Detection and Cleaning of Strike-out Texts in Offline Handwritten Documents

Bidyut B. Chaudhuri
INAE Distinguished Professor
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
www.isical.ac.in/~bbc

On OCR Problems

- OCR of printed text is considered a solved problem.
- OCR of handwritten text is still challenging.
- Major progress has been made on English handwriting recognition; but for Indian scripts, we have a long way to go.
- Abundant English handwriting databases (IAM, Univ. of Washington) are available for research. On Indian scripts, the database generation process is advancing slowly (e.g. ISI, JU database).
- Methods based on SVM, HMM and BLSTM have pushed the English handwriting accuracy to respectable level. More recently, experiments have started on Indian scripts.(Sankaran & Jawahar, 2012, Garain et al, 2015, Adak et al. 2016).

Handwriting Recognition Issues

- Almost all handwritten text recognition articles assume that the document texts are flawlessly written.
- In reality, chances of error in unconstrained handwriting are fairly high.
- There may be various kinds of writing errors. Perhaps the most common is the strike-out error. The writer strikes-out a wrong/inadequate word and writes the proper word next to it. This may be called First-draft correction.
- In general, strike-outs can be as small as one character and as big as multi-line or a full paragraph.
- Various editing operations may be done in the post-writing revision, which may be called On-revision correction.
- If such a document image is directly fed to OCR, then the output will be highly erroneous.
- So a preprocessing module is required to get high OCR accuracy. Else, a more complex recognition scheme is needed.

Editing in Handwritten Manuscript (Tagore)

Insertion with caret

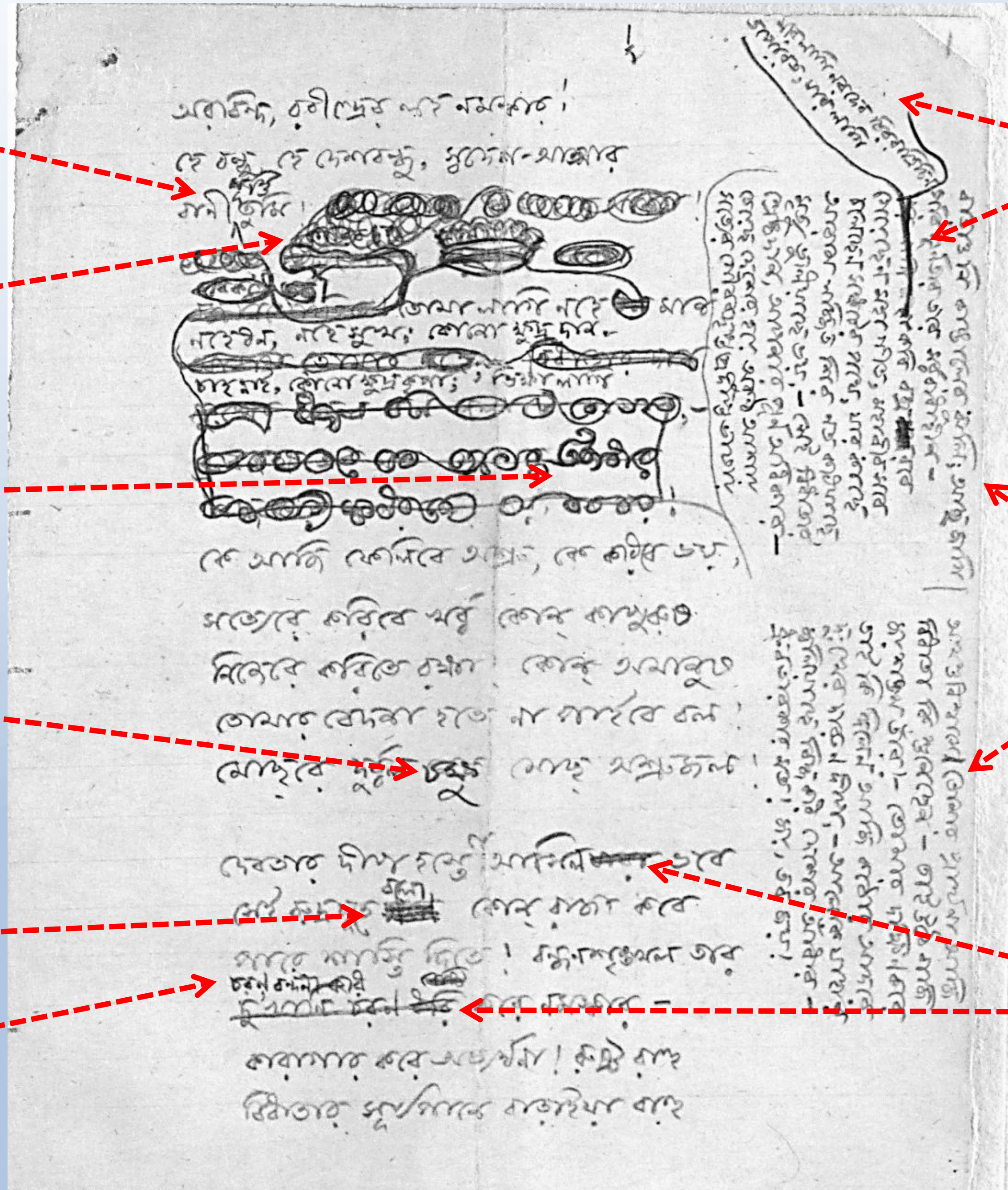
Tree-like Doodle

Ornamental struck-out

Overwriting

Blackened text

Insertion without caret



Struck-out & unconventional insertion

Vertically Oriented Text

Struck-out text

Struck-out Text Processing

- In this work we consider only strike-out text processing.

Motivation of the work:

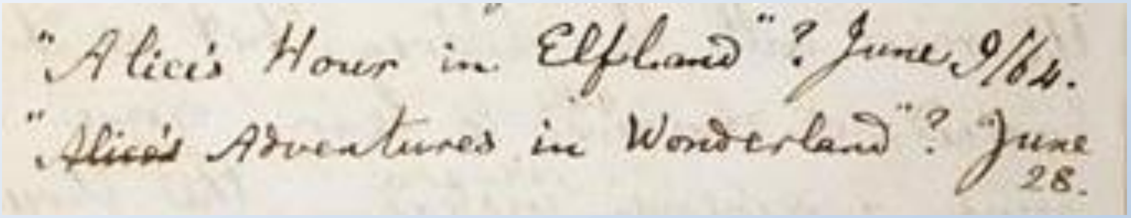
- *OCR Application:* Aid to OCR & digital transcription generation.
- *Forensic Application:* Detection of struck-out texts and their patterns may provide important psychological clues for the forensic experts.
- *Cognitive Application:* Examining the struck-out words and their replacements may shed light to the behavioral pattern of a writer, in general and mentally challenged patients, in particular.

Tasks to be done:

- Identification of Strike-out words.
- Localization of Strike-out Strokes (SS)
- Cleaning of struck-out words by deleting the SSs.

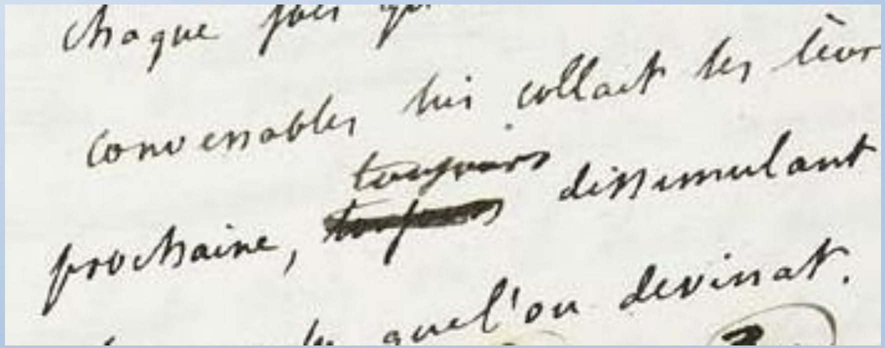
Typical Examples of Digital Transcription

(a)



(b)

“Alice's Hour in Elfland”? June 9/64.
~~“Alice's~~ Adventures in Wonderland”? June 28.



(c)

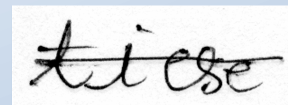
mais, chaque fois qu'elle s'offrit, l'embarras
convenables lui collait les lèvres, et il s'ajour
toujours
prochaine, ~~toujours~~ dissimulant ce qu'il eût
bien voulu que l'on devinât.

(d)

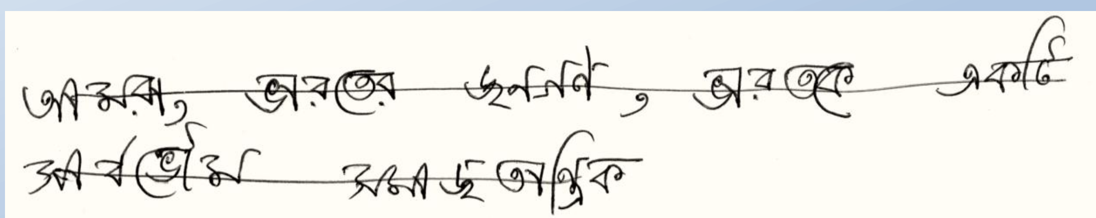
Snippets of (a) ***Lewis Carroll***, (c) **Gustave *Flaubert*** manuscripts and their transcriptions (b), (d).

