22/1/2017

Need to try nearest centroid classsifier

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=583D87DDC7C65F3861D143294033140D?doi=10.1.1.53.1450&rep=rep1&type=pdf for handwritten script identification

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1261096 for online handwritten script recognition

Knn - http://www.kdnuggets.com/2016/01/implementing-your-own-knn-using-python.html/3

LBP features increase DONE

24/9/2016

Handwritten ocr:

1) Obtaining handwritten documents
    a) Class notes
    b) Letters (applications, etc)
2) Preprocessing
    a) Removing irregularities through photoshop and few through code
    b) Postal markings, doodles, foreign characters, sideway writings etc.,
3) Method
    a) Extract connected components
    b) Filtering
        i) Unusually small or large
        ii) Small(EX : height of bounding box < 3 or width of bounding box < 3 or area of bounding box < 30)
    c) Features extraction
        i) Relative Y centroid, Relative X centroid, Number of white holes, Sphericity, Aspect Ratio is calculated for all connected components.
        ii) Find mean, standard deviation and skew for all this five connected component features. 15 element vector is formed
    d) Language identification
        i) Fischer linear discriminant analysis is done in paper.
        ii) We can try other classifiers like neural networks etc.

Tried centroid classifier but as there may be centroid deviation due to some of language images, we are not getting accuracy at all. So we need to try k means classifier and change k for better accuracy.

25/9/2016

Trying knn classifier with varying k

K	wrong
1	1
3	2
5	2
7	2
9	2

Need to increase data for document images.
Mean time we will concentrate on handwritten images and will apply this methods for them and increase this data and handwritten methods for this finally. Will make the poster then.

28/1/2017

Not useful Research paper on datasets-
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6424486

IPC dataset for english, urdu

http://www.visionias.in/beta/blog/toppers-answer-booklet IAS notes for english
http://ajayvision.com/beta/sites/all/themes/momentum/files/toppersanswerbooklet/170_Premsukh%20Delu_15301_634_GS.pdf IAS notes for hindi

Telugu notes using crawler:
http://appscgroup.blogspot.in/2014/04/andhra-history-for-appsc-group-1-and-group-2-exams-material-download-part1.html

Hindi https://drive.google.com/drive/folders/0BzwIRRCeTIopZGUxMTBrajJWa2s

30/1/2017

Finalised datasets

English - IAS dataset - http://www.visionias.in/beta/blog/toppers-answer-booklet from different writers
Hindi - IAS dataset, Geography dataset -
http://www.visionias.in/beta/blog/toppers-answer-booklet
https://drive.google.com/drive/folders/0BzwIRRCeTIopZGUxMTBrajJWa2s
Telugu -
http://appscgroup.blogspot.in/2014/04/andhra-history-for-appsc-group-1-and-group-2-exams-material-download-part1.html - search fot TM

31/1/2017

---

https://chrome.google.com/webstore/detail/save-all-images/jmolegopjlipmoedaoijpjaddhjjckal
extension for saving images

Paper we are implementing :
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=583D87DDC7C65F3861D143294033140D?doi=10.1.1.53.1450&rep=rep1&type=pdf

2/2/2017

Watermark removed

Need to manually process data

4/2/2017

Sheet link
https://docs.google.com/spreadsheets/d/1sv58-mxujP7jS_hqhGzg_Kt1dpFr1vKtYE9BdnEjijQ/edit

Papers:
Connected components:
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=583D87DDC7C65F3861D143294033140D?doi=10.1.1.53.1450&rep=rep1&type=pdf

Survey:
http://visgraph.cs.ust.hk/biometrics/Papers/Signature/pami2000-01-01.pdf

Gabor filter:
http://ceur-ws.org/Vol-758/paper_12.pdf

3 papers going to implement:
Word level multi - script identification - Gabor filter
http://mile.ee.iisc.ernet.in/mile/publications/softCopy/DocumentAnalysis/ScriptRecog_PRL.pdf
Connected components:
http://download.springer.com/static/pdf/98/art%253A10.1007%252Fs100320050036.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1007%2Fs100320050036&token2=exp=1486477687~acl=%2Fstatic%2Fpdf%2F98%2Fart%25253A10.1007%25252Fs100320050036.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Farticle%252F10.1007%252Fs100320050036*~hmac=7b24f097a88286c8bd432075ef9f9c92f0306f20b71fd7b17f4a8cbccf70d476

Handwritten Script Recognition using DCT and Wavelet Features at Block Level:
https://www.researchgate.net/profile/Ganapatsingh_Rajput/publication/45718429_Handwritten_Script_Recognition_using_DCT_and_Wavelet_Features_at_Block_Level/links/00b49531adb4def700000000.pdf

