Empirical Evaluation of Character Classification Schemes

Neeba N.V and C.V Jawahar

Centre for Visual Information Technology,
International Institute of
Information Technology, Hyderabad, India - 500 032

Motivation

- Are the state of the art classifiers suitable/ sufficient to solve Large Class problems ?
  - Most of the classifiers designed for smaller number of classes.
  - But a large number of real world problems are large class (in the order of hundreds) in nature.
- Will the character classification problem for Indian languages be solved successfully ?
  - Large number of classes.
  - Unavailability of bench-mark datasets.

State of the Art

- STATLOG was considered to be the most comprehensive empirical comparative study for pattern classifier 12 years back.
- A recent study focusing on empirical comparison of recent approaches presented by Carunana.
- Lecun et.al reported a comparative study of various Convolutional Neural Network architectures and other classifiers for handwritten digit recognition.

R. King, C. Feng, and A. Shutherland, .Statlog: comparison of classication algorithms on large real-world problems,.*AAI, vol. 9, pp. 259.287, June 1995.*

Indian Language Character Recognition

- **Bangla and Devanagiri OCR :** B.B. Chaudhuri, U. Pal, “An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi) “, IEEE Computer Society, 1997.
- **Telugu OCR :** A. Negi, C.Bhagavathi and B. Krishna, "An OCR System for Telugu", ICDAR-2001.
- **Kannada OCR :** T.V Ashwin and P.S Sastry, “A font and size independent OCR System for Printed Kannada documents using Support Vector Machines.” Sadhana Vol.27, 36-58.
- **A survey on Indian language character recognition :** is presented by B.B. chaudhuri. [U. Pal and B. B. Chaudhuri, .Indian script character recognition: a survey,. *Pattern Recognition, vol. 37, no. 9, 2004.*]

Focus of the Study

- **Experiment 1** : Comparison of classifiers and features.
- **Experiment 2** : Scalability of classifiers.
- **Experiment 3** : Richness in the feature space.
- **Experiment 4** : Sensitivity of features to degradation.
- **Experiment 5** : Generalization across fonts.
- **Experiment 6** : Applicability across scripts.

Classifiers

Classifiers Used

- Multi-layer Perceptron (MLP).
- Convolutional Neural Networks(CNN).
- K-Nearest Neighbour (KNN).
- Approximate Nearest Neighbour(ANN).
- SVM-Majority Voting (SVM-1).
- SVM-DDAG (SVM-2).
- Naive Bayes(NB).
- Decision Tree Classifier (DTC).