Strike-out Strokes of Different Sizes

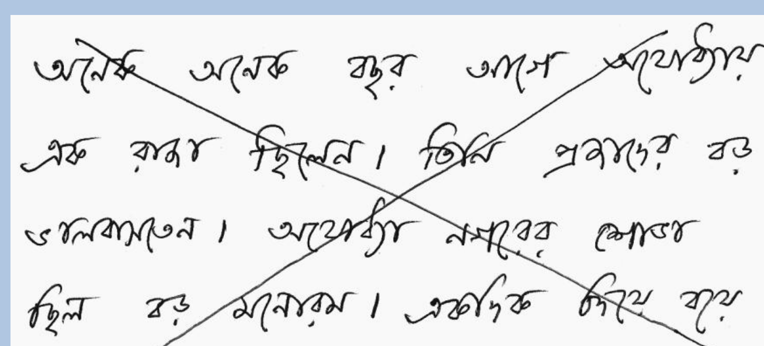
Character level strike-out



Word level strike-out

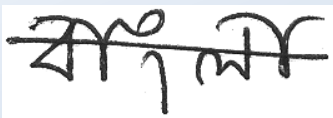


Successive multi-words strike-out

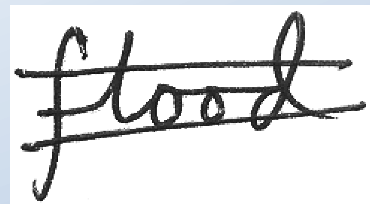


Successive multi-line strike-out

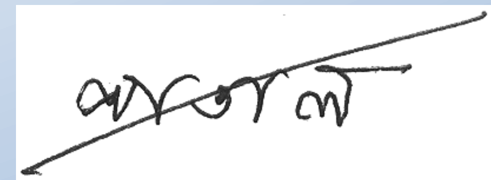
Strike-out Strokes of Different Styles



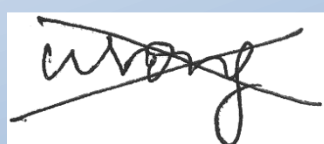
(a) Single



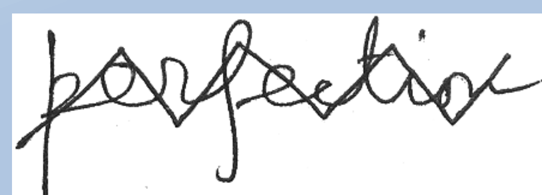
(b) Multiple



(c) Slanted



(d) Crossed



(e) Zig-zag

(f) Wavy

Related Works in Literature

Method	Description
Arlandis et al. [ICPR-2002]	Mentioned the SS problem, but did not provide any solution.
L-Sulem et al. [ICFHR-2008]	Used Markov Random Field (MRF) based method to identify Struck-out. No % accuracy was reported. SS was not detected.
Nicolas et al. [IWFHR-2006]	Hidden Markov Model (HMM)-based method of word recognition. SS was simulated by artificially made superimposed strokes. Identification or cleaning of such simulated SSs was not reported.
Banerjee et al. [CVPR-2009]	The machine-printed documents vandalized by longer ink-strokes in different directions were reinstated using a MRF-based document learning model. Neither struck-out word detection nor SS identification were performed.
* Brink et al. [DRR-2008]	Used binary classifier to remove struck-out text. Automatic removal of 47.5% struck-out words with 99.1% preservation of normal text were reported, but no fair-copy generation was done.
ABBYY [US patent #847271925, 2013]	Recognized crossed out English characters by a feature-based classifier. Word/line level strike-outs were not considered. Detailed approach was unavailable.

SS : Strike-out Stroke

Possible Problem-solving Ways

- Design a single recognizer that can generate correct transcription including strike-out using some deep-learning based method (e.g. BLSTM). But we could not design a BLSTM system with high Bangla OCR accuracy.
- Sub-divide the problem into modules of (a) finding strike-out text, (b) locating the SSs, (c) cleaning the SS, (d) generate the transcription.
- The advantage of second method is that different methods can be used at different modules.

BLSTM-CTC based Unconstrained Handwritten Bangla Text Recognition

System 1:

- 2338 handwritten lines from 100 writers. Training : Validation : Test = 3 : 1 : 1 .
- 30X8 window with 8-directional HOG feature in 2X4 sub-windows, i.e. 64 features.
- BLSTM input layer contains 64 nodes. 2 hidden layers are of 200 neurons. CTC layer is of 917 output nodes.
- 917 corresponds to the same number of semi-orthosyllables (*semi-Akshara*) of Bangla text.
- Semi-orthosyllable level accuracy = 75.40% . Substitution, deletion, insertion errors are 18.91%, 4.69% and 0.98%, respectively.

System 2:

- Instead of HOG feature, we extracted features from (LeNet-5). The number of features = 128, standardized using z-score.
- The semi-ortho-syllable level accuracy = 86.13% . Substitution, deletion, insertion errors are 9.54%, 3.10% and 1.23%, respectively

1. U. Garain, L. Mioulet, B. B. Chaudhuri, C. Chatelain, T. Paquet, “Unconstrained Bengali handwriting recognition with recurrent models”, Proc. ICDAR, pp. 1056-1060, 2015.

2. C. Adak , B. B. Chaudhuri , M. Blumenstein , “Offline Cursive Bengali Word Recognition using CNNs with a Recurrent Model”, Proc. ICFHR, pp. 429-434, 2016.

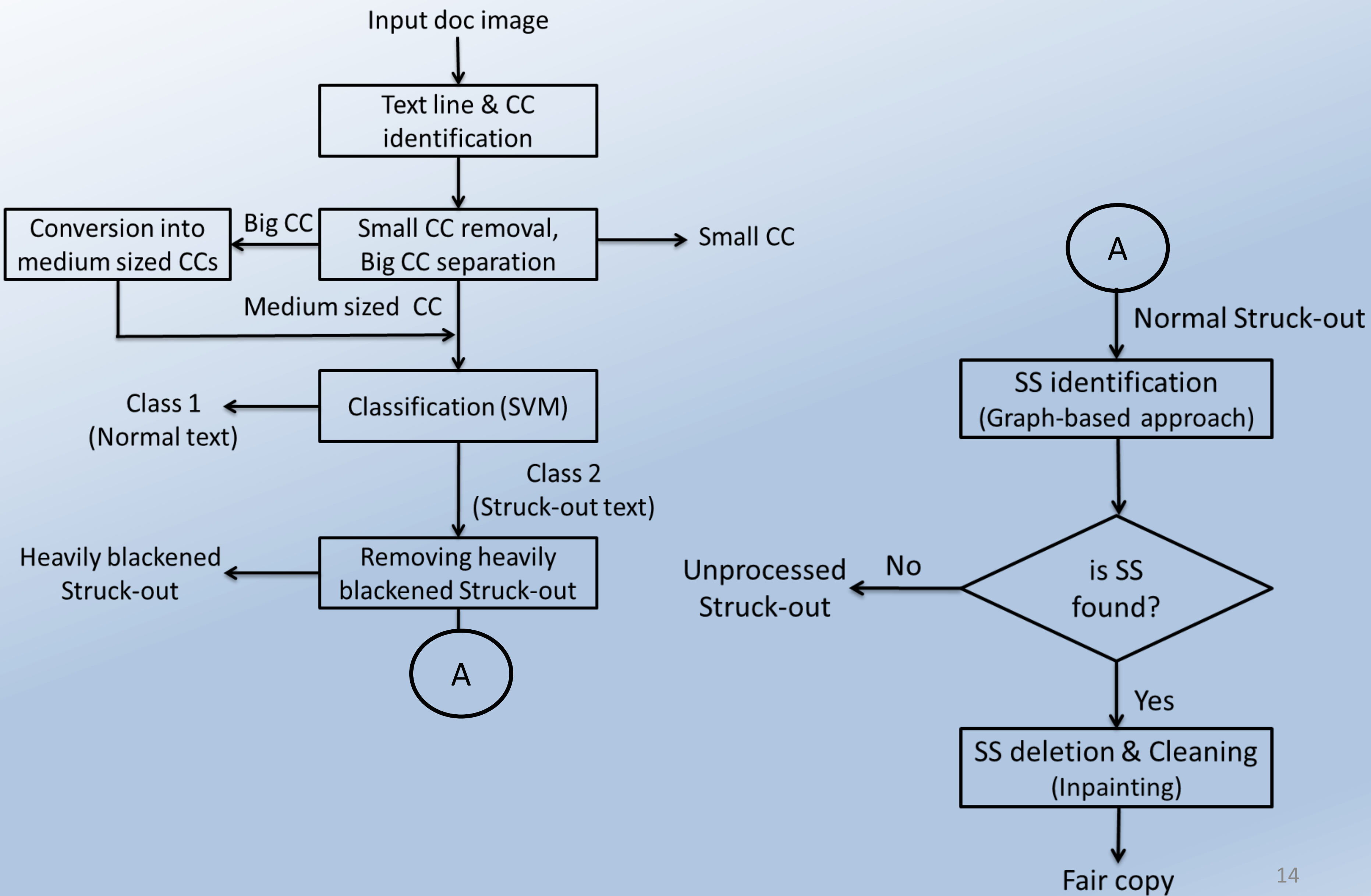
Proposed Struck-out Text Processing Approach

- Document image pre-processing.
- Strike-out word detection by SVM.
- Strike-out Stroke (SS) identification by graph path finding.
- Cleaning of strike-out words by image inpainting.

