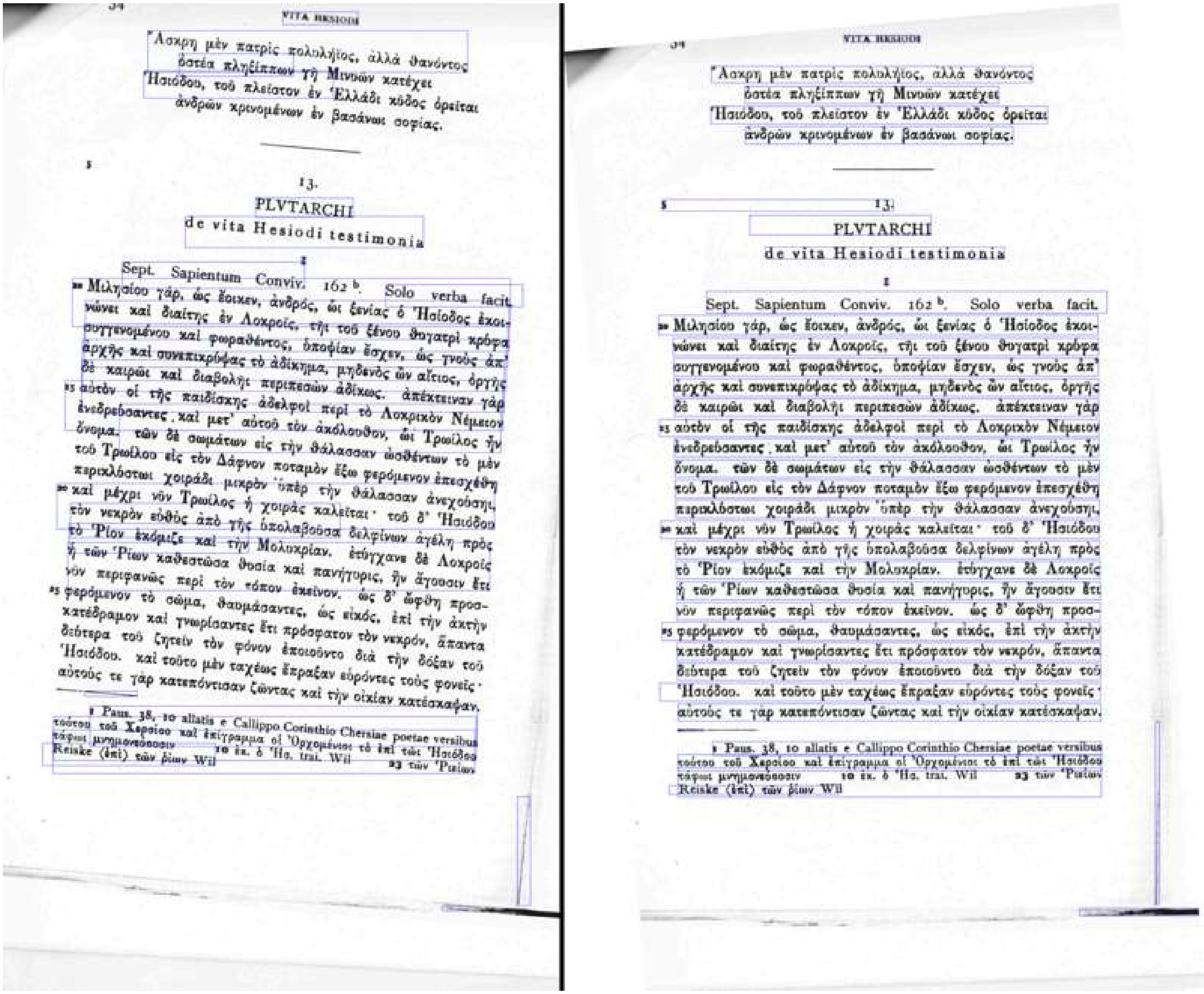
| | |
|---|---|
| Cattle .................................. | No.. |
| Horses.......... ......... .................. | No.. |
| Sheep ................................ .. | No.. |
| All other, including fowls .................... | |
| Total ........................................ | |