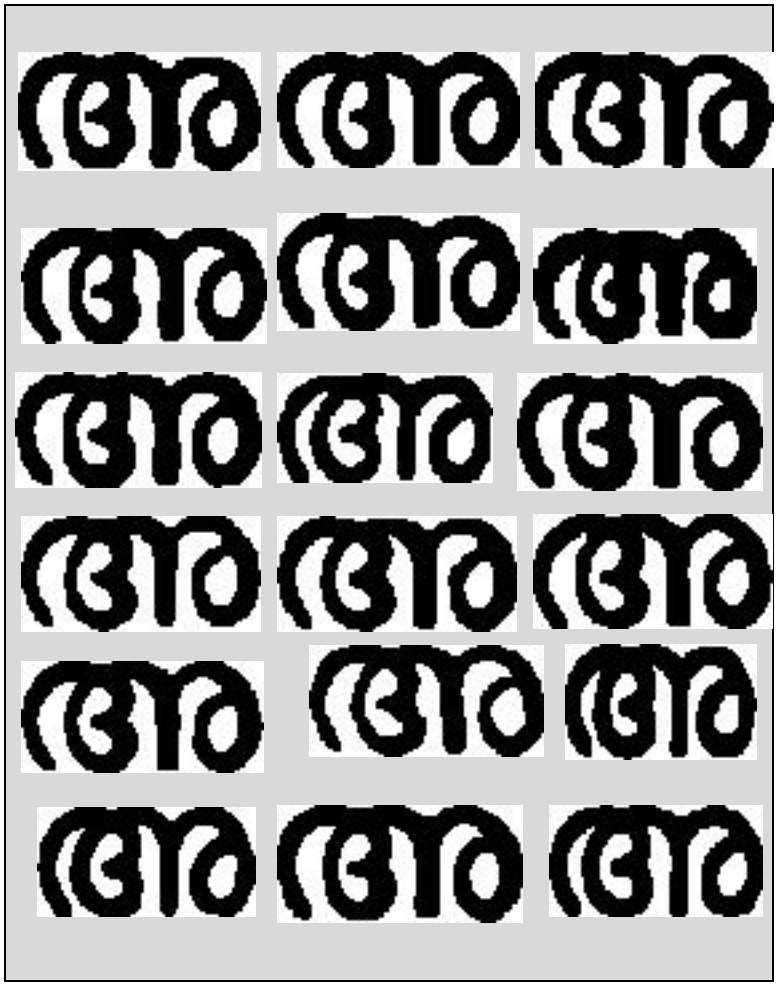
Features Used

- Central Moment (CM).
- Zernike Moment (ZM).
- Discrete Cosine Transform(DCT).
- Discrete Fourier Transform(DFT).
- Principal Component analysis(PCA).
- Linear Discriminant Analysis(LDA).
- Random Projections (RP).
- Distance Transform (DT).
- Raw Image (IMG).

Dataset Used

Dataset

- From annotated books printed primarily in Malayalam.
- 5,00,000 real characters (Symbols) from 5 Books.
- Other scripts used for the experiments are : *Telugu and English.*
- Dataset Generation.



C. V. Jawahar and A. Kumar, .Content-level annotation of large collection of printed document images,. in *ICDAR*, pp. 799.803, 2007.

Experiment: 1

Comparison of Classifiers and Features:

Experimental Settings

- The focus of the study is to find out the set of classifiers and features that can be used to solve the problem successfully.
- Parameters of classifiers :-
  - MLP – no. of nodes in the hidden layer: 60, momentum: 0.6, no. of epochs: 30.
  - SVM –with linear kernel.
  - KNN and ANN – with K = 5.
- Scale size used : 20 X 20.
- Train : Test Ratio => 5:95

Experiment: 1

Comparison of Classifiers and Features:

Results

Feature	Dim	Classifiers						
		MLP	KNN	ANN	SVM-1	SVM-2	NB	DTC
C.M	20	12.04	4.16	5.86	10.04	9.19	11.93	5.57
DFT	16	8.35	8.96	9.35	7.88	7.86	15.33	13.85
DCT	16	5.43	5.11	5.92	5.25	5.24	8.96	7.89
ZM	47	1.30	1.98	2.34	1.24	1.23	3.99	8.04
PCA	350	1.04	1.14	2.39	0.37	0.35	4.83	5.97
LDA	350	0.55	0.52	1.04	0.35	0.34	3.20	4.77
RP	350	0.33	0.50	0.74	0.34	0.34	3.12	8.04
DT	400	1.94	1.27	1.98	1.84	1.84	4.28	2.20
IMG	400	0.32	0.56	0.78	0.32	0.31	1.22	2.45

Error rates on Malayalam dataset.

Error rate using CNN : 0.93

Experiment: 1

Comparison of Classifiers and Features:

Observations

- SVM classifiers outperforms other classifiers, because of its high generalization capability.
- SVM-2 with a class of feature extraction techniques based on raw images and their projection on uncorrelated set of vectors resulted in the best performance.
- DTC and NB performed the worst of all.
- KNN performed moderately well, but with a higher computational requirement compared to SVM.

