Author Identification in Noisy Handwritten Documents

Karl Ni
Pat Callier
Brad Hatch

An IQT Lab. Private—Do not disseminate without approval of Lab41 or In-Q-Tel

Agenda

- Background
- State of the Art
- Contributions
- Results

Background

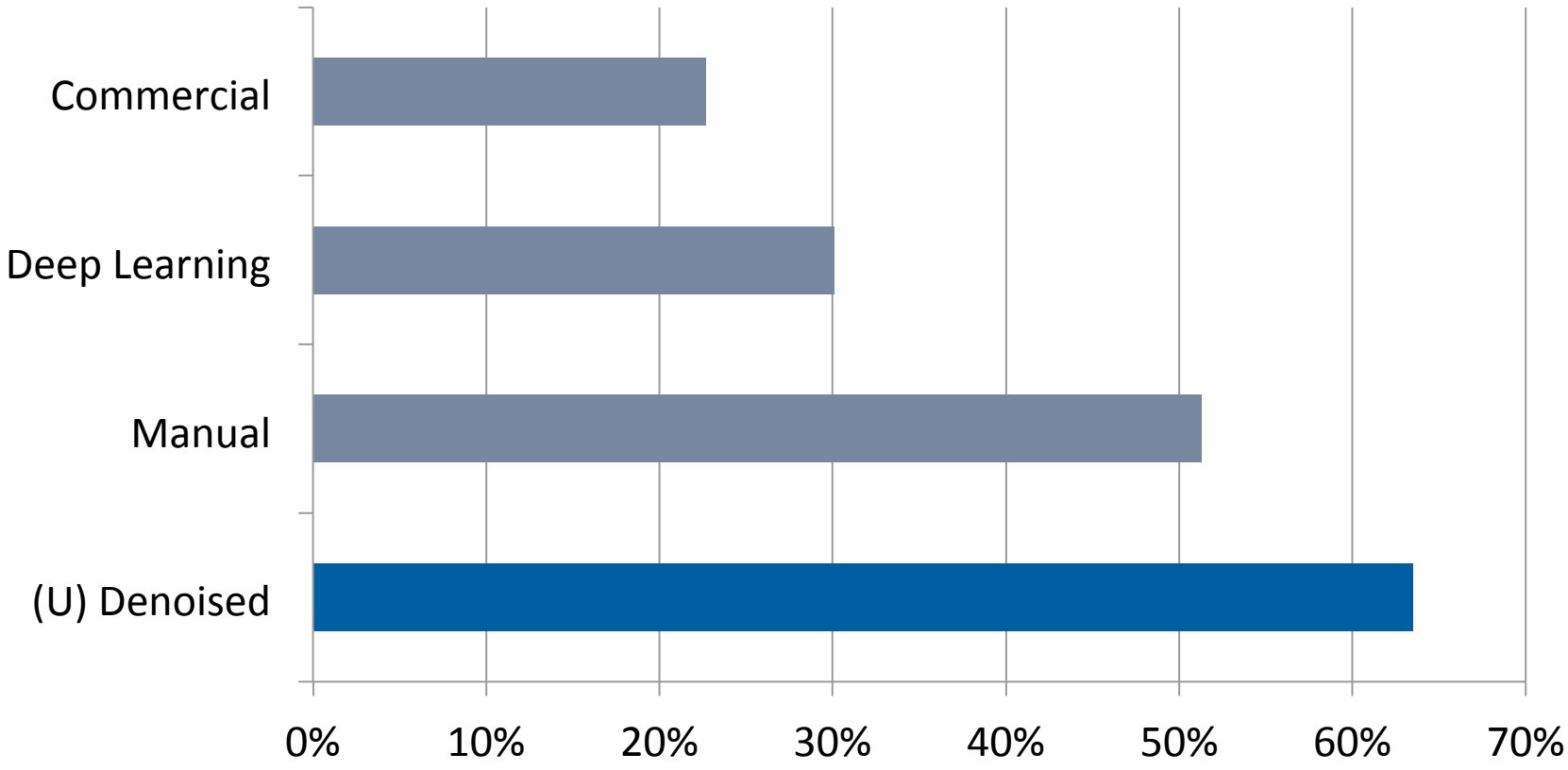
- Tradecraft: handwriting analysis for biometrics
  - Identify the author of a handwritten document
  - Subject matter expert workflow compares query to documents in reference database to assign authorship