Erosion applied:



| | |
|---|---|
| Cattle .................................. | No.. |
| Horses.......... ......... .................. | No.. |
| Sheep ................................ .. | No.. |
| All other, including fowls .................... | |
| Total ........................................ | |

## Rotation / Deskewing



54 VITA HESIODI

Ἄσκρη μὲν πατρὶς πολυλήιος, ἀλλὰ θανόντος
ὀστέα πληξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλεῖστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνωι σοφίας.

13.

PLVTARCHI

de vita Hesiodi testimonia

1

Sept. Sapientum Conviv. 162 b. Solo verba facit.

Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὧι ξενίας ὁ Ἡσίοδος ἐκοινώνει καὶ διαίτης ἐν Λοκροῖς, τῆι τοῦ ξένου θυγατρὶ κρύφα συγγενομένου καὶ φωραθέντος, ὑποψίαν ἔσχεν, ὡς γνοὺς ἀπ' ἀρχῆς καὶ συνεπικρύψας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῆς δὲ καιρῶι καὶ διαβολῆι περιπεσὼν ἀδίκως. ἀπέκτειναν γὰρ αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον ἐνεδρεύσαντες, καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὧι Τρωίλος ἦν ὄνομα. τῶν δὲ σωμάτων εἰς τὴν θάλασσαν ὠσθέντων τὸ μὲν τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσχέθη περικλύστωι χοιράδι μικρὸν ὑπὲρ τὴν θάλασσαν ἀνεχούσηι, καὶ μέχρι νῦν Τρωίλος ἡ χοιρὰς καλεῖται· τοῦ δ' Ἡσιόδου τὸν νεκρὸν εὐθὺς ἀπὸ γῆς ὑπολαβοῦσα δελφίνων ἀγέλη πρὸς τὸ Ῥίον ἐκόμιζε καὶ τὴν Μολυκρίαν. ἐτύγχανε δὲ Λοκροῖς ἡ τῶν Ῥίων καθεστῶσα θυσία καὶ πανήγυρις, ἣν ἄγουσιν ἔτι νῦν περιφανῶς περὶ τὸν τόπον ἐκεῖνον. ὡς δ' ὤφθη προσφερόμενον τὸ σῶμα, θαυμάσαντες, ὡς εἰκός, ἐπὶ τὴν ἀκτὴν κατέδραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἅπαντα δεύτερα τοῦ ζητεῖν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ Ἡσιόδου. καὶ τοῦτο μὲν ταχέως ἔπραξαν εὑρόντες τοὺς φονεῖς· αὐτούς τε γὰρ κατεπόντισαν ζῶντας καὶ τὴν οἰκίαν κατέσκαψαν.

1 Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τοῦ Χερσίου καὶ ἐπίγραμμα οἱ Ὀρχομένιοι τὸ ἐπὶ τῶι Ἡσιόδου τάφωι μνημονεύουσιν 10 ἐκ. ὁ Ἡσ. trai. Wil 23 τῶν Ῥιίων Reiske (ἐπὶ) τῶν βίων Wil

54 VITA HESIODI

Ἄσκρη μὲν πατρὶς πολυλήιος, ἀλλὰ θανόντος
ὀστέα πληξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλεῖστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνωι σοφίας.

13.

PLVTARCHI

de vita Hesiodi testimonia

1

Sept. Sapientum Conviv. 162 b. Solo verba facit.

Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὧι ξενίας ὁ Ἡσίοδος ἐκοινώνει καὶ διαίτης ἐν Λοκροῖς, τῆι τοῦ ξένου θυγατρὶ κρύφα συγγενομένου καὶ φωραθέντος, ὑποψίαν ἔσχεν, ὡς γνοὺς ἀπ' ἀρχῆς καὶ συνεπικρύψας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῆς δὲ καιρῶι καὶ διαβολῆι περιπεσὼν ἀδίκως. ἀπέκτειναν γὰρ αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον ἐνεδρεύσαντες, καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὧι Τρωίλος ἦν ὄνομα. τῶν δὲ σωμάτων εἰς τὴν θάλασσαν ὠσθέντων τὸ μὲν τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσχέθη περικλύστωι χοιράδι μικρὸν ὑπὲρ τὴν θάλασσαν ἀνεχούσηι, καὶ μέχρι νῦν Τρωίλος ἡ χοιρὰς καλεῖται· τοῦ δ' Ἡσιόδου τὸν νεκρὸν εὐθὺς ἀπὸ γῆς ὑπολαβοῦσα δελφίνων ἀγέλη πρὸς τὸ Ῥίον ἐκόμιζε καὶ τὴν Μολυκρίαν. ἐτύγχανε δὲ Λοκροῖς ἡ τῶν Ῥίων καθεστῶσα θυσία καὶ πανήγυρις, ἣν ἄγουσιν ἔτι νῦν περιφανῶς περὶ τὸν τόπον ἐκεῖνον. ὡς δ' ὤφθη προσφερόμενον τὸ σῶμα, θαυμάσαντες, ὡς εἰκός, ἐπὶ τὴν ἀκτὴν κατέδραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἅπαντα δεύτερα τοῦ ζητεῖν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ Ἡσιόδου. καὶ τοῦτο μὲν ταχέως ἔπραξαν εὑρόντες τοὺς φονεῖς· αὐτούς τε γὰρ κατεπόντισαν ζῶντας καὶ τὴν οἰκίαν κατέσκαψαν.

1 Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τοῦ Χερσίου καὶ ἐπίγραμμα οἱ Ὀρχομένιοι τὸ ἐπὶ τῶι Ἡσιόδου τάφωι μνημονεύουσιν 10 ἐκ. ὁ Ἡσ. trai. Wil 23 τῶν Ῥιίων Reiske (ἐπὶ) τῶν βίων Wil

A skewed image is when a page has been scanned when not straight. The quality of Tesseract's line segmentation reduces significantly if a page is too skewed, which severely impacts the quality of the OCR. To address this rotate the page image so that the text lines are horizontal.