Experiment: 2

Richness in the Feature Space:

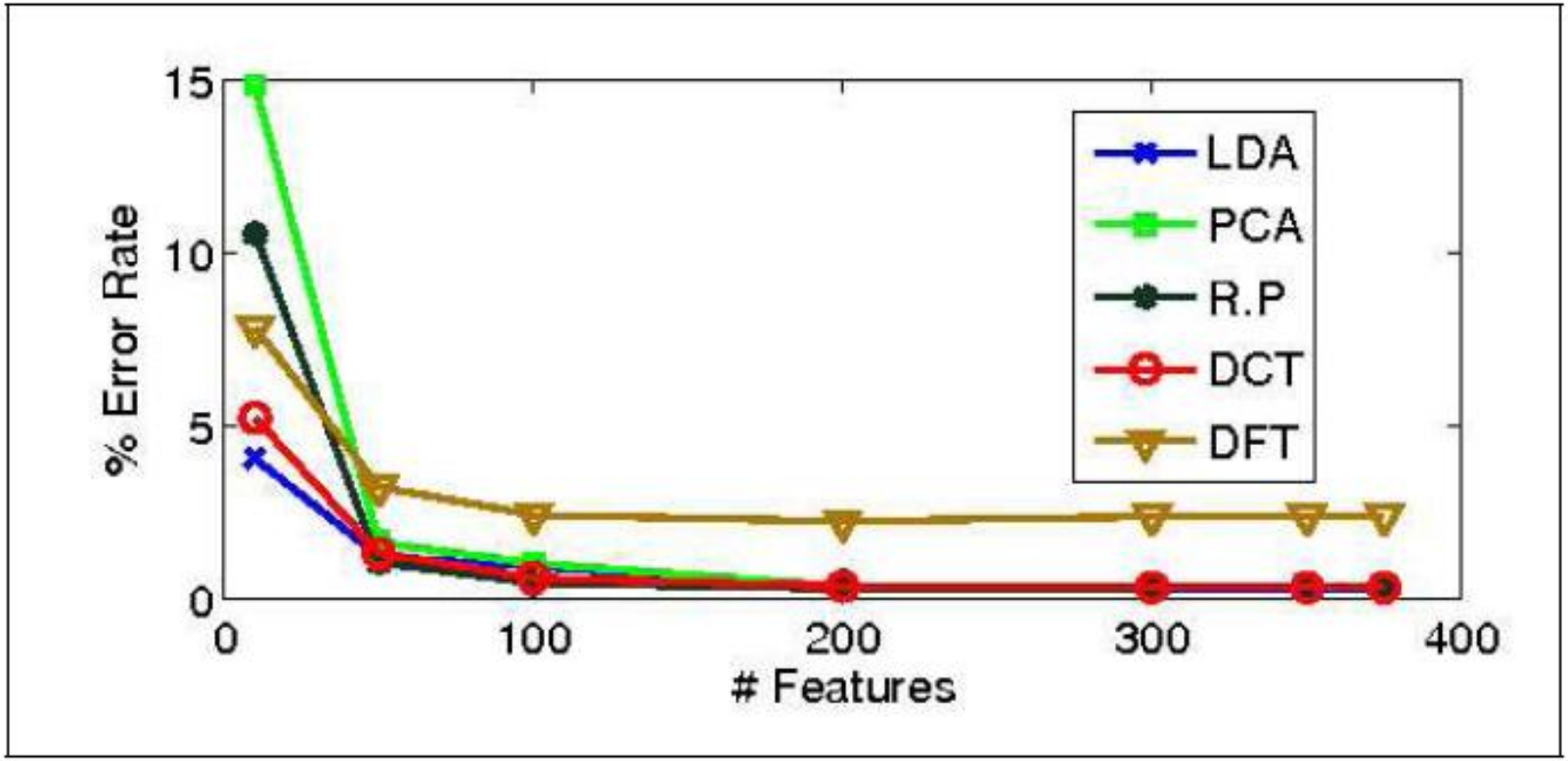
Experimental Settings

- What should be the ideal feature vector length for the problem to get solved successfully ?
- With a large number of features accuracy can be improved.
- We conducted the experiments by varying the feature vector length from 10 to 375.
- Features used for this study are, LDA, PCA, RP, DCT and DFT.

Richness in the Feature Space:

Results

Experiment: 2



Error rates of SVM-2 classifiers with varying number of features.