Preprocessing for Struck-out Recognition

- Document noise cleaning and binarization.
- Skew correction and text region isolation.
- Individual Text lines segmentation and word isolation.
- Connected Component (CC) identification.
- Very small-sized CCs (dot, comma, colon etc.) deletion.
- Abnormally Big-sized CCs identification.
- Word formation by medium CCs (to send to SVM classifier).
- Segmenting Big CCs into small CCs.

Work Flow of Proposed Method



Primary Strike-out Detection by SVM

- Each word is subject to a **SVM with RBF kernel** based 2-class classifier.
- **2 Classes** : *non-struck-out* (*class-1*) words and *struck-out* (*class-2*) words.
- **7 features**: **3** branch-point based, **2** density based and **2** hole based features.
- A factor called elongation (E_{cc}) is computed from the height (H_{cc}) and width (W_{cc}) of a word bounding box as:

$$E_{cc} = \frac{\min\{H_{cc}, W_{cc}\}}{\max\{H_{cc}, W_{cc}\}}$$

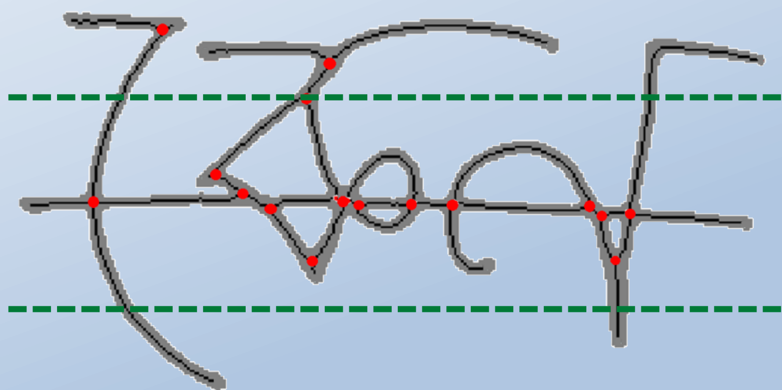
E_{cc} is used as *normalizing factor* for the features as follows.

Hand-Crafted Features

1. Branch point (F_{BP}): The skeleton of the word image is found. Here, the pixel-points where three or more strokes intersect are called *branch points*. Feature F_{BP} is defined as

$$F_{BP} = N_B/E_{cc}$$

where N_B is the total number of branch points. The SS intersects text strokes, increasing the number of branch points.



2. Weighted branch points (F_{BPW}): The word is partitioned into three horizontal zones and the branch points in the middle zone are given more weight since the SS is more likely to lie in this zone. Thus, the zone-weighted branch point based feature is given by

$$F_{BPW} = (\omega_u.N_{BU} + \omega_m.N_{BM} + \omega_l.N_{BL})/E_{cc}$$

where N_{BU}, N_{BM} and N_{BL} are the number of branch points in the upper, middle and lower zone. The weights ω_u, ω_m and ω_l are found by a data-driven approach.

Hand-Crafted Features (Contd…)

3. ×-like branch points (F_{BPX}): When the SS cuts through another stroke, a ×-like branch point with four edges are formed. Let the number of such branch points be (N_{BX}). Then

$$F_{BPX} = N_{BX} / E_{cc}$$

4. Normalized black pixel density (F_D): The number of foreground pixels N_F, divided by the total number of pixels N_T in the bounding box (BB) is normalized by E_{cc} to get

$$F_D = \frac{N_F / N_T}{E_{cc}}$$

5. Standard deviation of density (F_{SD}): Sub-divide a component BB into T_s equal horizontal strips & count the number of black pixels (n_i, for $i = 1, 2, \ldots, T_s$) in each strip.

$$F_{SD} = \frac{\sigma}{E_{cc}}, \quad \text{where } \sigma = \sqrt{\frac{1}{T_s}\sum_{i=1}^{T_s}(n_i - \mu)^2} \text{ and } \mu = \frac{1}{T_s}\sum_{i=1}^{T_s} n_i$$

The parameter T_s is fixed by experimental analysis.

Hand-Crafted Features (Contd…)

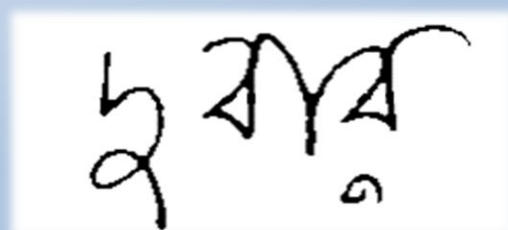
6. Normalized number of holes (F_H): Let N_H be the total count of holes in the word. Then

$$F_H = N_H / E_{cc}$$

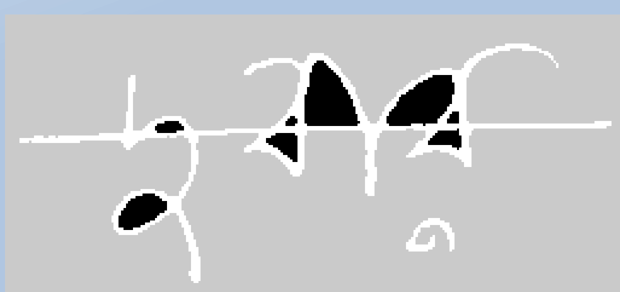
7. Hole pairs with common straight side (F_{CS}): When an SS passes through a hole, it creates two holes. One side of each hole is fairly straight and is common to the other hole on the opposite side. Count of such hole pairs (N_{CS}) gives the feature

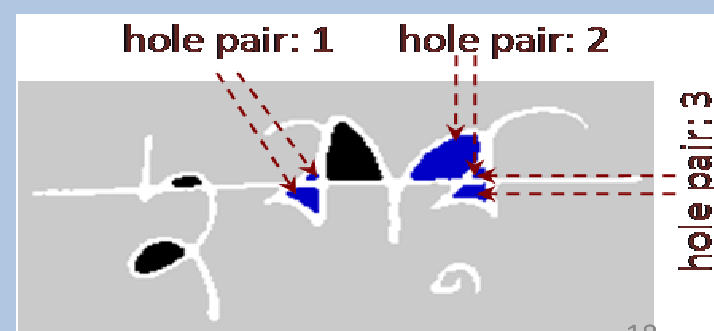
$$F_{CS} = N_{CS} / E_{cc}$$

initial hole = 3



hole increased to = 8





Auto-Extracted Features by CNN

- The features are extracted automatically using CNN.
- We use the *LeNet-5* CNN architecture*.
- The features are obtained after 2 (convolution + sub-sampling layers) followed by a fully-connected layer.

- The input word image is normalized into 32 X 92 pixels.
- The CNN produces a feature vector with dimension 480.

*Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition”, Proceedings of the IEEE, vol.86, no.11, pp.2278-2324, 1998.

SVM Classifier Characteristics & Training

- SVM is a powerful supervised classifier. Here SVM with **RBF** kernel is used for its efficiency in text processing.

- The hyper-parameters gamma (***γ***) and cost (***C***) are to be set for best results:
  The parameter ***γ*** avoids *overfitting*
  The parameter ***C*** controls *decision boundary*

- Hyper-parameters ***γ*** , ***C*** are tuned from a subset of data from the training database.
  (We use **20%** of the database for tuning)

- ***K***-fold cross validation is used (we chose *K*=5) for training.

N. Cristianini, J. S.-Taylor, "An Introduction to Support Vector Machines and other kernel-based learning methods", Cambridge University Press, 2000. ISBN 0-521-78019-5.