## Borders

### Missing borders

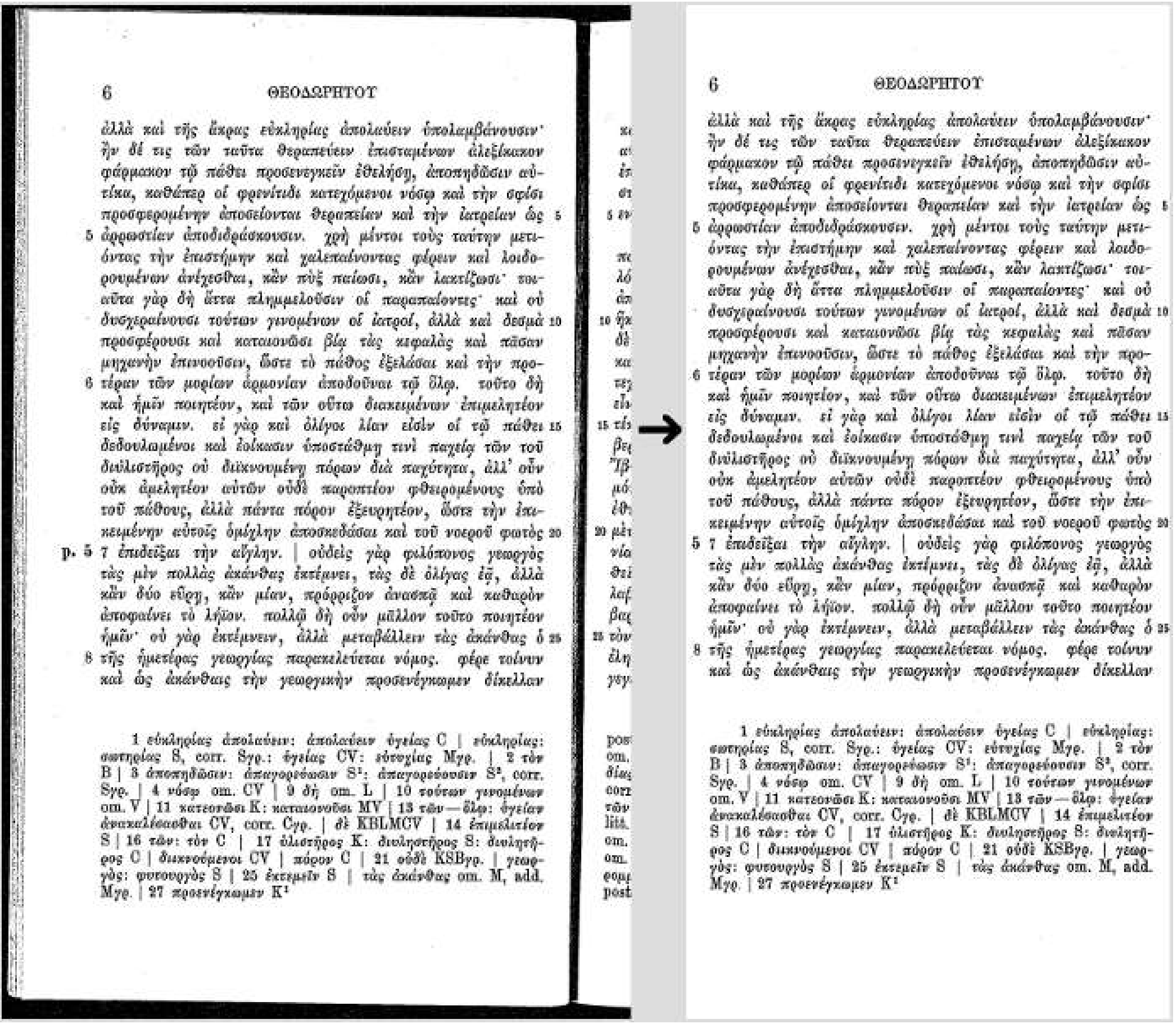
If you OCR just text area without any border, tesseract could have problems with it. See for some details in tesseract user forum#427 . You can easy add small border (e.g. 10 px) with ImageMagick®:

```
convert  427-1.jpg  -bordercolor White -border 10x10 427-1b.jpg
```

## Too big borders

Big borders (especially when processing a single letter/digit or one word on a large background) can cause problems ("empty page"). Please try to crop you input image to a text area with reasonable border (e.g. 10 px).