Experiment: 2

Richness in the Feature Space:

Observations

- Error rates rapidly decreases with the increase in number of features initially and then saturates after a point.
- When the number of features are small, LDA outperforms PCA.
- However with a large number of features PCA, LDA, RP performs more or less similarly.
- A rich feature space is needed to solve the classification problem successfully.

Experiment: 3

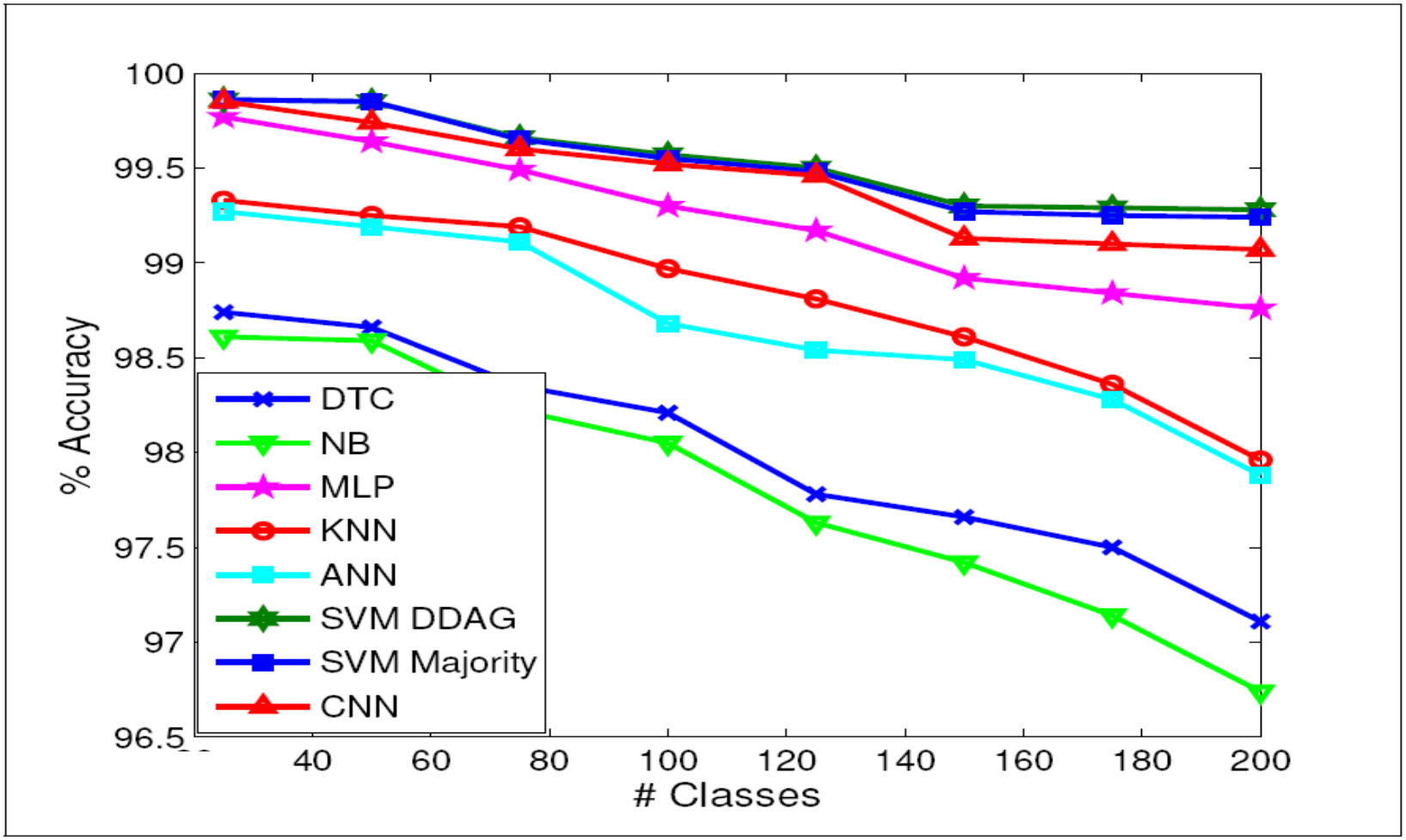
Scalability of Classifiers:
Experimental Settings

- Most of the publicly available datasets have small number of classes (in the order of a few tens).
- One of the major challenges in Indian language character recognition is the large number of classes (in the order of hundreds).
- How the performance of the classifiers effected with the increase in size of the problem (as the number of classes increases)?
- We conducted the experiments by varying the number of classes from 10 to 200.