- Current challenges:
  - Scalability: large quantities of reference documents collected over time. Reliant on human memory.
  - Limited resources: personnel time is valuable
  - Optimized workflow: attention to relevant documents
  - Noise: academic & commercial work focus on clean data

- Proposed solution:
  - For a given query, automatically rank reference database according to writing style similarity

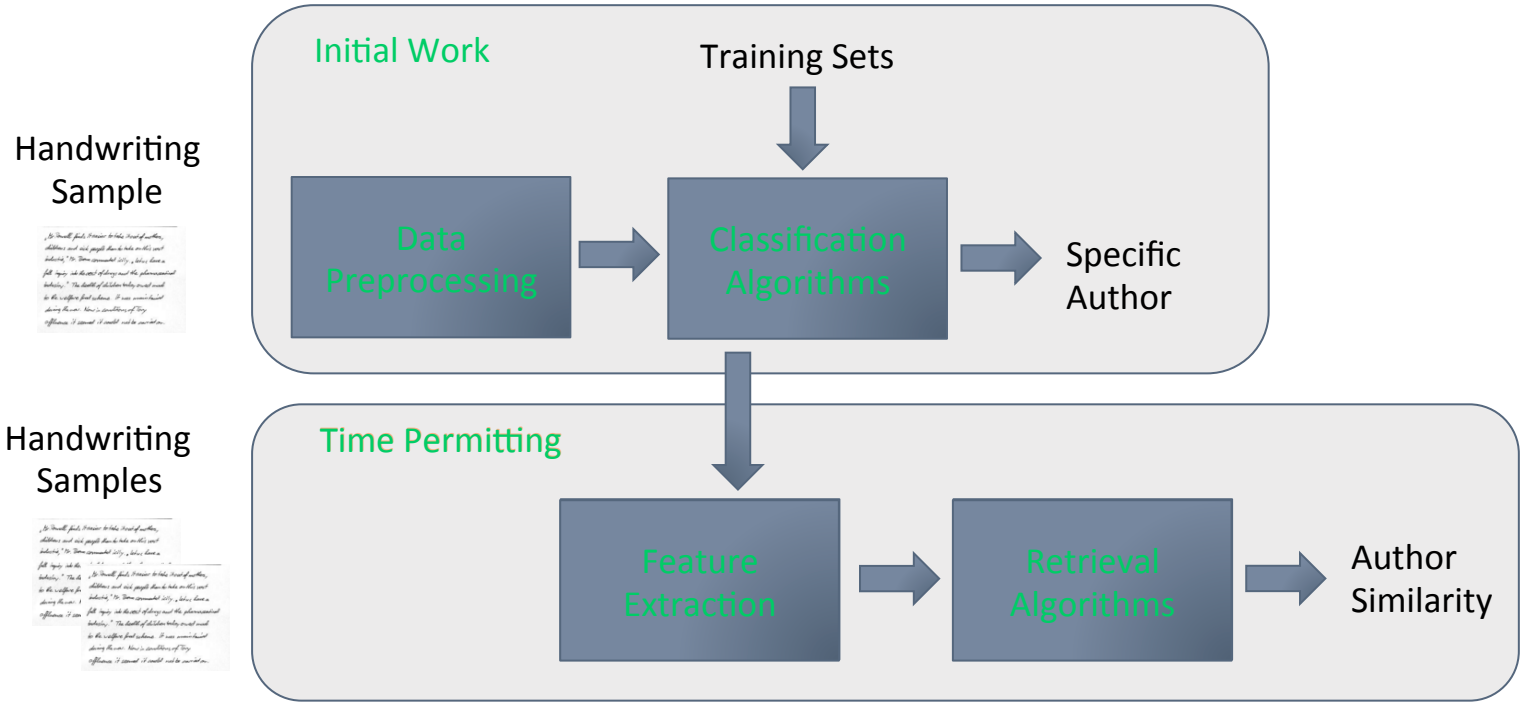
Preview of Results

Aggregate Top 1 Accuracy Rates



Scope

- Classifier focused versus feature focused:
  - Most CV apps: Train classifiers on known authors
  - Author ID: Rank document according to features