## Scanning border Removal



6 ΘΕΟΔΩΡΗΤΟΥ

ἀλλὰ καὶ τῆς ἄκρας εὐκληρίας ἀπολαύειν ὑπολαμβάνουσιν·
ἢν δέ τις τῶν ταῦτα θεραπεύειν ἐπισταμένων ἀλεξίκακον
φάρμακον τῷ πάθει προσενεγκεῖν ἐθελήσῃ, ἀποπηδῶσιν αὐτίκα,
καθάπερ οἱ φρενίτιδι κατεχόμενοι νόσῳ καὶ τὴν σφίσι
προσφερομένην ἀποσείονται θεραπείαν καὶ τὴν ἰατρείαν ὡς
5 ἀρρωστίαν ἀποδιδράσκουσιν. χρὴ μέντοι τοὺς ταύτην μετιόντας
τὴν ἐπιστήμην καὶ χαλεπαίνοντας φέρειν καὶ λοιδορουμένων
ἀνέχεσθαι, κἂν πὺξ παίωσι, κἂν λακτίζωσι· τοιαῦτα
γὰρ δὴ ἄττα πλημμελοῦσιν οἱ παραπαίοντες· καὶ οὐ
δυσχεραίνουσι τούτων γινομένων οἱ ἰατροί, ἀλλὰ καὶ δεσμὰ
προσφέρουσι καὶ καταιονῶσι βίᾳ τὰς κεφαλὰς καὶ πᾶσαν
μηχανὴν ἐπινοοῦσιν, ὥστε τὸ πάθος ἐξελάσαι καὶ τὴν προ-
6 τέραν τῶν μορίων ἁρμονίαν ἀποδοῦναι τῷ ὅλῳ. τοῦτο δὴ
καὶ ἡμῖν ποιητέον, καὶ τῶν οὕτω διακειμένων ἐπιμελητέον
εἰς δύναμιν. εἰ γὰρ καὶ ὀλίγοι λίαν εἰσὶν οἱ τῷ πάθει
δεδουλωμένοι καὶ ἐοίκασιν ὑποστάθμῃ τινὶ παχείᾳ τῶν τοῦ
διϋλιστῆρος οὐ διϊκνουμένῃ πόρων διὰ παχύτητα, ἀλλ' οὖν
οὐκ ἀμελητέον αὐτῶν οὐδὲ παροπτέον φθειρομένους ὑπὸ
τοῦ πάθους, ἀλλὰ πάντα πόρον ἐξευρητέον, ὥστε τὴν ἐπικειμένην
αὐτοῖς ὁμίχλην ἀποσκεδάσαι καὶ τοῦ νοεροῦ φωτὸς
p. 5 7 ἐπιδεῖξαι τὴν αἴγλην. | οὐδεὶς γὰρ φιλόπονος γεωργὸς
τὰς μὲν πολλὰς ἀκάνθας ἐκτέμνει, τὰς δὲ ὀλίγας ἐᾷ, ἀλλὰ
κἂν δύο εὕρῃ, κἂν μίαν, πρόρριζον ἀνασπᾷ καὶ καθαρὸν
ἀποφαίνει τὸ λήϊον. πολλῷ δὴ οὖν μᾶλλον τοῦτο ποιητέον
ἡμῖν· οὐ γὰρ ἐκτέμνειν, ἀλλὰ μεταβάλλειν τὰς ἀκάνθας ὁ
8 τῆς ἡμετέρας γεωργίας παρακελεύεται νόμος. φέρε τοίνυν
καὶ ὡς ἀκάνθαις τὴν γεωργικὴν προσενέγκωμεν δίκελλαν

1 εὐκληρίας ἀπολαύειν: ἀπολαύσιν ὑγείας C | εὐκληρίας: σωτηρίας S, corr. Sγρ.: ὑγείας CV: εὐτυχίας Mγρ. | 2 τὸν B | 3 ἀποπηδῶσιν: ἀπαγορεύωσιν $S^1$: ἀπαγορεύουσιν $S^2$, corr. Sγρ. | 4 νόσῳ om. CV | 9 δὴ om. L | 10 τούτων γινομένων om. V | 11 κατεονῶσι K: καταιονοῦσι MV | 13 τῶν — ὅλῳ: ὑγείαν ἀνακαλέσασθαι CV, corr. Cγρ. | δὲ KBLMCV | 14 ἐπιμελιτέον S | 16 τῶν: τὸν C | 17 ὑλιστῆρος K: διυλιστῆρος S: διυλητῆρος C | διικνούμενοι CV | πόρον C | 21 οὐδὲ KSBγρ. | γεωργός: φυτουργὸς S | 25 ἐκτεμεῖν S | τὰς ἀκάνθας om. M, add. Mγρ. | 27 προενέγκωμεν $K^1$

Scanned pages often have dark borders around them. These can be erroneously picked up as extra characters, especially if they vary in shape and gradation.

## Transparency / Alpha channel

Some image formats (e.g. png) can have an alpha-channel for providing a transparency feature.