Experiment: 3

Scalability of Classifiers:

Results



Accuracy of different classifiers Vs no. of classes, Feature used : LDA.

Scalability of Classifiers:

Observations

- Performance of all the classifiers goes down as the number of classes increases.
- SVM classifiers degrade gracefully with the increase in size of the problem.

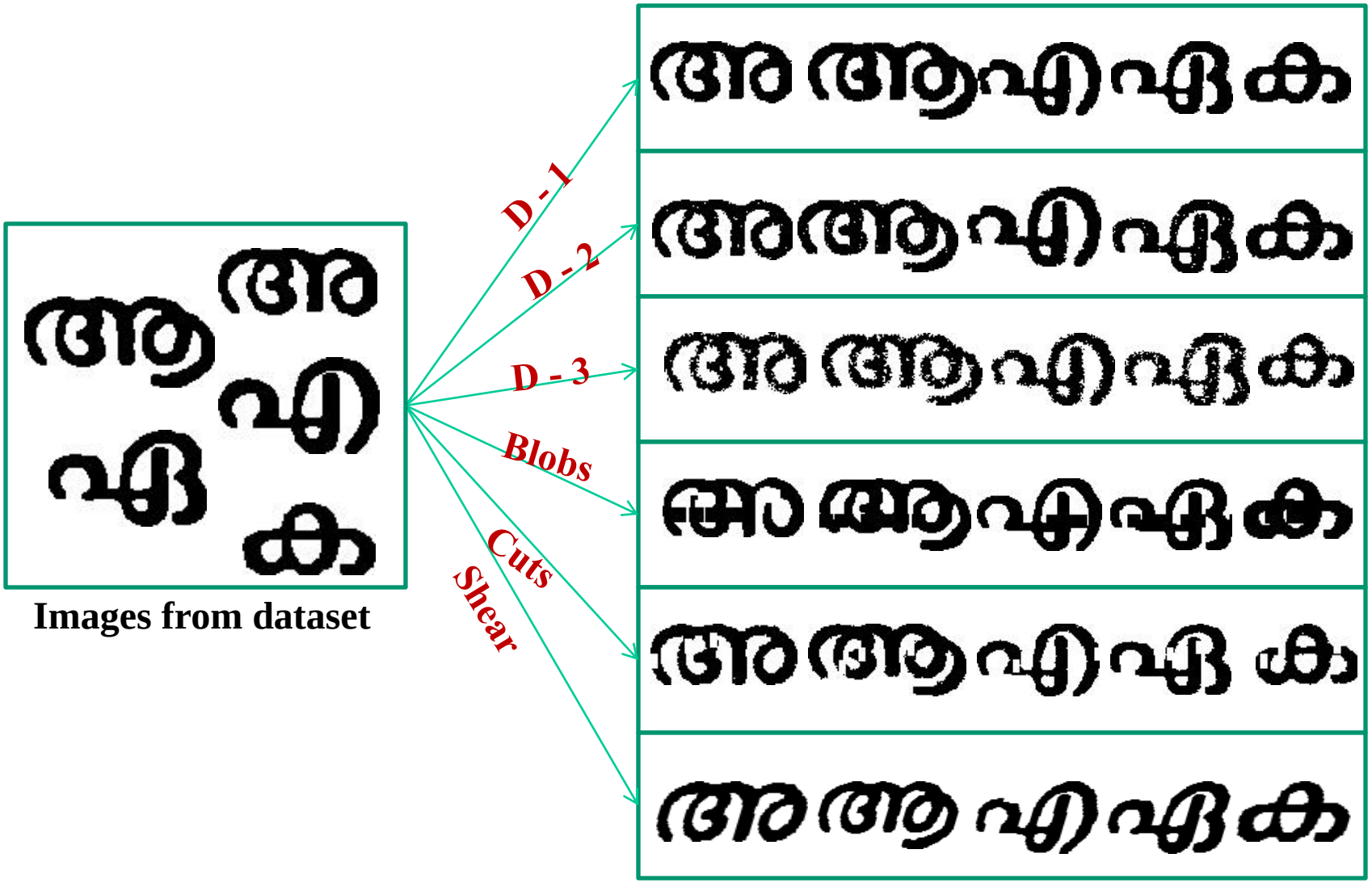
  (For 10 class problem, accuracy = 99.9, For 200 class problem = 99.3).
- The second best performing classifiers are Neural Networks.