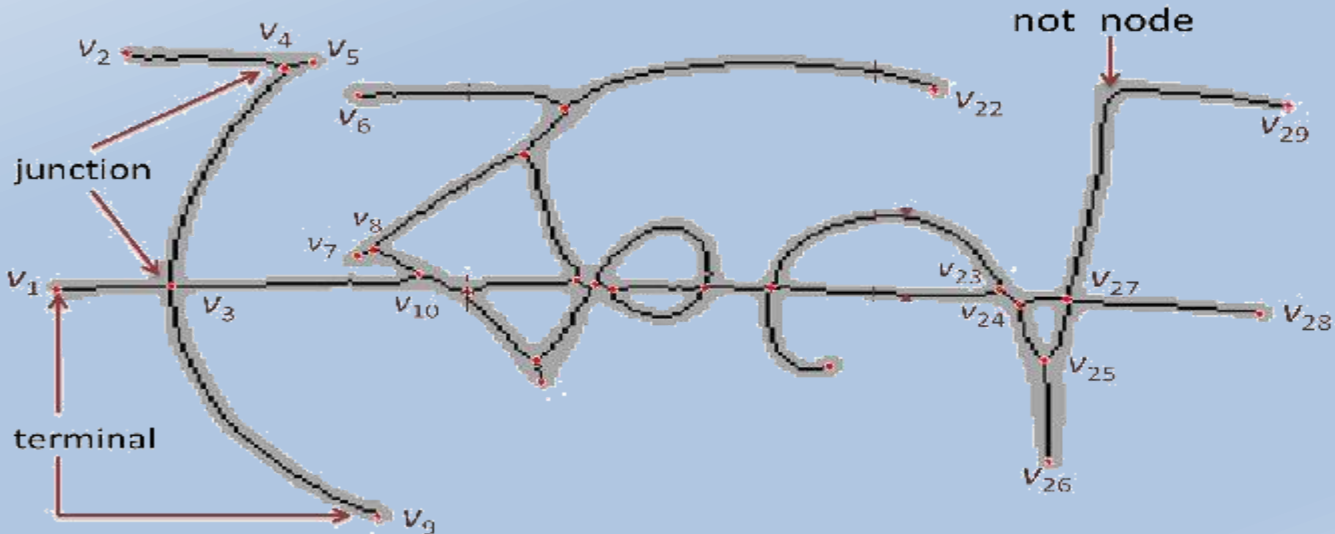
Text Skeleton Graph-based SS identification

The struck-out word skeleton I_{sk} is considered as an undirected graph $G = (V, E)$, where V and E are the set of *nodes* and *edges*, respectively.

- The *end* and *intersection* pixels of I_{sk} are the *terminal* and *junction* nodes in V.
- An edge (e_{ij}) exists between 2 node pixels v_i and v_j, if they are directly connected.
- The weight (w_{ij}) of an edge (e_{ij}) is found from the number of diagonal moves (N_d) and h/v moves (N_{hv}) in the edge between v_i to v_j , where (w_{ij}) is given by:

$$w_{ij} = \omega_d . N_d + \omega_{hv} . N_{hv} \quad \text{where } \omega_d = \sqrt{2} \text{ and } \omega_{hv} = 1.$$

- The straight and shortest path passing through middle of the skeleton is chosen as SS skeleton.

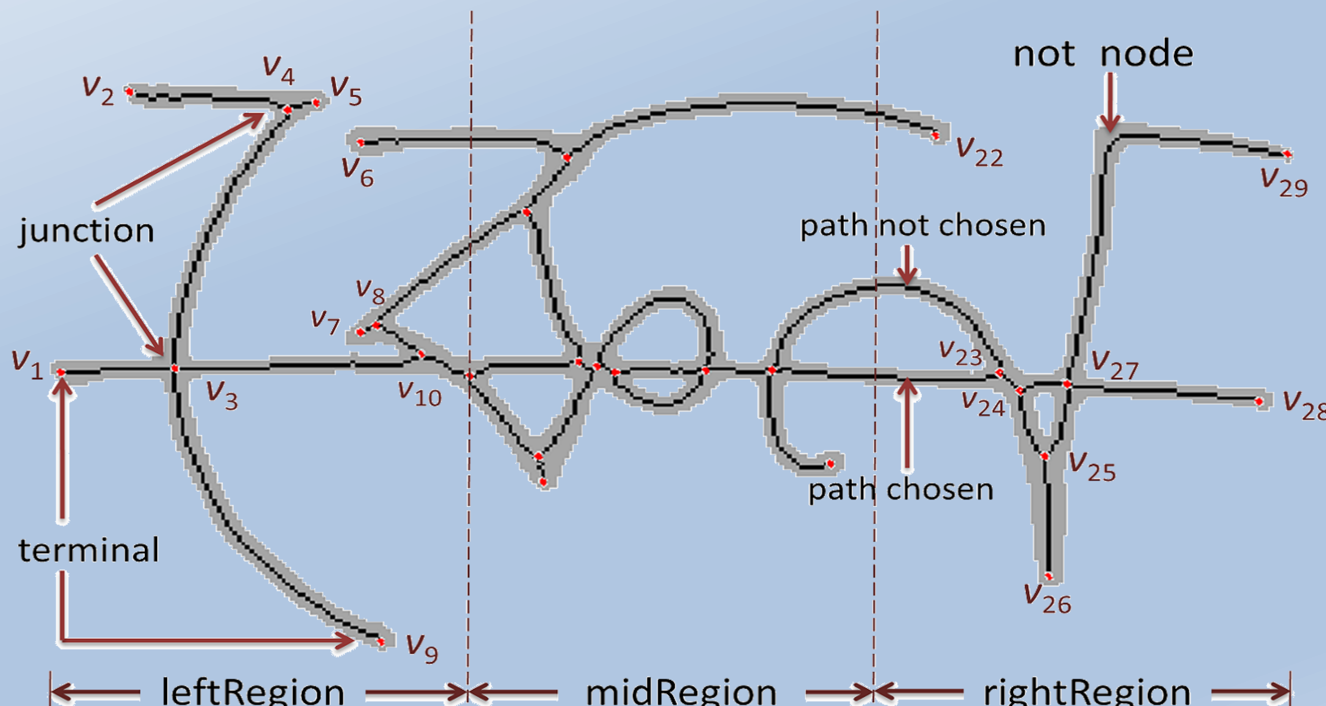


Localization of Strike-out Stroke (SS)

- Shortest Paths (SP) from terminal nodes of left region to the terminal nodes of right region are found using Dijkstra's algorithm.

Speed-up of Path Finding Algorithm:

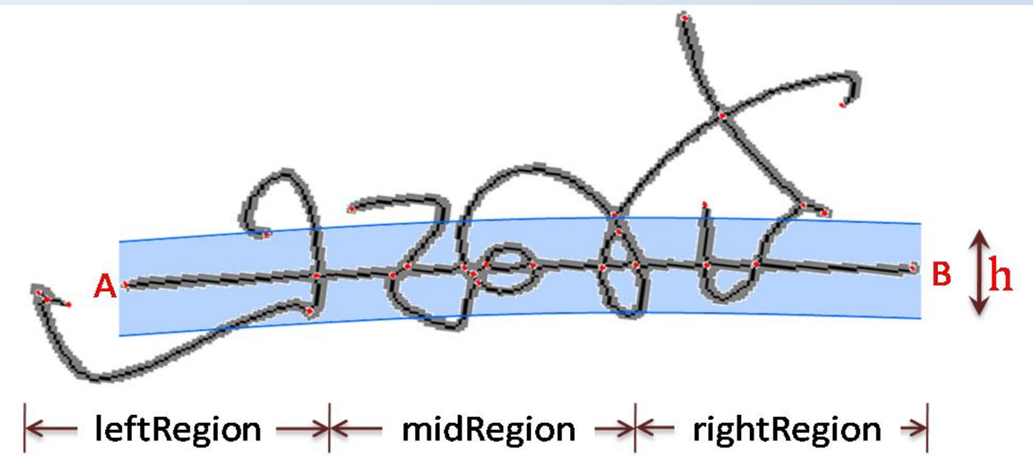
- The skeleton is partitioned vertically into three equi-spaced regions. The SPs are found mostly between nodes of leftmost region to the nodes of rightmost region.
- The self-loop at any node is deleted.
- If multiple edges between two neighboring nodes exist, then only the shortest edge is considered in calculation.
- In SS, usually no retrograde move is seen. So, path with backward move is disallowed.

