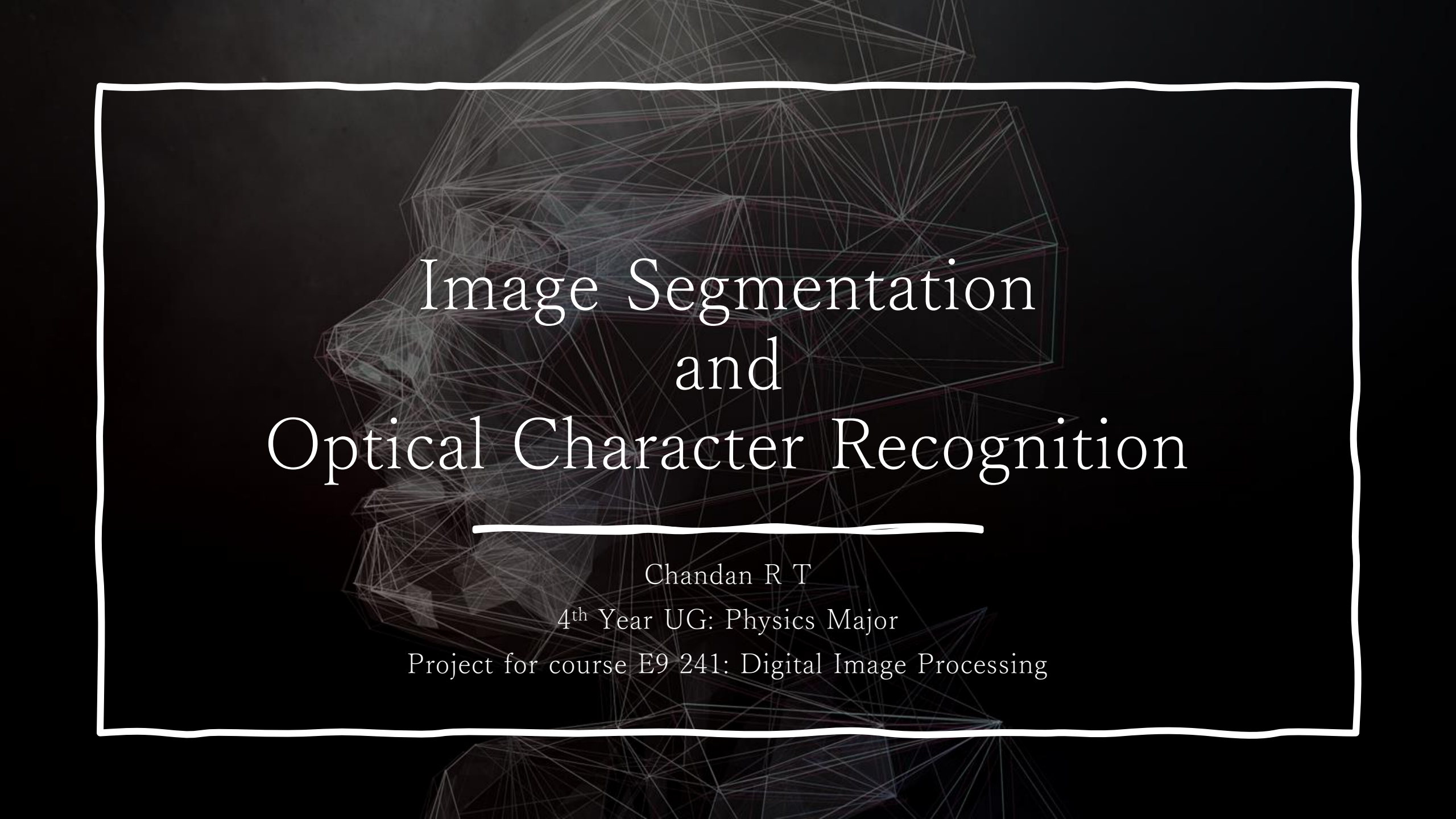
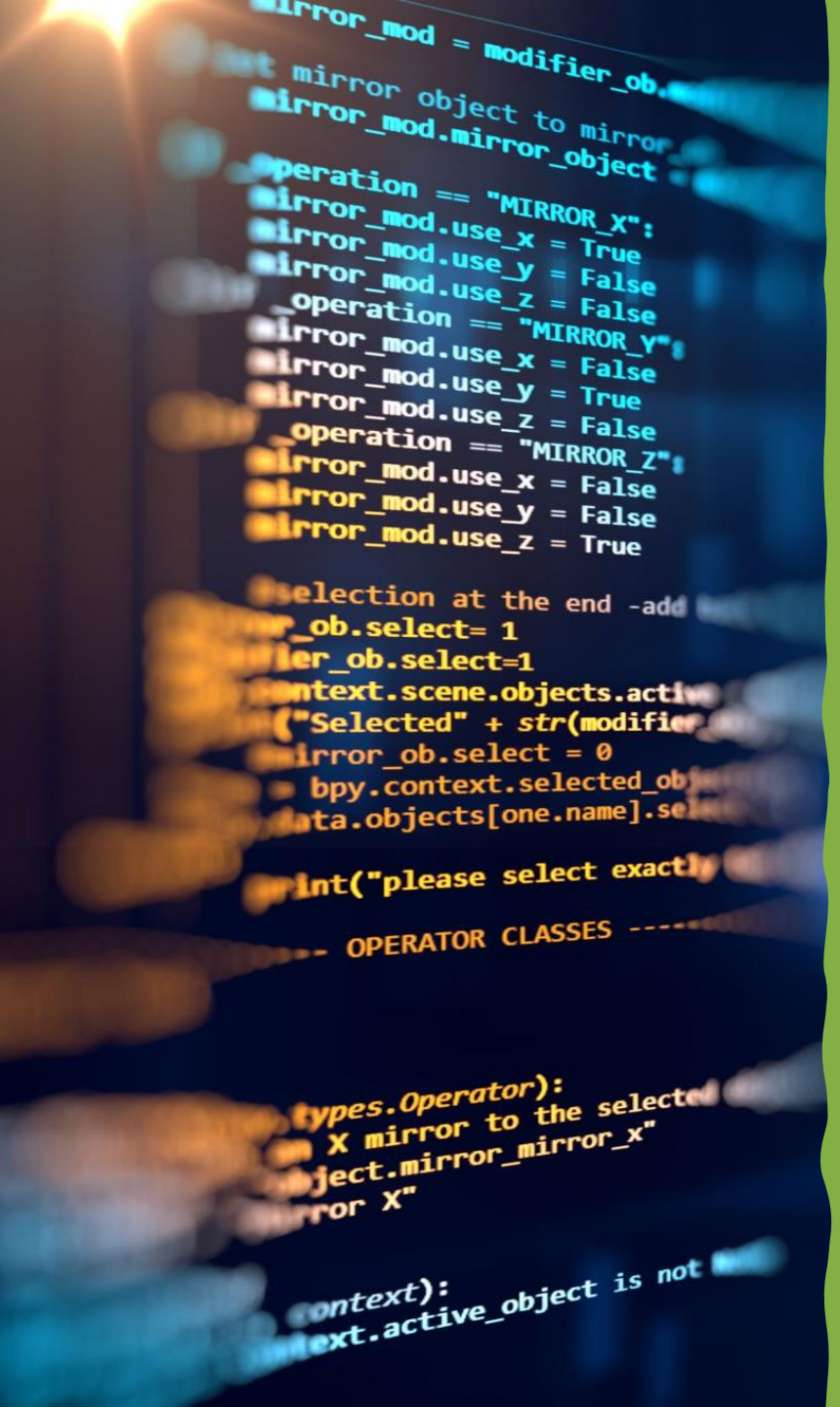


Image Segmentation and Optical Character Recognition

Chandan R T

4th Year UG: Physics Major

Project for course E9 241: Digital Image Processing



Problem

Given an image containing printed text, perform the required pre-processing and segmentation required, before utilizing an OCR model to identify characters.