Tesseract 3.0x expects that users remove the alpha channel from the image before using the image in tesseract. This can be done e.g. with ImageMagick command:

```
convert input.png -alpha off output.png
```

Tesseract 4.00 removes the alpha channel with leptonica function pixRemoveAlpha(): it removes the alpha component by blending it with a white background. In some cases (e.g. OCR of movie subtitles) this can lead to problems, so users would need to remove the alpha channel (or pre-process the image by inverting image colors) by themselves.

## Tools / Libraries

- Leptonica
- OpenCV
- ScanTailor Advanced
- ImageMagick
- unpaper
- ImageJ
- Gimp
- PRLib - Pre-Recognize Library with algorithms for improving OCR quality

## Examples

If you need an example how to improve image quality programmatically, have a look at this examples:

- OpenCV - Rotation (Deskewing) - c++ example
- Fred's ImageMagick TEXTCLEANER - bash script for processing a scanned document of text to clean the text background.
- rotation_spacing.py - python script for automatic detection of rotation and line spacing of an image of text
- crop_morphology.py - Finding blocks of text in an image using Python, OpenCV and numpy
- Credit card OCR with OpenCV and Python
- noteshrink - python example how to clean up scans. Details in blog Compressing and enhancing hand-written notes.
- uproject text - python example how to recover perspective of image. Details in blog Unprojecting text with ellipses.
- page_dewarp - python example for Text page dewarping using a "cubic sheet" model. Details in blog Page dewarping.
- How to remove shadow from scanned images using OpenCV

## Page segmentation method

By default Tesseract expects a page of text when it segments an image. If you’re just seeking to OCR a small region, try a different segmentation mode, using the `--psm` argument. Note that adding a white border to text which is too tightly cropped may also help, see issue 398.

To see a complete list of supported page segmentation modes, use `tesseract -h`. Here’s the list as of 3.21:

```
  0    Orientation and script detection (OSD) only.
  1    Automatic page segmentation with OSD.
  2    Automatic page segmentation, but no OSD, or OCR.
  3    Fully automatic page segmentation, but no OSD. (Default)
  4    Assume a single column of text of variable sizes.
  5    Assume a single uniform block of vertically aligned text.
  6    Assume a single uniform block of text.
  7    Treat the image as a single text line.
  8    Treat the image as a single word.
  9    Treat the image as a single word in a circle.
 10    Treat the image as a single character.
 11    Sparse text. Find as much text as possible in no particular order.
 12    Sparse text with OSD.
 13    Raw line. Treat the image as a single text line,
                        bypassing hacks that are Tesseract-specific.
```

## Dictionaries, word lists, and patterns

By default Tesseract is optimized to recognize sentences of words. If you’re trying to recognize something else, like receipts, price lists, or codes, there are a few things you can do to improve the accuracy of your results, as well as double-checking that the appropriate segmentation method is selected.

Disabling the dictionaries Tesseract uses should increase recognition if most of your text isn’t dictionary words. They can be disabled by setting both of the configuration variables `load_system_dawg` and `load_freq_dawg` to `false`.

It is also possible to add words to the word list Tesseract uses to help recognition, or to add common character patterns, which can further help to improve accuracy if you have a good idea of the sort of input you expect. This is explained in more detail in the Tesseract manual.

If you know you will only encounter a subset of the characters available in the language, such as only digits, you can use the `tessedit_char_whitelist` configuration variable. See the FAQ for an example.

## Tables recognition

It is known tesseract has a problem to recognize text/data from tables (see issues tracker) without custom segmentation/layout analysis. You can try to use/test Sintun proposal or get some ideas from Text Extraction from a Table Image, using PyTesseract and OpenCV/code for Text-Extraction-Table-Image

## Still having problems?

If you’ve tried all the above and are still getting low accuracy results, ask on the forum for help, ideally posting an example image.

---

**tessdoc is maintained by tesseract-ocr.**
This page was generated by GitHub Pages.