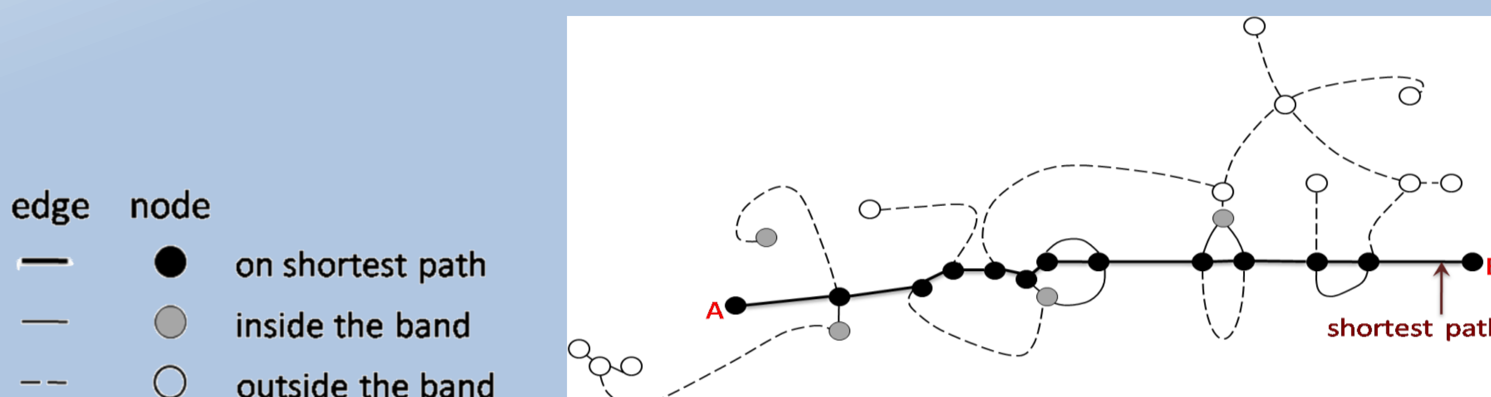


Localization of SS (Contd…)

Speed-up Approach (Contd…) :

- To find SP between v_i and v_j let $\overline{v_i v_j}$ be the straight line joining them. The nodes and edges which entirely lie within a band of thickness h around this line are only considered. This is because we want the SP to be reasonably straight.





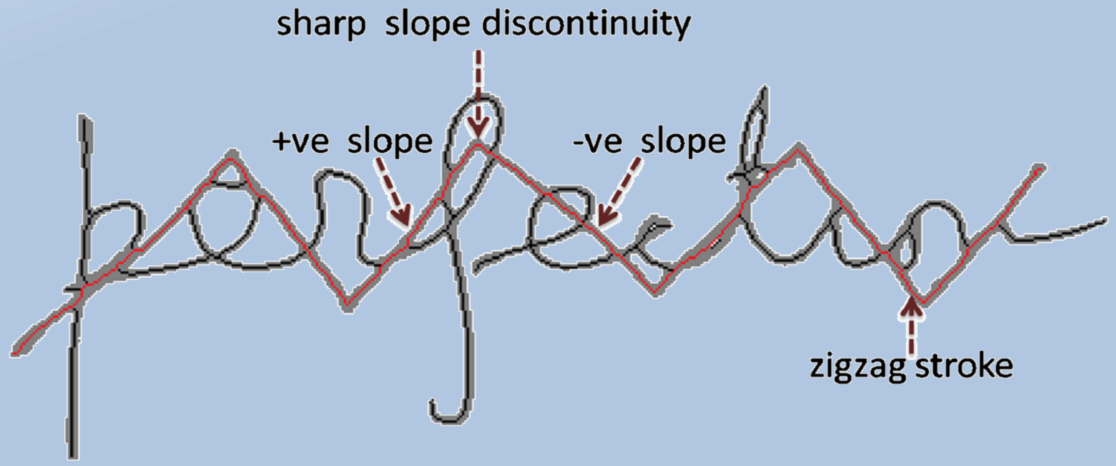
Zig-Zag Strike-out Stroke Detection

Characteristics:

(1) Zigzag stroke is usually drawn on words longer than two characters.
(2) The stroke usually covers most characters of the word.
(3) The stroke passes through the middle, and zigzag span almost covers the characters.
(4) The zigzag shape is characterized by points having sharp slope discontinuity.
(5) The stroke is quite linear between two consecutive slope discontinuity points.

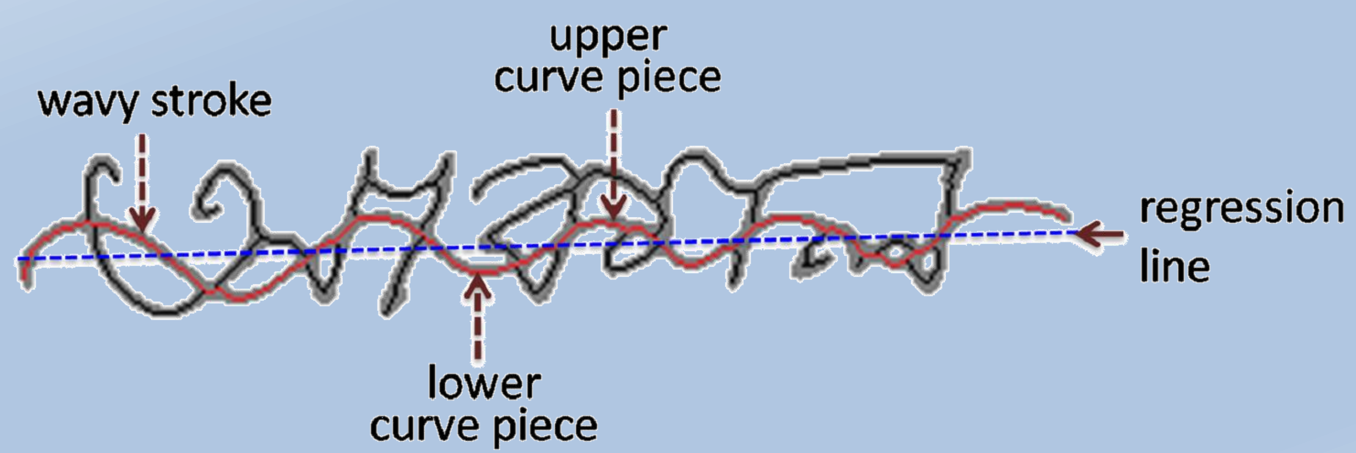
Detection method:

- Find all paths between left and right zone terminal nodes (Not shortest path).
- Look for curvature discontinuities. Arrange them in increasing x-values.
- The successive discontinuity will have high to low or low to high y-value.
- The path between successive discontinuity should be almost linear.
- A path having the above properties is considered as zig-zag SS.

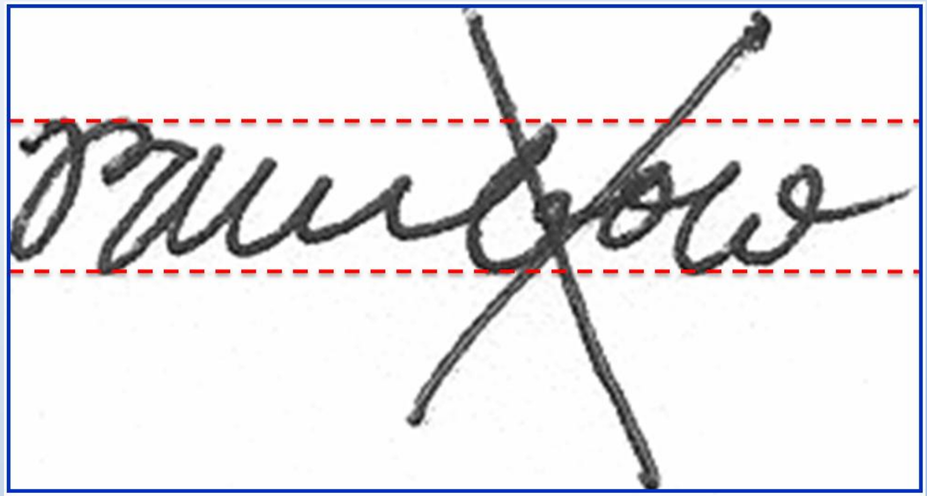


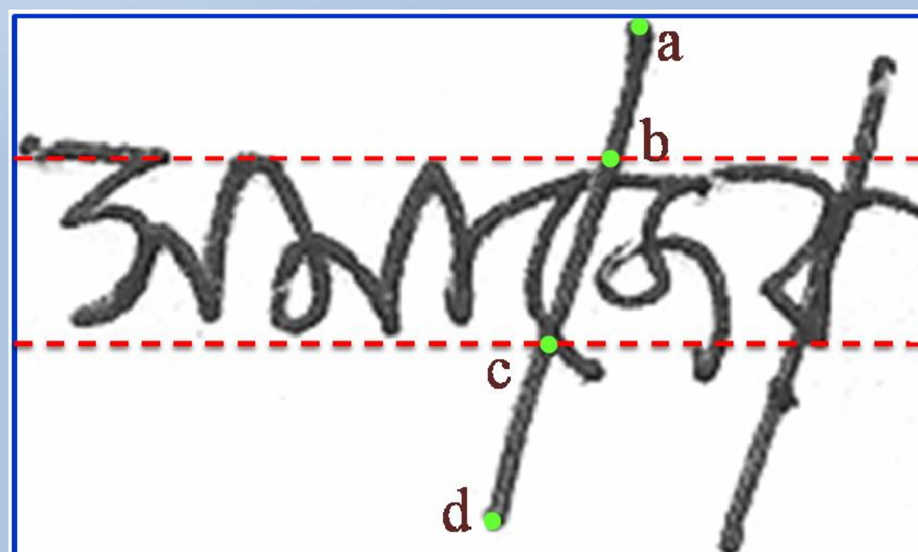
Wavy Strike-out Stroke Detection

- Find all paths between left and right zone terminal nodes (Not shortest path).
- Draw a near-horizontal regression line, segmenting the curve into pieces.
- If the path is a wavy stroke then
  1. A curve piece has a maximum (for the upper piece) followed by a piece with a minimum (for the lower piece) or vice versa.
  2. The average width of the upper pieces and lower pieces are nearly equal.
  3. The av. height of the upper pieces and the depth of lower pieces are nearly equal.
  4. The regression line is near midway to the CC bounding box height.
  5. The average area of upper pieces bounded by regression line is nearly equal to the average area of lower pieces.
- A score S is computed based on the above properties. The path with maximum S above a threshold T indicates wavy SS.