This project has been extended to work with hand-written text too.

Pipeline

1. Image pre-processing
2. Image segmentation
3. Character pre-processing
4. Character recognition
5. Post-processing on output

1: Image pre-processing

1

Adaptive/local thresholding

2

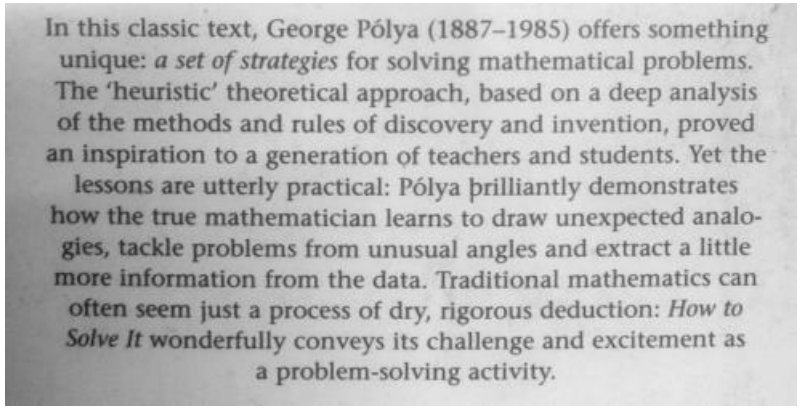
Denoising (optional) using Non-Local means

3

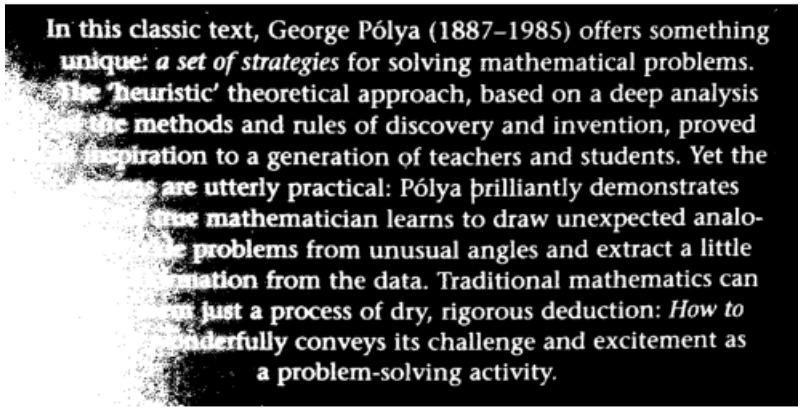
Skew correction (optional) using projection profile method

Adaptive Thresholding (demo)

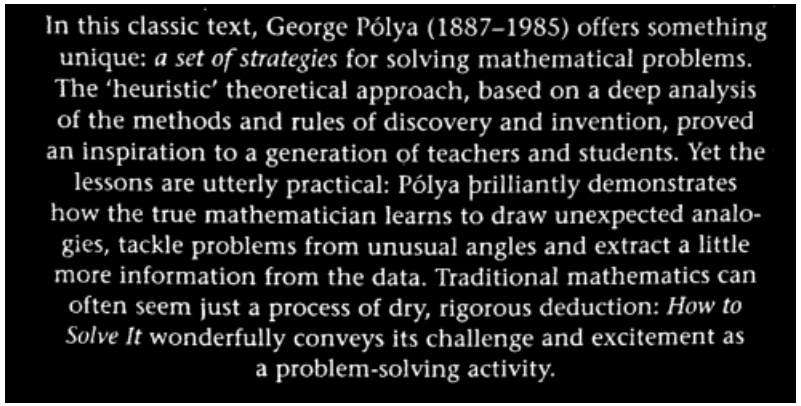
Original Image



Global Thresholding (Otsu)