5/2/2017

First implementing:
https://www.researchgate.net/profile/Ganapatsingh_Rajput/publication/45718429_Handwritten_Script_Recognition_using_DCT_and_Wavelet_Features_at_Block_Level/links/00b49531adb4def700000000.pdf
Steps:

Obtaining handwritten documents DONE
Preprocessing DONE
Extract connected components
Filtering
Features extraction
Language identification

https://in.mathworks.com/help/images/ref/dct2.html
https://in.mathworks.com/help/wavelet/ref/dwt2.html
https://in.mathworks.com/help/images/ref/imopen.html
https://in.mathworks.com/help/images/ref/imdilate.html

7/2/2017

---

https://in.mathworks.com/help/stats/knnsearch.html
https://in.mathworks.com/help/bioinfo/ref/knnclassify.html

Very worst accuracy with KNN about 50%
So trying svm about 98.5

DONE implementing 1st paper

Gabor filter
https://in.mathworks.com/matlabcentral/answers/104759-apply-gabor-filter-to-an-input-image-using-matlab
https://in.mathworks.com/help/images/ref/imgaborfilt.html

Will implement this
http://download.springer.com/static/pdf/98/art%253A10.1007%252Fs100320050036.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1007%2Fs100320050036&token2=exp=1486477687~acl=%2Fstatic%2Fpdf%2F98%2Fart%25253A10.1007%25252Fs100320050036.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Farticle%252F10.1007%252Fs100320050036*~hmac=7b24f097a88286c8bd432075ef9f9c92f0306f20b71fd7b17f4a8cbccf70d476

12/2/2017

---

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6628641   good paper will implement
https://sites.google.com/site/aacruzr/conferences good conferences

Got an accuracy of 97.3856 % accuracy with new features for cleaned images
Got an accuracy of 88.8889 % accuracy with new features with lined images

GABOR Filter paper:
http://mile.ee.iisc.ernet.in/mile/publications/softCopy/DocumentAnalysis/ScriptRecog_PRL.pdf

Gabor features method :-
0.125,0.25,0.5 radial frequencies
0,30,60,90,120,150 orientations
Sine and cosine filters
We get 36 filtered images
Find energy of this image(sum squared) and divided with original image energy)
These are our 36 features.
Apply any classifier(knn,svm,linear discriminant analysis)

Implementing gabor paper
Got an accuracy of 96.7320% accuracy with new features for cleaned images
Got an accuracy of 96.0784% accuracy with new features with lined images

Connected components paper:
http://download.springer.com/static/pdf/98/art%253A10.1007%252Fs100320050036.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Farticle%2F10.1007%2Fs100320050036&token2=exp=1486907706~acl=%2Fstatic%2Fpdf%2F98%2Fart%25253A10.1007%25252Fs100320050036.pdf%3ForiginUrl%3Dhttp%253A%252F%252Flink.springer.com%252Farticle%252F10.1007%252Fs100320050036*~hmac=3eeb12d817328cced168894c88f55527e95f5cc87b90f04f854ee17befb4cd60

Method:

4) Extract connected components
5) Filtering
    a) Unusually small or large
    b) Small(EX : height of bounding box < 3 or width of bounding box < 3 or area of bounding box < 30)
6) Features extraction
    a) Relative Y centroid, Relative X centroid, Number of white holes, Sphericity, Aspect Ratio is calculated for all connected components.
    b) Find mean, standard deviation and skew for all this five connected component features. 15 element vector is formed
7) Language identification
    a) Fischer linear discriminant analysis is done in paper.
    b) We can try other classifiers like neural networks etc.

https://wikidev.in/wiki/matlab/image_processing/imgaborfilt

https://lost-contact.mit.edu/afs/cs.stanford.edu/pkg/matlab-r2015b/matlab/r2015b/toolbox/images/images/

https://in.mathworks.com/help/stats/skewness.html

Implementing connected components paper
Got an accuracy of 88.889% accuracy with new features for cleaned images
Got an accuracy of 91.5033% accuracy with new features with lined images

If we have time will implement stroke paper and also increase dataset.

12/3/2017

Remove horizontal lines DONE
Google images download and test and make data

15/3/2017

Running on google images

With connected components:
Training 3 writers and testing google images 51.2% accuracy
Training 20 writers and testing google images 78.8%accuracy
Training 25 writers and testing google images 89.4%accuracy

With dct components:
Training 3 writers and testing google images 58.33% accuracy
Training 20 writers and testing google images 71.4%accuracy
Training 25 writers and testing google images 92.71%accuracy

With gaborcomponents:
Training 3 writers and testing google images 47.36% accuracy
Training 20 writers and testing google images 65.78%accuracy
Training 25 writers and testing google images 82.04%accuracy