Degradation of Characters :
Experimental Settings

- Characters in a real document are generally degraded.
- Most of the feature extraction techniques will have difficulties to extract the right features, in the presence of degradations.
- We modeled 6 degradations.
  - D1, D2, D3 (based on boundary erosions).
  - Cuts, Ink blobs, Shear.

Q. Zheng and T. Kanungo, .Morphological degradation models and their use in document image restoration,. in *ICIP, pp. 193.196, 2001.*

Examples of degraded images



Images from dataset

Experiment: 1

Degradation of Characters:

Results

Feature	D-1	D-2	D-3	Blobs	Cuts	Shear
C.M	9.45	9.46	10.97	16.28	12.33	30.07
DFT	7.89	7.93	7.98	26.70	8.73	18.90
DCT	5.71	5.72	6.07	19.80	7.93	16.46
ZM	1.96	1.98	2.10	8.41	4.35	17.75
PCA	0.39	0.39	0.40	2.17	0.64	8.59
LDA	0.30	0.31	0.32	2.01	0.61	7.32
RP	0.48	0.67	1.04	3.61	0.71	6.75
DT	1.75	1.98	2.21	10.33	5.07	12.34
IMG	0.32	0.33	0.33	2.78	0.66	6.84

Error rates of different features on various degradations using SVM-2 classifier.

Experiment: 4

Degradation of Characters :

Observations

- Statistical features are reasonably insensitive to the small degradations (D1, D2 and D3).
- Features like DT, which works well with clean images fails with cuts and ink blobs in the character.
- A better performance is observed for features PCA, LDA, RP and even raw images(IMG) on degradation, compared to others.
- Shear is a more challenging problem, need more consideration in this aspect.

Generalization Across Fonts:

Experimental Settings

- How sensitive is the classifier performance on an unseen font ?
- The study included 5 popular fonts in Malayalam.
  - MLTTRevathi, MLTTKarthika, MLTTMalavika, MLTTAmbili, MLTTKaumudi.
- Train the classifier with 4 fonts and test on the 5th font.

Experiment: 5

Generalization Across Fonts:

Results and Observations

	Font -1	Font -2	Font -3	Font - 5	Font-4
S1	98.15	95.49	92.52	94.27	92.22
S2	98.97	97.14	95.22	94.59	94.65

Accuracies of SVM-2 classifier when trained with 4 fonts and tested on the 5th font. S1 : Dataset without degradation, S2: Dataset with degradation.

- Generalization across fonts can be achieved by having a wide variety of fonts in the training set.
- A better performance can be achieved by adding a little degradation to the training data.

Experiment: 6

Applicability across Scripts:

Experimental Settings

- Can the previous experimental results be extended to other scripts ?
- Experiments conducted on Telugu and English scripts.
- Around 50,000 real character images from each scripts used for the experiments.