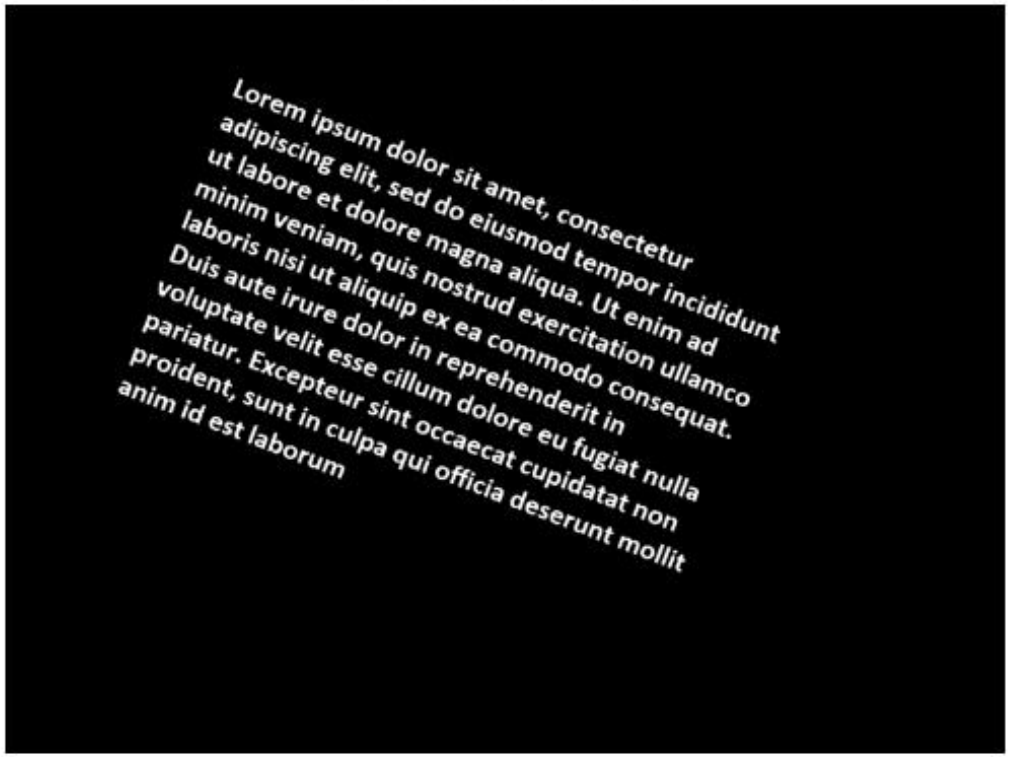


Adaptive Thresholding

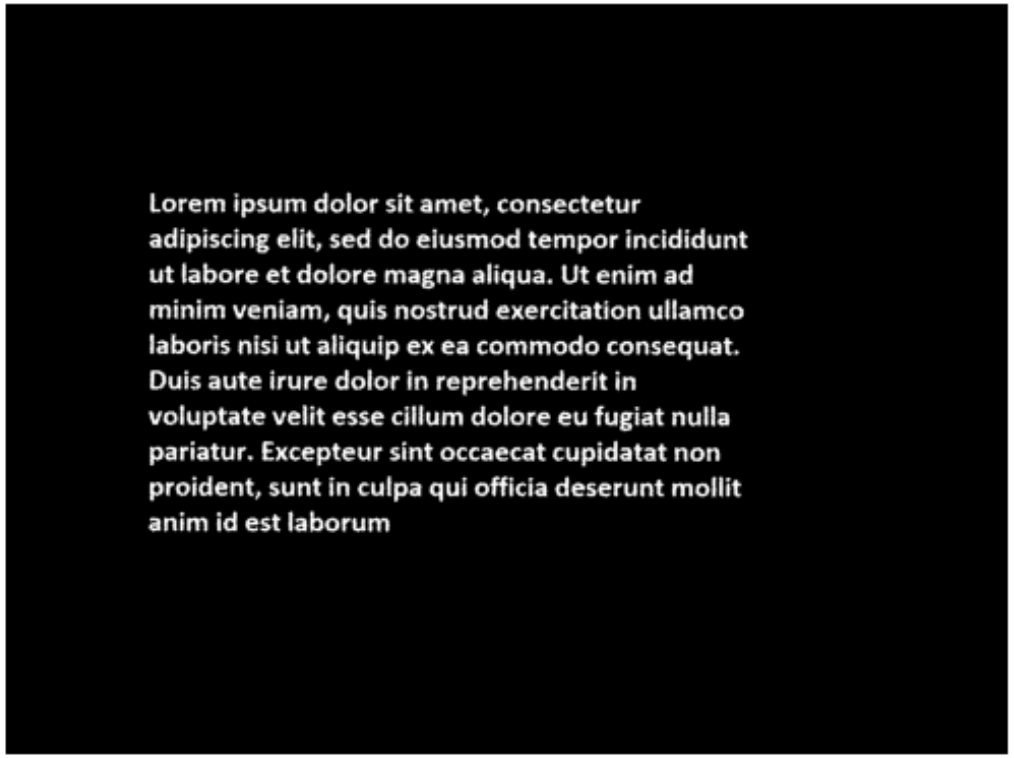


Skew Correction (demo)

Thresholded image



Skew corrected image



2: Image Segmentation

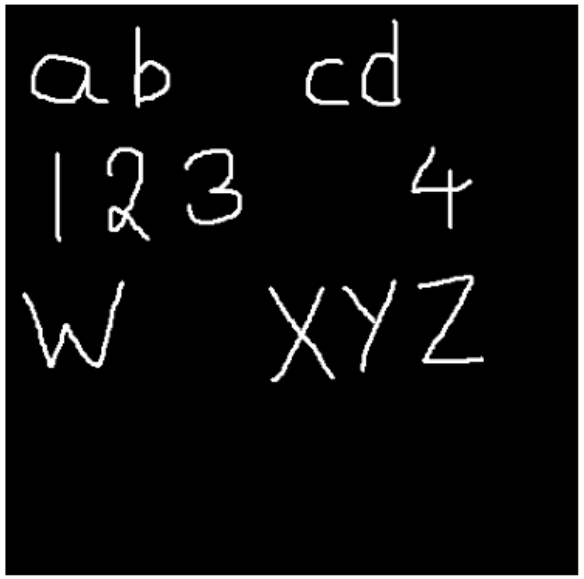
Line Segmentation is done by looking for local minima in row-wise histogram

Word Segmentation is achieved by looking at contiguous spaces in the column-wise histogram. I used k-means clustering to differentiate between the word spacing and the line spacing

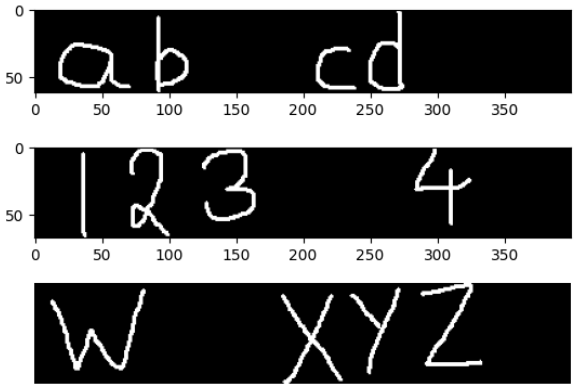
Character Segmentation is carried out via connected component analysis, since printed text involves separated characters

Image Segmentation (demo)