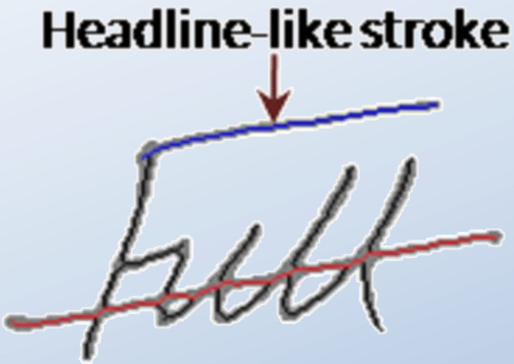


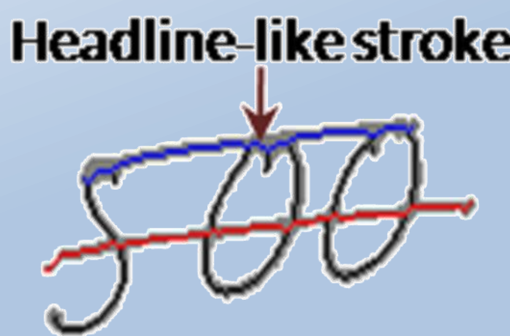
Recognizing Non-horizontal Strike-outs





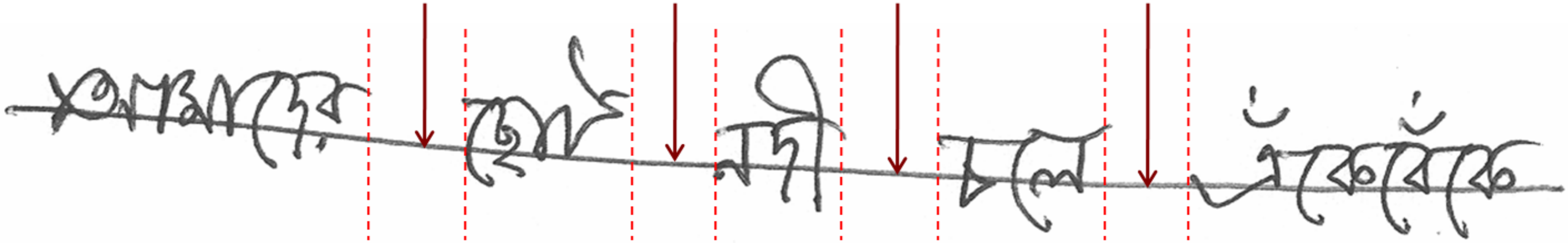
Handling Multiple SS & Script-specific SS



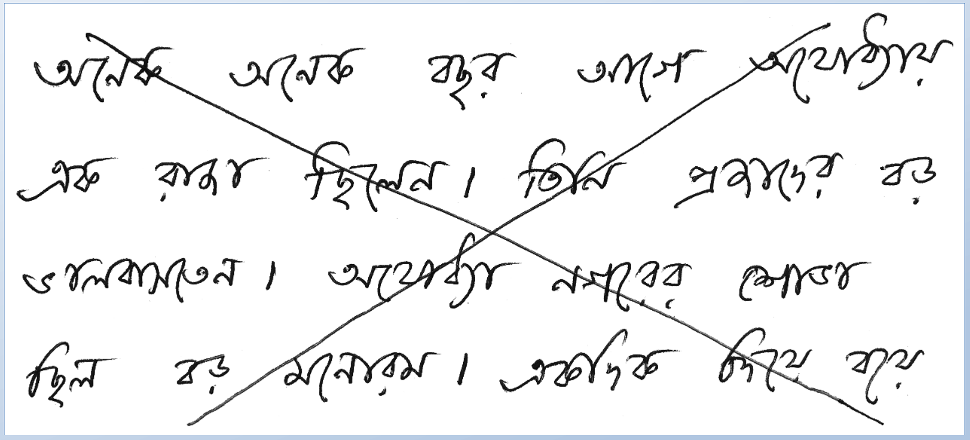


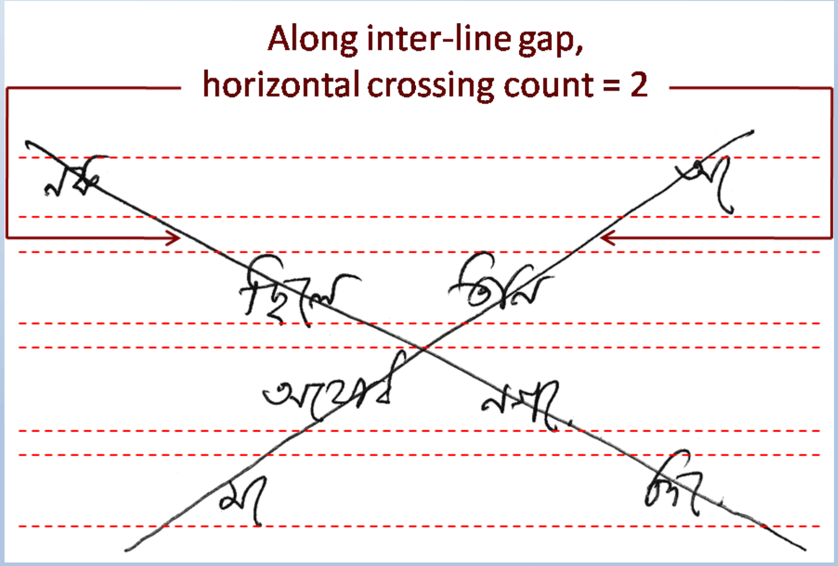
Multi-word Strike-out Detection

Along the arrows, vertical crossing count = 1



Multi-line Strike-out Detection

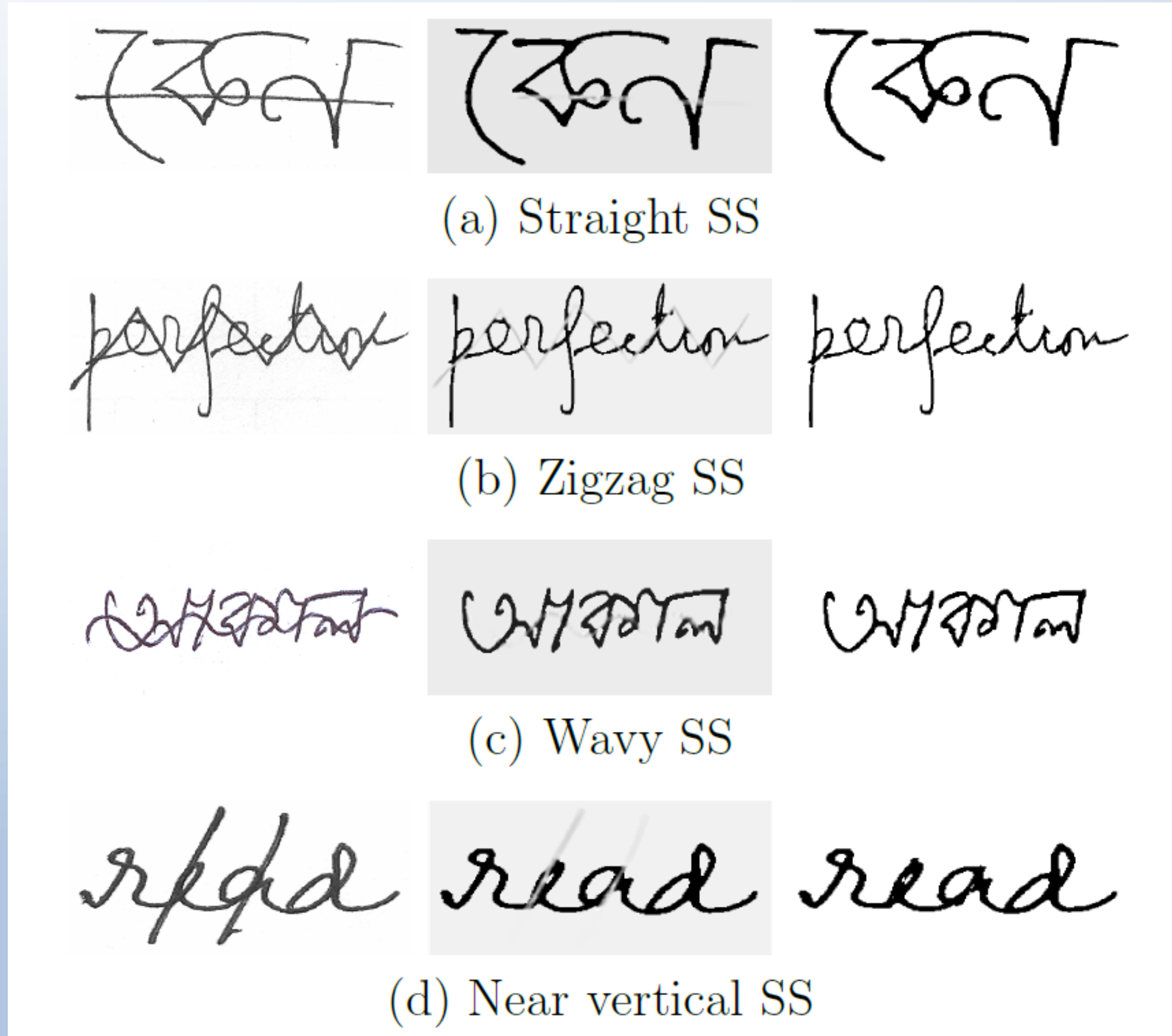




Cleaning of Strike-out Stroke (SS)