Applicability across Scripts:

Results

Experiment: 6

Features	Telugu (350 class)		English (72 class)	
	20X20	40X40	20X20	40X40
C.M	20.78	12.32	7.25	6.48
DFT	8.45	5.48	2.04	1.12
DCT	9.67	2.71	2.14	1.04
ZM	15.71	6.71	5.37	3.31
PCA	4.62	2.93	0.86	0.46
LDA	2.56	1.67	0.29	0.23
RP	2.49	1.66	0.28	0.23
DT	3.48	3.17	0.98	0.87
IMG	3.18	2.84	0.28	0.23

Error rates on Telugu and English Datasets, with SVM-2 classifier.

Applicability across Scripts:
Observations

- The conclusions on character classification are highly script/ language independent.
- More complex scripts can be approached with a richer feature space, which gives more discriminative features.

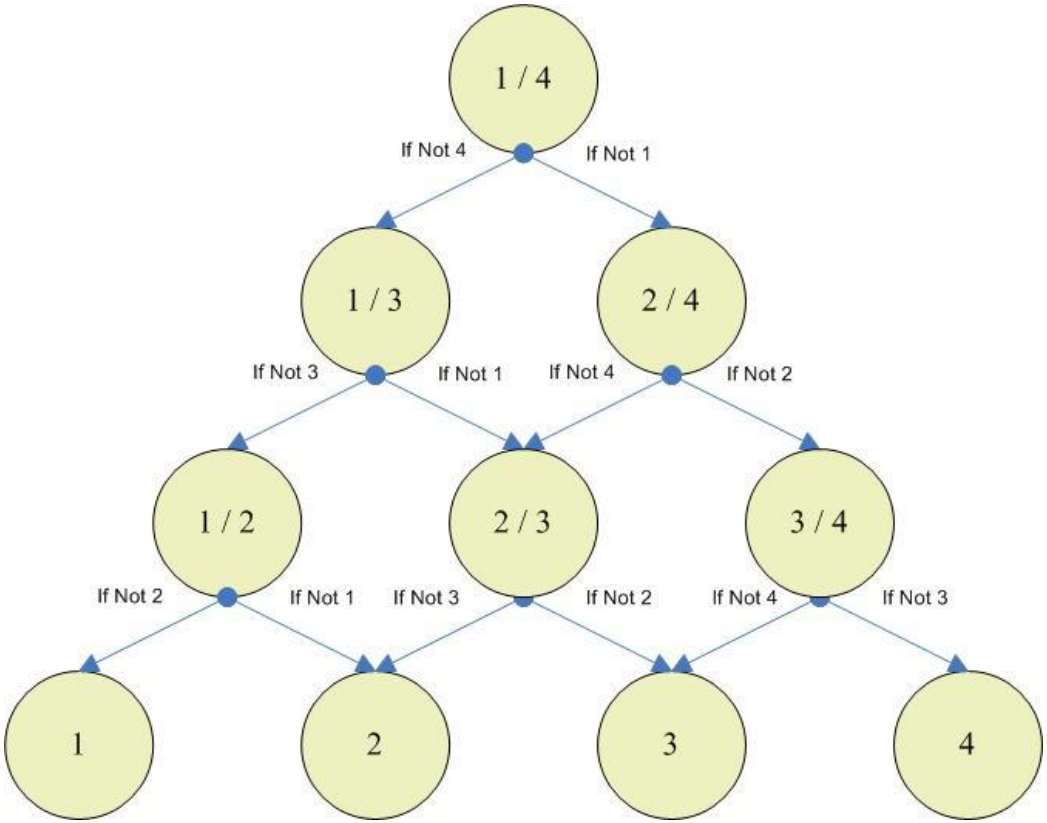
Summary

- We presented the results of empirical study on character classification problem focusing on Indian languages.
- Dimensions of the study included:
  - Comparison of classifiers and features.
  - Richness in the feature space.
  - Scalability of classifiers.
  - Sensitivity of features to degradation.
  - Generalization across fonts.
  - Applicability across scripts
- Error rates are reported on isolated segmented characters/symbols
- Indian Language character classification problem can be successfully solved
  - With the use of rich feature space.
  - Using a set of statistical features.
  - And state of the art modular classifiers like SVM.

Thank you ☺

Questions ?

SVM-DDAG:Architecture



DAG for 4 class components