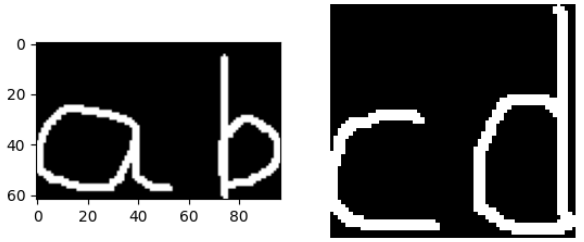
Thresholded image



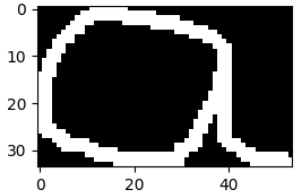
Line Segmentation

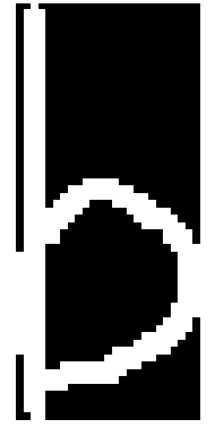


Word Segmentation

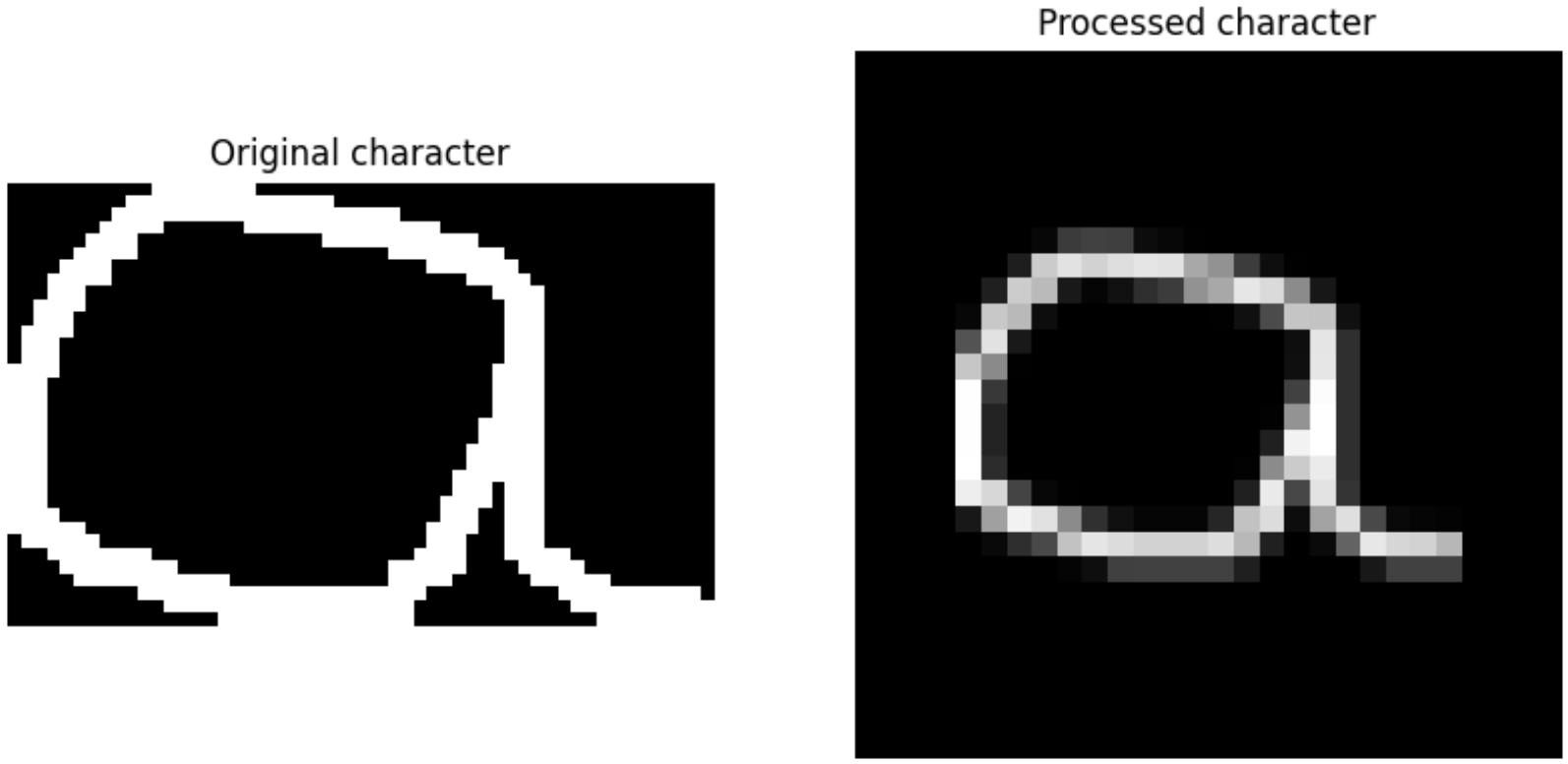


Character Segmentation





3: Character Pre-processing (demo)



Optionally: Dilation (in case of eroded or rotated images) or Skeletonization

4: Character Recognition

Two types of neural networks were considered:

- Dense/Artificial Neural Networks
- Convolutional Neural Networks

Two datasets were used for training:

- Handwritten (from extended MNIST) with ~7 lakh characters
- Printed characters (from Chars74K) with ~70k characters

Accuracy of Trained Models

Models	Accuracy on training data	Accuracy on testing data
ANN on hand-written characters	85.34 %	84.40 %
CNN on hand-written characters	87.97 %	86.10 %
ANN on printed characters	91.38 %	87.67 %
CNN on printed characters	94.55 %	90.18 %

Observations:

1) Better accuracy on printed characters
2) CNNs perform better than ANNs
3) Models trained on printed characters have a slight tendency to overfit

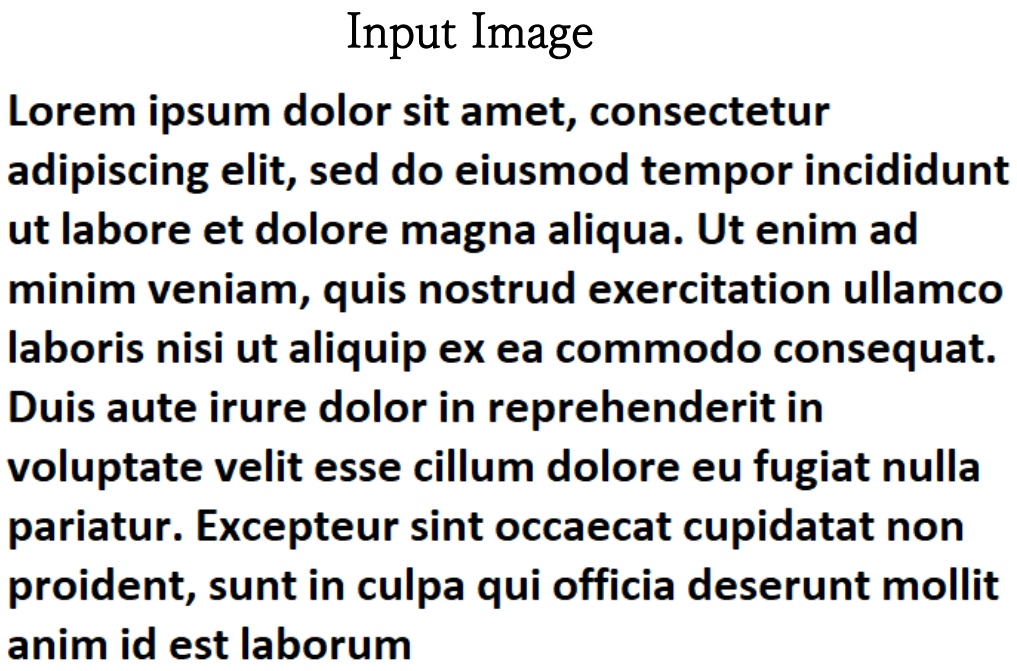
5: Post-processing

Number Correction (optional): It is very common for numbers to be recognized in place of letters (‘0’ for ‘o’, ‘1’ for ‘l’, ‘2’ for ‘z’, ‘5’ for ‘s’, ‘6’ for ‘b’, ‘9’ for ‘q’ or ‘g’, etc). In places where a number is unusual, post-processing will replace them with the most probable letter.

Spelling Correction (optional): should be used when text is in English. Sometimes involves multiple corrected spellings.

Number Correction (demo)

Input Image



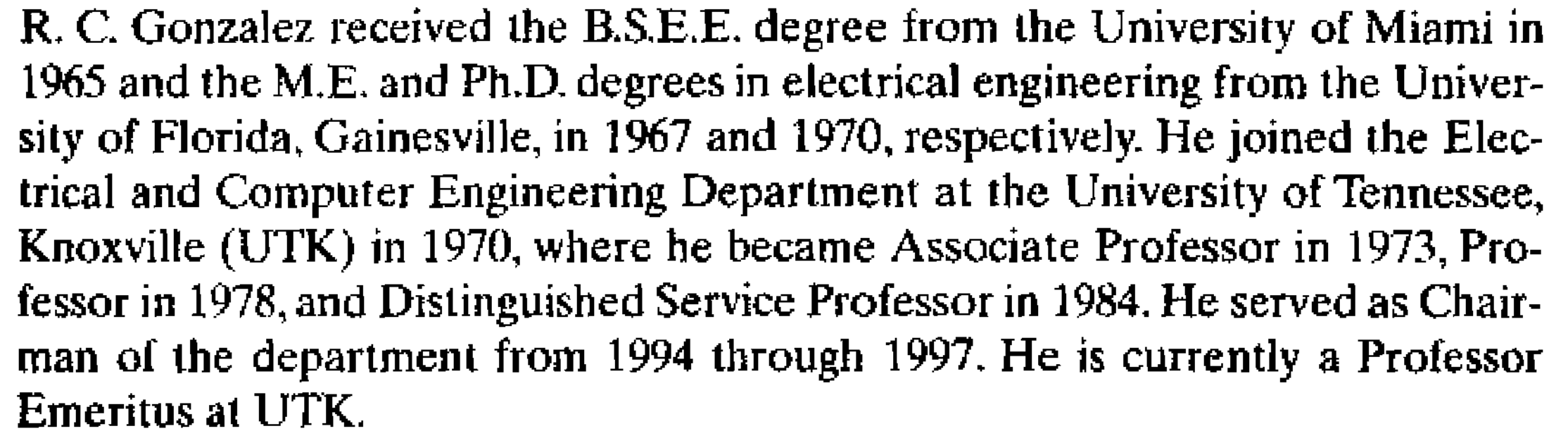
Raw output:
l0rem ipsum d010r sit amet c0nsectetur adipiscin8 e1it sed d0 eiusm0d temp0r incididunt ut 1ab0re et d010re ma8na a1i9ua ut enim ad minim veniam ....