- The SS skeleton, detected by graph-based approach is morphologically dilated.
- We use Image Inpainting approach for cleaning this stroke portion.
- Inpainting requires a “mask” region, that is used to fill by the texture of neighborhood regions.
- The morphologically dilated version of SS is used as mask.
- After Inpainting, we binarize the inpainted image using an adaptive threshold.

Cleaning of SS by Inpainting



(a) Straight SS

(b) Zigzag SS

(c) Wavy SS

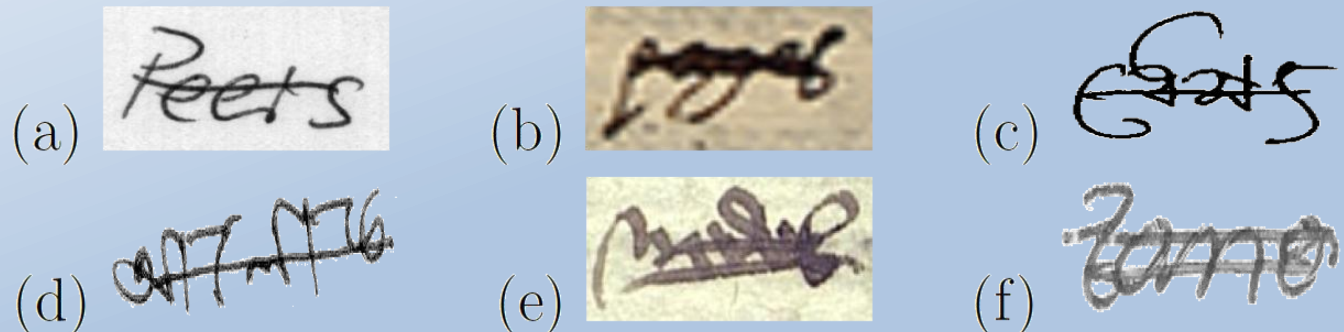
(d) Near vertical SS

Left column: original image, **Middle column**: image after inpainting, **Right column**: final binarized output.

Experimental Results

Publicly Available Databases:

IAM (IJDAR 2002): University of Bern, Switzerland.
BH2M (ICPR 2014): Barcelona Historical Handwritten Marriages database.
ICDAR segmentation contest (ICDAR 2013): Handwriting segmentation contest.
CMATERdb1 (IJDAR 2012): Jadavpur University database, Kolkata.
MLS (ICFHR 2014): Monk Line Segmentation (MLS) Dataset, Groningen.
BFL (IWFHR 2008): Brazilian Forensic Letter database.



Struck-out examples in (a) IAM, (b) BH2M, (c) ICDAR segmentation contest, (d) CMATERdb, (e) MLS and (f) BFL database.

These database contained very few struck-out text, So we needed to generate a new database.

Generated Database

(1) Uncontrolled data: The writings were collected from the publicly available databases. This data is called uncontrolled, since we did not have any control on the data generation.

(2) Semi-controlled data: The writers were allowed to make extempore writings of their choice, but were informed that some struck-outs should be there in a written page.

(3) Controlled data: The writers were given a brief lesson of various form of strike-out strokes by showing some examples. They were also asked to draw as many different examples as possible from the six major types (single, multiple, slanted, crossed, zigzag and wavy) of strike-outs in their write-up.

Primary database details

Database type	Uncontrolled		Semi-controlled		Controlled		Total		
Script (# documents)	$\mathcal{E}$ † (50)	$\mathcal{B}$ † (50)	$\mathcal{E}$ (100)	$\mathcal{B}$ (100)	$\mathcal{E}$ (100)	$\mathcal{B}$ (100)	$\mathcal{E}$ (250)	$\mathcal{B}$ (250)	$\mathcal{E}+\mathcal{B}$ (500)
# Normal words	5321	6018	12116	12893	13233	13851	30670	32762	63432
# Struck-out words	147	168	556	538	692	726	1395	1432	2827
Strike-out rate (%)	2.68	2.71	4.38	4.00	4.96	4.98	4.35	4.18	4.26

† $\mathcal{E}$: English and $\mathcal{B}$: Bengali script.

Frequency of Various SS Types

Frequency of SS types in the Database

Frequency of SS types in the Database

Strike-out types	Number of struck-out words	
	English	Bengali
Single	497	532
Multiple	266	261
Slanted	262	245
Crossed	189	197
Zigzag	92	87
Wavy	57	71
Others	32	39
Total	1395	1432

Struck-out Text Detection Performance

Performance of Struck-out Word Detection

Script	English (Bengali)			
Word length in #characters/OSs*	Precision %	Recall %	F-Measure %	Balanced Accuracy %
1	86.94 (84.73)	88.38 (87.83)	87.65 (86.25)	92.53 (92.12)
2	87.67 (85.64)	88.92 (88.67)	88.29 (87.12)	93.04 (92.59)
3	89.58 (88.45)	89.86 (89.76)	89.71 (89.10)	93.91 (93.20)
4	93.21 (91.72)	94.16 (93.35)	93.68 (92.52)	95.54 (94.76)
5	93.80 (94.67)	94.48 (95.41)	94.13 (95.03)	96.37 (97.15)
More than 5	94.67 (95.92)	96.37 (96.86)	95.51 (96.38)	97.23 (98.27)
Overall	90.94 (90.19)	92.18 (91.94)	91.56 (91.06)	94.77 (94.68)

* OS: Ortho-Syllable.

Training set = 100 document pages
Testing set = 400 document pages

Balanced Accuracy = (True Positive Rate + True Negative Rate) / 2

Struck-out Text Detection Performance (Contd…)

Overall Performance of Feature Subsets for Struck-out Detection

Features	F-Measure (%)	
	English	Bengali
Branch point based (F_{BP}, F_{BPW}, F_{BPX})	88.56	89.22
Density based (F_D, F_{SD})	79.52	78.12
Hole based (F_H, F_{CS})	84.37	85.48
Branch point + Hole based features	89.42	89.78
All 7 features	91.56	91.06

Overall Struck-out Detection Performance by Various Classifiers

Classifiers	F-Measure (%)	
	English	Bengali
Min. Distance	76.09	73.93
K-NN	79.45	78.16
SVM-linear	81.51	79.84
MQDF	88.97	88.24
MLP	89.81	89.13
SVM-RBF	91.56	91.06

Struck-out Text Detection Performance (Contd…)

Performance of Struck-out Word Detection

	English (Bengali)		
Method	**Precision %**	**Recall %**	**F-Measure %**
Hand-crafted feature + SVM	90.94 (90.19)	92.18 (91.94)	91.56 (91.06)
CNN feature + SVM	97.63 (97.25)	98.08 (97.84)	97.85 (97.54)

Struck-out Stroke (SS) Localization

Performance of Locating Various Types of SS

Strike-out types	F-Measure (%)	
	English	Bengali
Single	97.32	96.21
Multiple	93.84	94.62
Slanted	94.46	93.04
Crossed	91.59	92.78
Zigzag	82.16	81.74
Wavy	78.58	77.49
Overall	89.65	89.31

TP: area of SS region recognized by our system,
FN: area of SS region which is not recognized,
FP: area of non-SS region recognized as SS region,
Precision (P) = $TP/(TP + FP)$, **Recall** (R) = $TP/(TP + FN)$,
F-Measure (FM) = $(2 \times P \times R)/(P + R)$.

Performance of Struck-out Stroke (SS) Cleaning

Performance Evaluation for Stroke Removal

Script	DR (%)	RA (%)	F-Measure (%)
English	92.83	89.56	91.16
Bengali	90.39	88.23	89.29

N: # black (object) pixels in the ground-truth version,
M: # black pixels on automatically cleaned output by our method,
O2O: # matching black pixels between ground-truth and our method output,
Detection Rate (**DR**) = O2O/N,
Recognition Accuracy (**RA**) = O2O/M,
F-Measure (**FM**) = (2 × DR × RA) / (DR + RA).