After number correction:
lorem ipsum doior sit amet consectetur adipiscinb eiit sed do eiusmod tempor incididunt ut iabore et doiore mabna aiiqua ut enim ad minim veniam ....

In order to deliberately bring out the errors required for demonstrating this feature,
The error-prone handwritten ANN model was used for character prediction.

Spelling Correction (demo)

Input Image

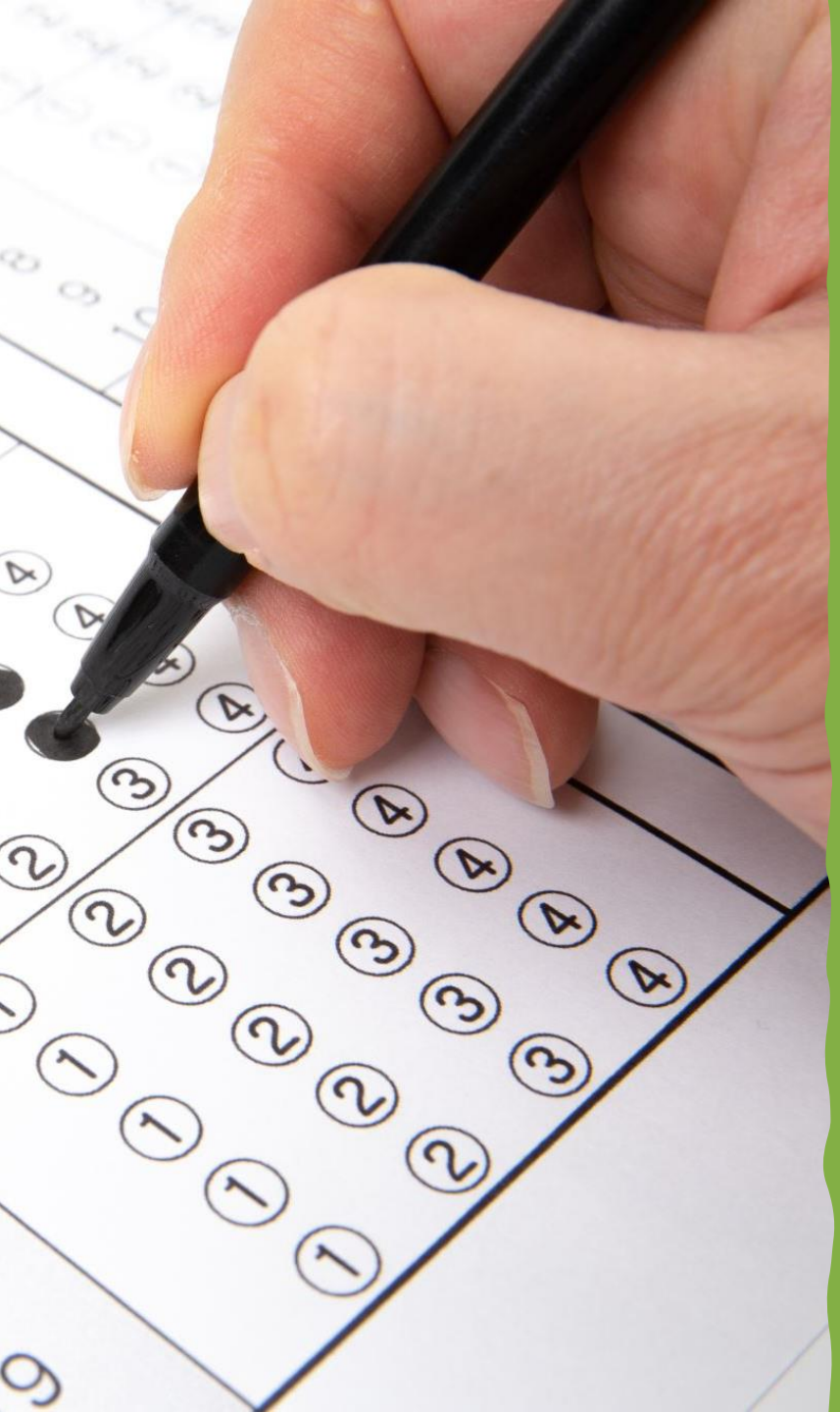


R. C. Gonzalez received the B.S.E.E. degree from the University of Miami in 1965 and the M.E. and Ph.D. degrees in electrical engineering from the University of Florida, Gainesville, in 1967 and 1970, respectively. He joined the Electrical and Computer Engineering Department at the University of Tennessee, Knoxville (UTK) in 1970, where he became Associate Professor in 1973, Professor in 1978, and Distinguished Service Professor in 1984. He served as Chairman of the department from 1994 through 1997. He is currently a Professor Emeritus at UTK.

Spelling Correction (demo)

r c gqnzale7 lecwived lhe bsee degree frqm lhe univebity of miami in
1mi and thc me 3nd phd degrees dn clectrica1 engineefing fr0m lhe univer
sity of flotida gaine8vij1e in 1s7 and 197q respectivejy he joined lhe elcc
ttica1 and computer engineeane ncparlment at the univc5sity oftdnnessee
knoxvi1le kfky in 197fj wheie he beczme asswciate prqfexsor in 1973 pro
fdkoi in l978 3nd bizlinguished service professor in 1984 hg semed as chair
man of the deparlmenl from 1994 thtough 1997 he ie cutrent1y a ptotescoi
emetitvs at utk

r c gunwales deceived/received lhe byee/blee/bree/usee degree fram/frim/from lhe univocity of miami in
lmi and thc me and phd degrees dn electrical/electrican engineering from lhe uniter
sity of florida baskerville/evanescible/garnishable in qsf and ljzq respectively he joined lhe elec
ttical and computer engineered/enginelike appallment/department/impartment at the university patronesses
knoxville kfky in lqffj whein/where he became/become associate professor in 1973 pro
askoi/kikoi in lqfg hnd distinguished service professor in 1984 hg semed as chair
man of the department from 1994 through 1997 he ie currently a protestor
emeritus/emetines at utk



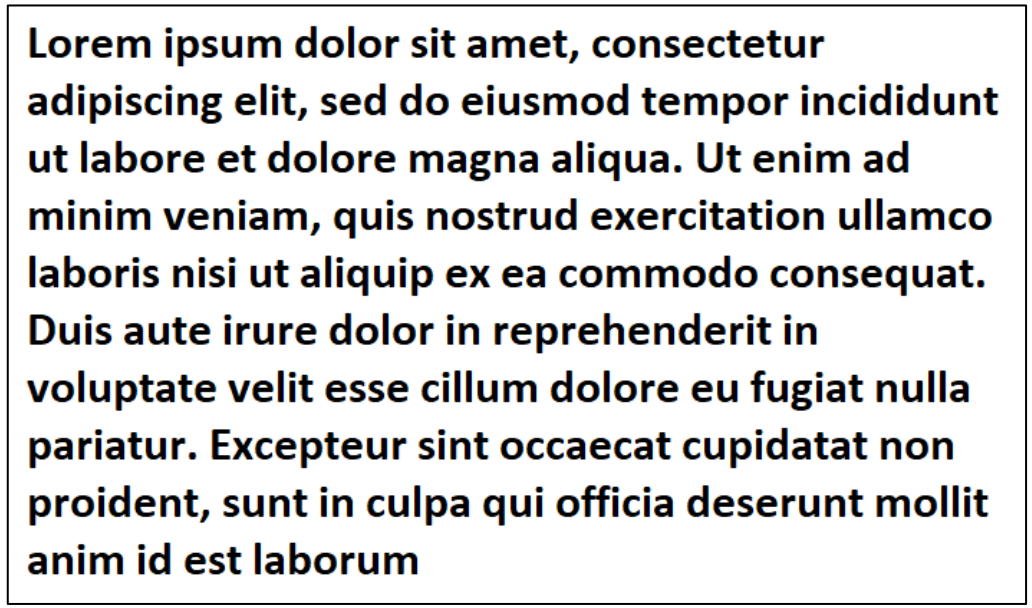
Evaluation Metrics

- Levenshtein distance is a commonly used metric for string similarity, it is based on the number of single-character edits required to transform one string into another.

- Overall error:

$$error\ \% = \frac{Levenshtein\ distance}{length\ of\ expected\ string} * 100$$

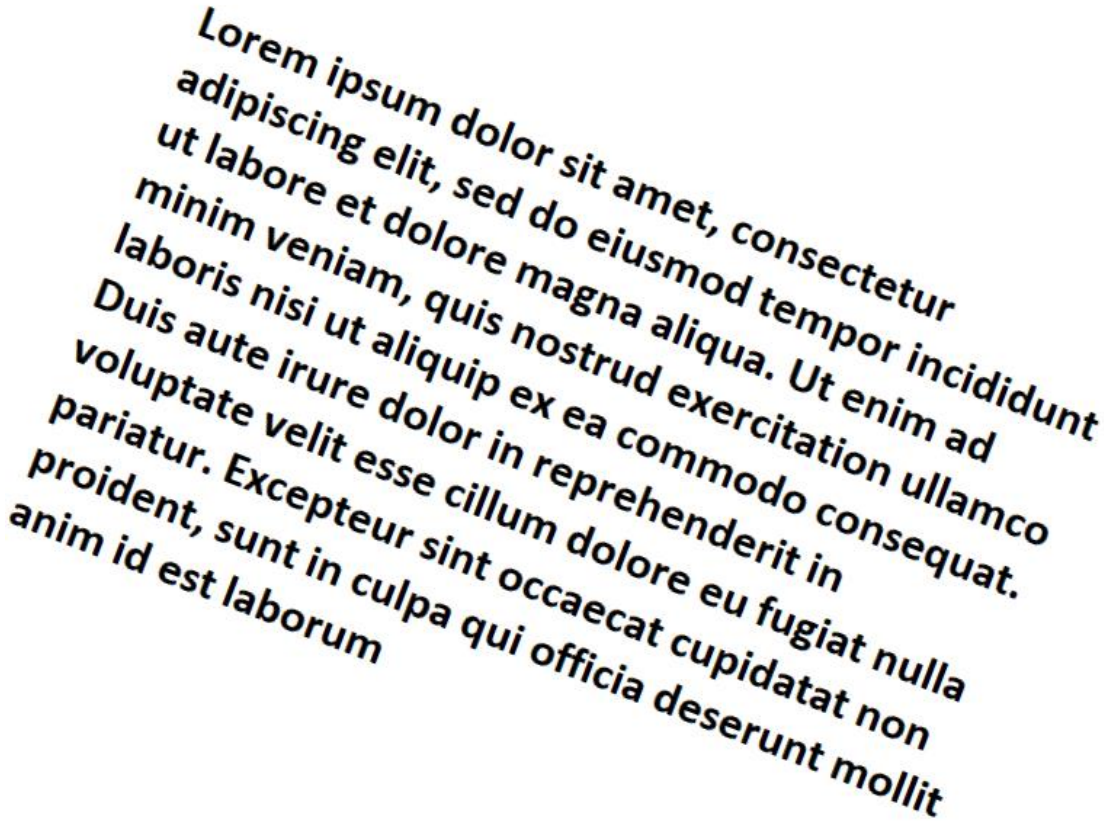
Evaluation 1: Synthetic image



Program output:
lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut enim ad minim veniam quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur excepteur sint occaecat cupidatat non proident sunt in culpa qui ohicia deserunt mollit anim id est laborum

Levenshtein distance: 3
Error: 0.69 %

Evaluation 2: Rotated image



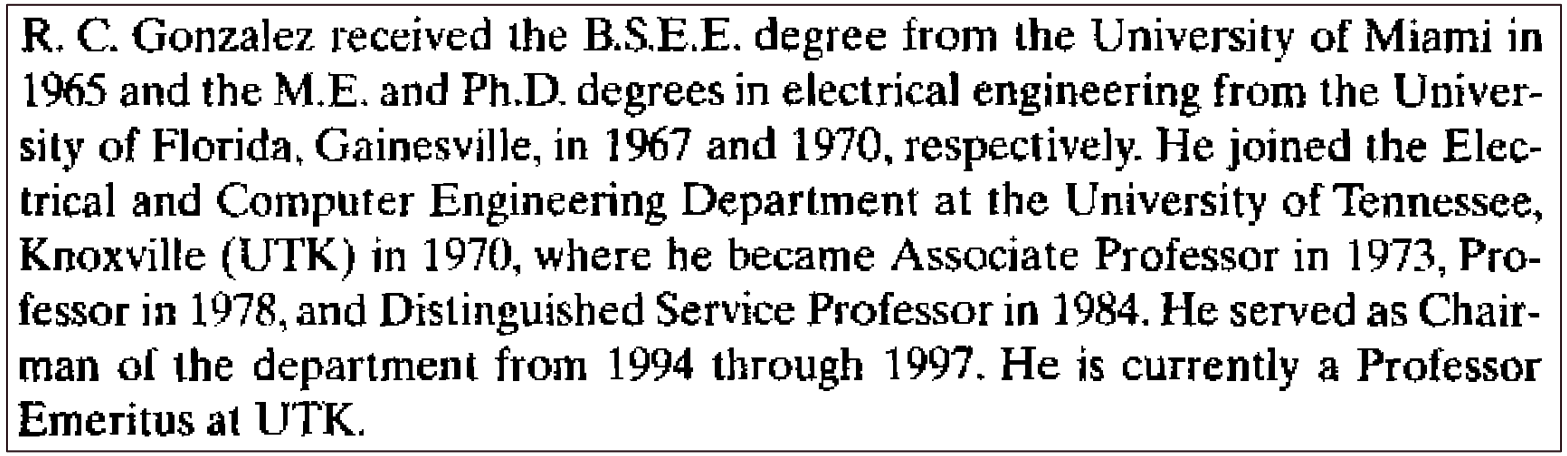
Program output:

lorqm ipsum dolor sdt amet consectetur adipiscinc elik sed do eiusmod tempor incididunt ut lahore et dojore macna ajiqua ut enim ad mdnsm vendam qubs aostrud exercitation ujiamco faboris absd ut ajiquip ex ea commodo consqquat ouis autr drurq doior in tepreheaderit ia voluptakq veiit qsse cifium dofore eu fugiat nujfa patdatur exceptqur xiat occaecat cupidatat aon proideak suak da cufpa qus okicia deserunk moltit aaim id est laborum