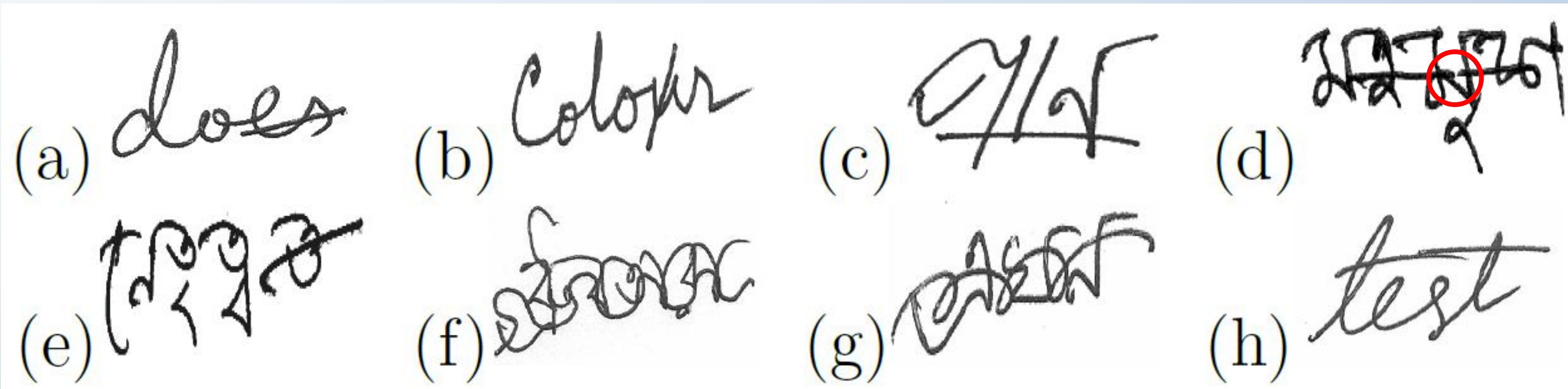
Overall Performance of Struck-out Text Processing

Relative Performance on 3 Database Types

	F-Measure (%)		
Script	English (Bengali)		
Database type	Struck-out word detection	SS* localization	SS* deletion
Uncontrolled	94.34 (93.76)	92.19 (92.33)	92.12 (91.35)
Semi-controlled	91.85 (91.37)	89.48 (89.07)	89.93 (89.17)
Controlled	88.48 (87.82)	87.12 (86.78)	88.15 (87.34)

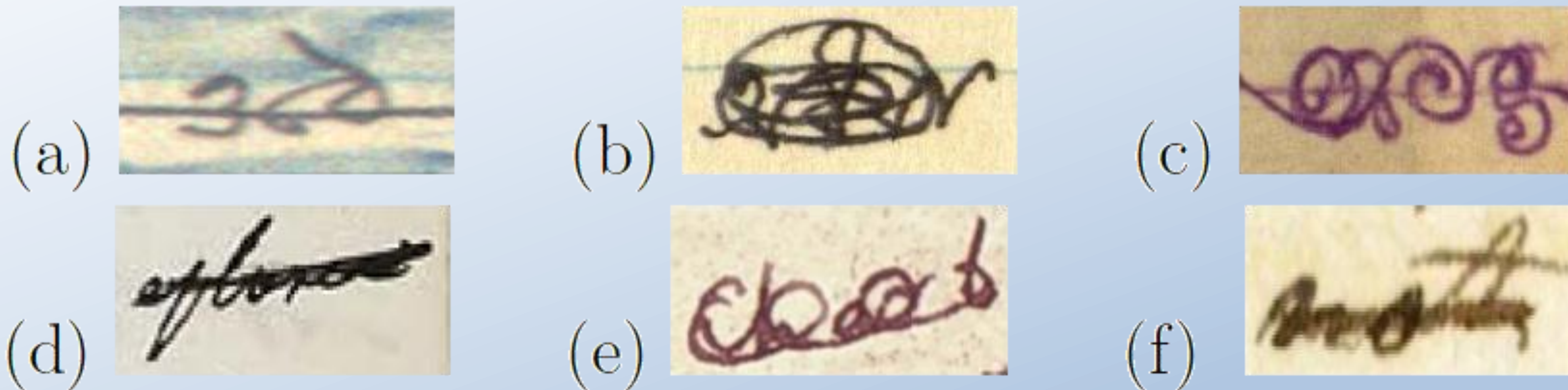
* SS: Strike-out Stroke.

Failure Instances on Contemporary Manuscript



(**a**), (**e**) Start/stop only in right region,
(**b**) SS height is less than main-body height,
(**c**) No hole generation,
(**d**) Pen lift in middle zone,
(**f**) Wavy with retrograding stroke,
(**g**) Spiral stroke,
(**h**) False positive.

Results on Historical Manuscripts



Instances on some manuscripts of (**a-c**) R. Tagore,
(**d**) G. Flaubert, (**e**) H. Balzac and (**f**) J. Austen.

D: Detection of struck-out text,
L: Identication/Localization of strike-out stroke.

(**a**) *D*:- Hit, *L*:- Hit, (**b**) *D*:- Hit, *L*:- Miss, (**c**) *D*:- Miss, *L*:- Miss,
(**d**) *D*:- Hit, *L*:- Hit, (**e**) *D*:- Hit, *L*:- Miss, (**f**) *D*:- Hit, *L*:- Miss.

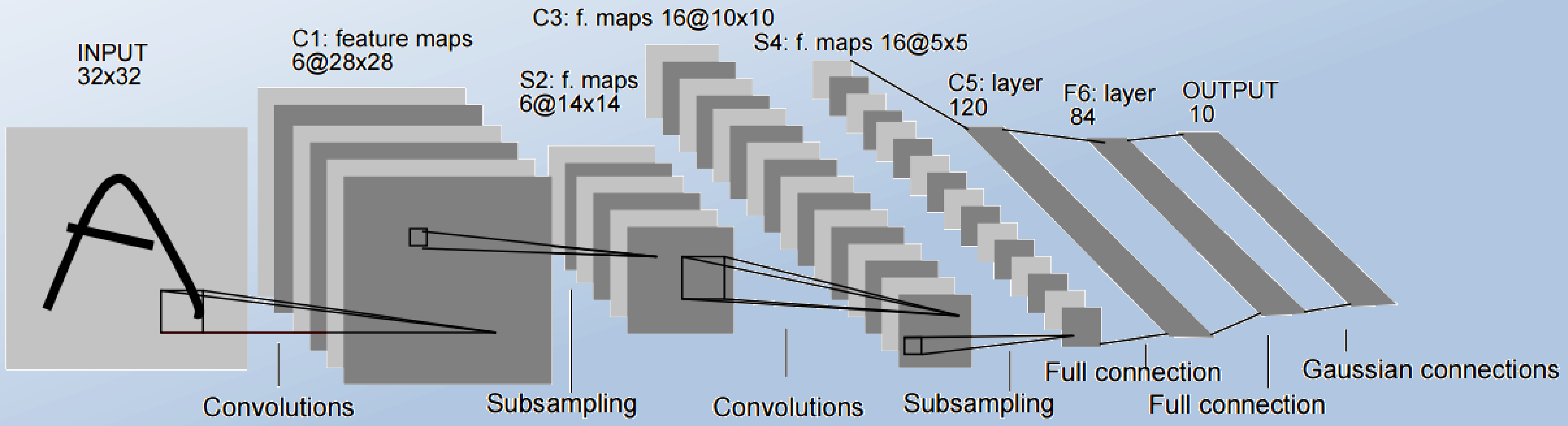
References

- J. Arlandis, J.C.P.-Cortes, J. Cano, “Rejection strategies and confidence measures for a K-NN classifier in an OCR task”, in: ICPR, vol. 1, 2002, pp. 576–579.
- L.L.-Sulem, A. Vinciarelli, “HMM-based offline recognition of handwritten words crossed out with different kinds of strokes”, in: ICFHR, 2008, pp. 70–75.
- S. Nicolas, T. Paquet, L. Heutte, “Markov random field models to extract the layout of complex handwritten documents”, in: IWFHR, 2006.
- J. Banerjee, A.M. Namboodiri, C.V. Jawahar, “Contextual restoration of severely degraded document images”, in: CVPR, 2009, pp. 517–524.
- A. Brink, H. van der Klauw, L. Schomaker, “Automatic removal of crossed-out handwritten text and the effect on writer verification and identification”, in: DRR, XV, 2008.
- D. Tuganbaev, D. Deriaguine, “Method of Stricken-out Character Recognition in Handwritten Text”, Patent US 8,472,719, 25 June 2013.
- B.B. Chaudhuri, C. Adak, “An Approach for Detecting and Cleaning of Struck-out Handwritten Text”, Pattern Recognition, vol.61, pp.282-294, January 2017.

Questions / Comments ?

Thank You

LeNet-5



* Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition", Proceedings of the IEEE, vol.86, no.11, pp.2278-2324, 1998.

Various Types of Insertions (Tagore)