Levenshtein distance: 58

Error: 13.27 %

Evaluation 3: Eroded pixels

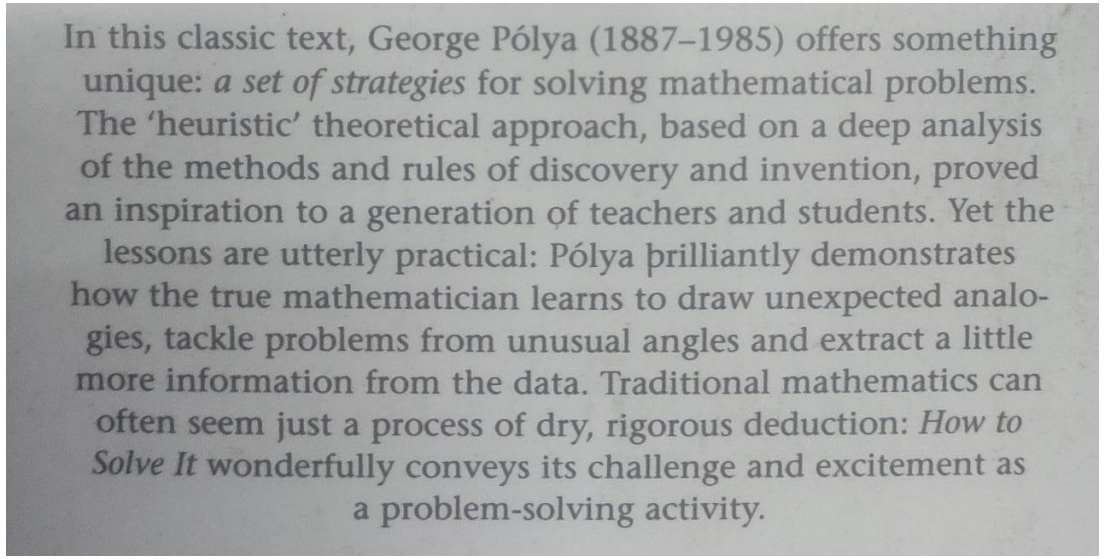

R. C. Gonzalez received the B.S.E.E. degree from the University of Miami in 1965 and the M.E. and Ph.D. degrees in electrical engineering from the University of Florida, Gainesville, in 1967 and 1970, respectively. He joined the Electrical and Computer Engineering Department at the University of Tennessee, Knoxville (UTK) in 1970, where he became Associate Professor in 1973, Professor in 1978, and Distinguished Service Professor in 1984. He served as Chairman of the department from 1994 through 1997. He is currently a Professor Emeritus at UTK.

Levenshtein distance: 84
Error: 15.82 %

Program output:

r c gonzalez leceived lhe bsee deeree frqm lhe unsvcasty of miami sn
lms hnd thc me and phd degrees in electrical engineefing from che unsyct
sity of blotsda gainesvijle in lhi and lqzo respectivejy he joined lhe elgc
crical and computer eneineeanc dcpartment at the univefsity oftdnncssce
knoxville cwkj in lqfcj whete he beczme assqciate professor in 1973 pro
fkmot sn lqze and dislinguished servsce professor in 1984 hc sewed as chair
man of lhe departmenc from 1994 thtough 1997 he is cuttently a professot
emetitus al utk

Evaluation 4:
A real-life image

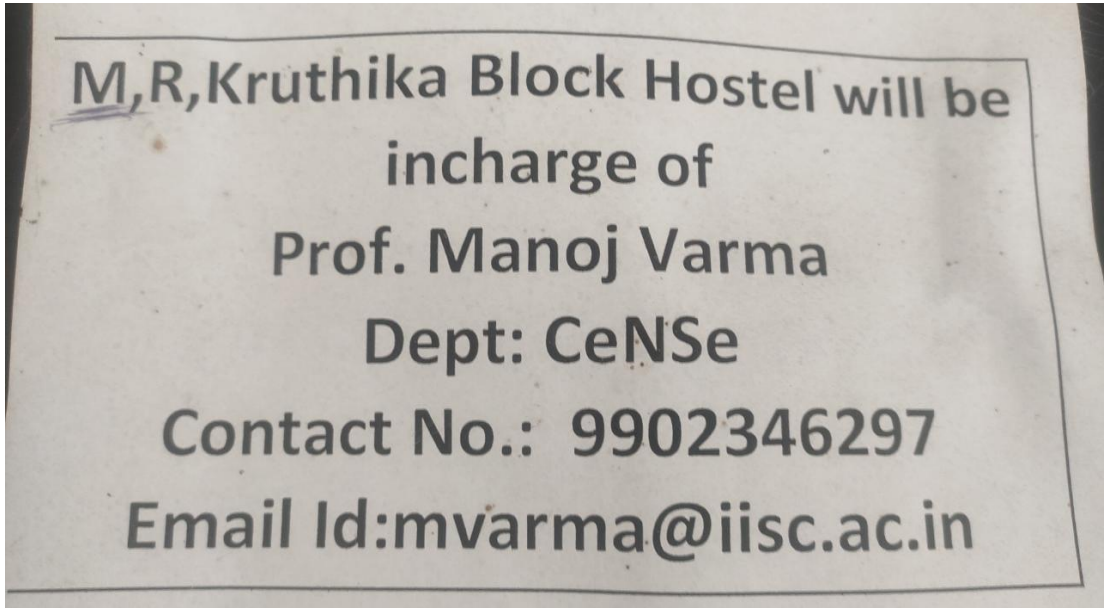

In this classic text, George Pólya (1887–1985) offers something unique: *a set of strategies* for solving mathematical problems. The 'heuristic' theoretical approach, based on a deep analysis of the methods and rules of discovery and invention, proved an inspiration to a generation of teachers and students. Yet the lessons are utterly practical: Pólya þrilliantly demonstrates how the true mathematician learns to draw unexpected analogies, tackle problems from unusual angles and extract a little more information from the data. Traditional mathematics can often seem just a process of dry, rigorous deduction: *How to Solve It* wonderfully conveys its challenge and excitement as a problem-solving activity.

Program output:
jn this classic text george pdlya k18871985j offers something unique a set ofstratelies for solving mathelnatical problems the heutistic theotetical approach based on a deep analysis of the methods and mles of discovery and invention ptoved an inspitation to a venetation of teachers and students yet the lessons are uttetfy practical pdlya vtilliantly demonsttates how the true mathematician learns to dtaw unexpected analo gies tackle ptoblems ftom unusual angles and extiact a little moie infotmation from the
data tiaditional mathematics can often seem just a piocess of dcy rigotous deduction hobv to solve it wonderfuliy conveys its challenge and excitement as a problemsoling activity

Levenshtein distance: 40
Error: 5.78 %

(This output can benefit a lot from spelling correction. However, for the purpose of evaluation, this feature has been turned off)

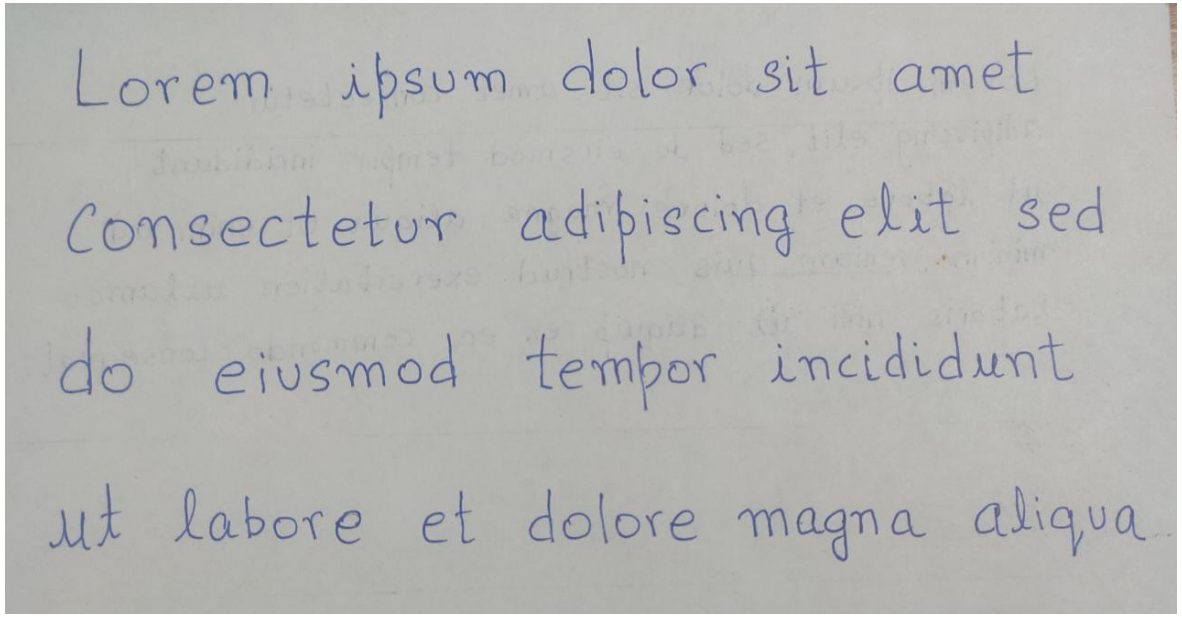
Evaluation 5:
Noise, Background and Lines



Program output:
i mrlkruthilka bloclk hkmiddk l
l o inchargeof o ll
j pfmsvly
l m deptcense l
l cttn9902346297 l
emailldmvdrmaqiiscacdn l

Levenshtein distance: 58
Error: 47.93 %

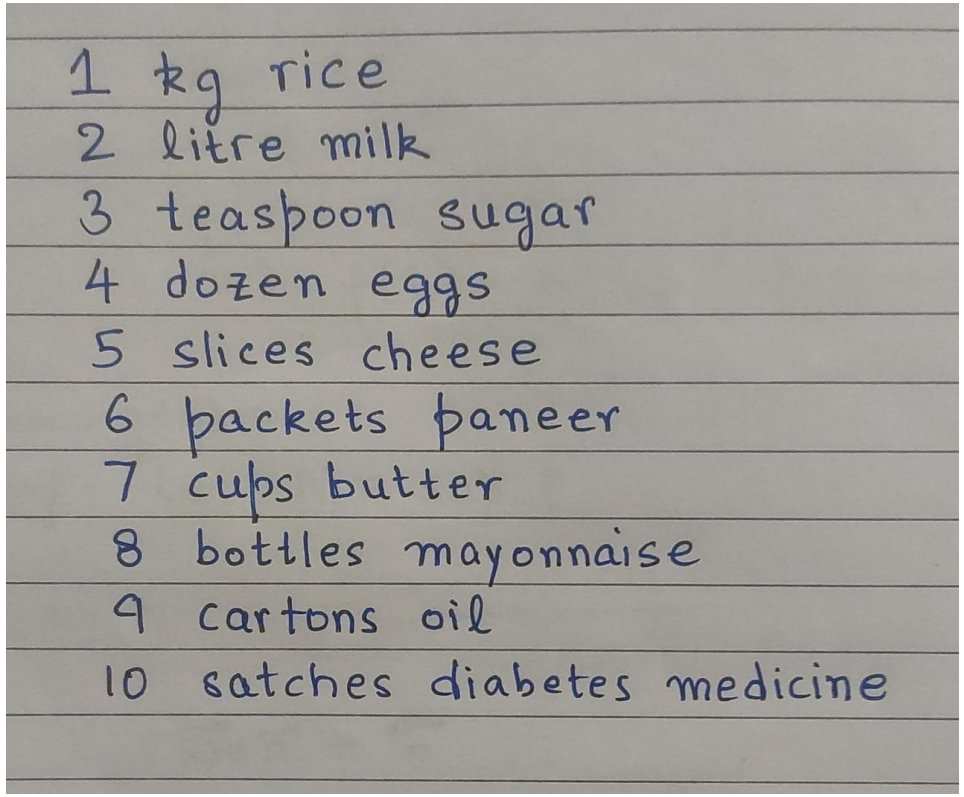
Evaluation 6: Hand-written note



Program output:
lotem hisum doirsit amet
consectetur adliisclng elht sed
d0 eiusmod temror incididxnt
d1 labote et d010re magna aliqua

Levenshtein distance: 21
Error: 17.50 %

Evaluation 7: Grocery List



Program output:

1 h9 tice
2 ditre milk
3 teasbon suhaf
4 dozen eqgs
5 siices cheese
6 hackets hner
1 culos butter
8 bttles mayse
q cartons oil
10 satches diabetes medicine

Levenshtein distance: 27
Error: 19.62 %

Shortcomings

Extensive tweaking of various parameters was carried out in order to minimize error during evaluation. Hence, generalization is minimal.

Problems like rotated images, eroded pixels, horizontal and vertical lines, etc, should be dealt with in an ad-hoc manner by the user. The program must be able to detect these problems automatically.

Future Improvements

- This program can be extended to work on hand-written text in cursive, via skeletonization and segmentation around ligatures.
- Instead of dealing with problems on an ad-hoc basis, if the training data itself contains these problematic artifacts, then the model will learn to detect characters regardless of any problems. This will also increase generalization and reduce the need for fine tuning of parameters.

Citations

- Hand-written characters dataset: Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). EMNIST: an extension of MNIST to handwritten letters. https://www.nist.gov/itl/products-and-services/emnist-dataset

- Printed characters dataset: T. E. de Campos, B. R. Babu and M. Varma (2009). Character recognition in natural images. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/

Thank You

Follow this project's development on GitHub!