Agenda

- Background
- State of the Art
- Contributions
- Results

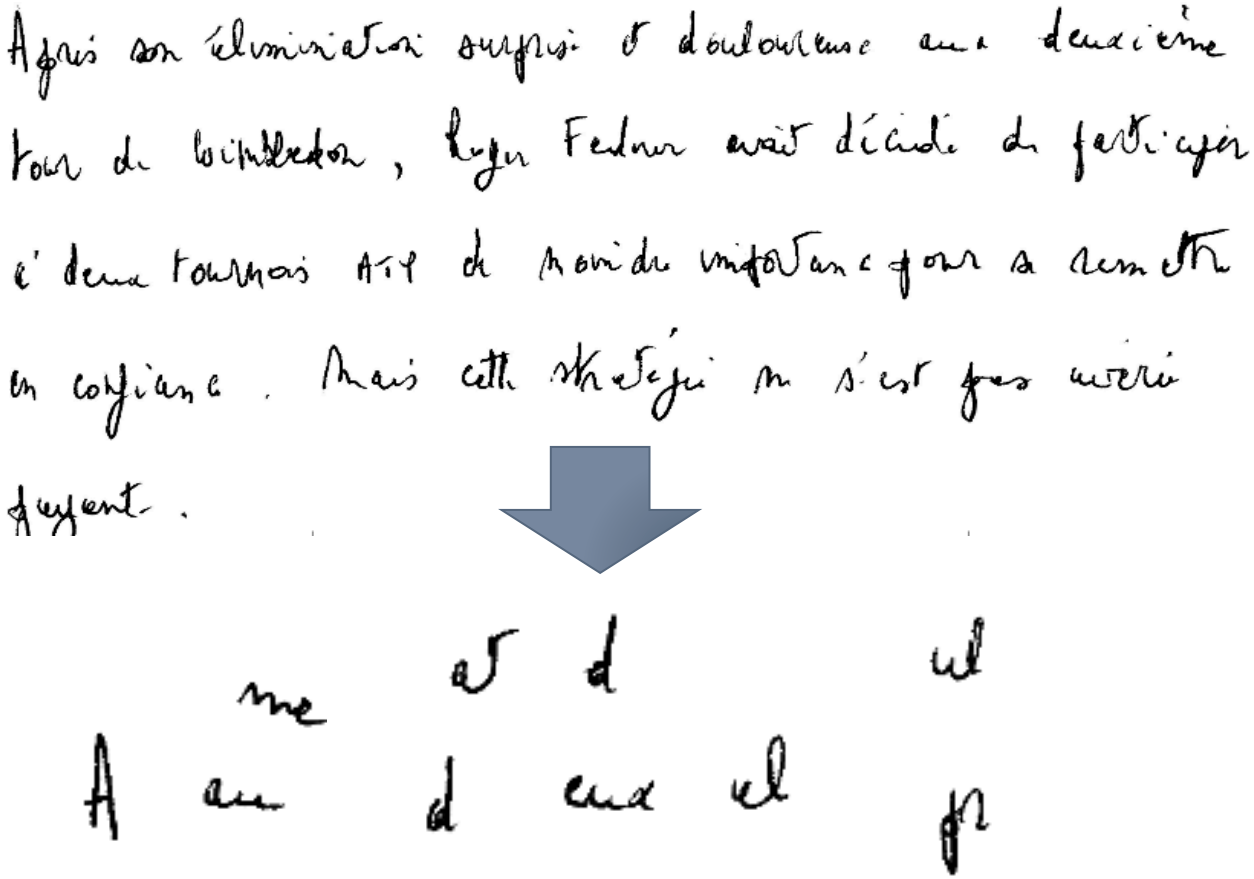
Related Work

- Related Work
  - Commercial Product Evaluated by Gov't Tech Lab 2015
  - Conference ICDAR competitions: 2011, 2013
  - Conference IAM competition: 2011
- State of the Art Investigation
  - Offline Writer ID (ICDAR'13): R. Jain & D. Doermann
  - CNN-DNN (2016): S. Fiel & R. Sablatnig
  - Encoder-Decoder Networks (CVPR, 2015)
- Investigative Research
  - Joint Optimization of Denoising CNNs
  - DRAW RNN Extensions to Writer Identification
  - Spatial Transformer Networks

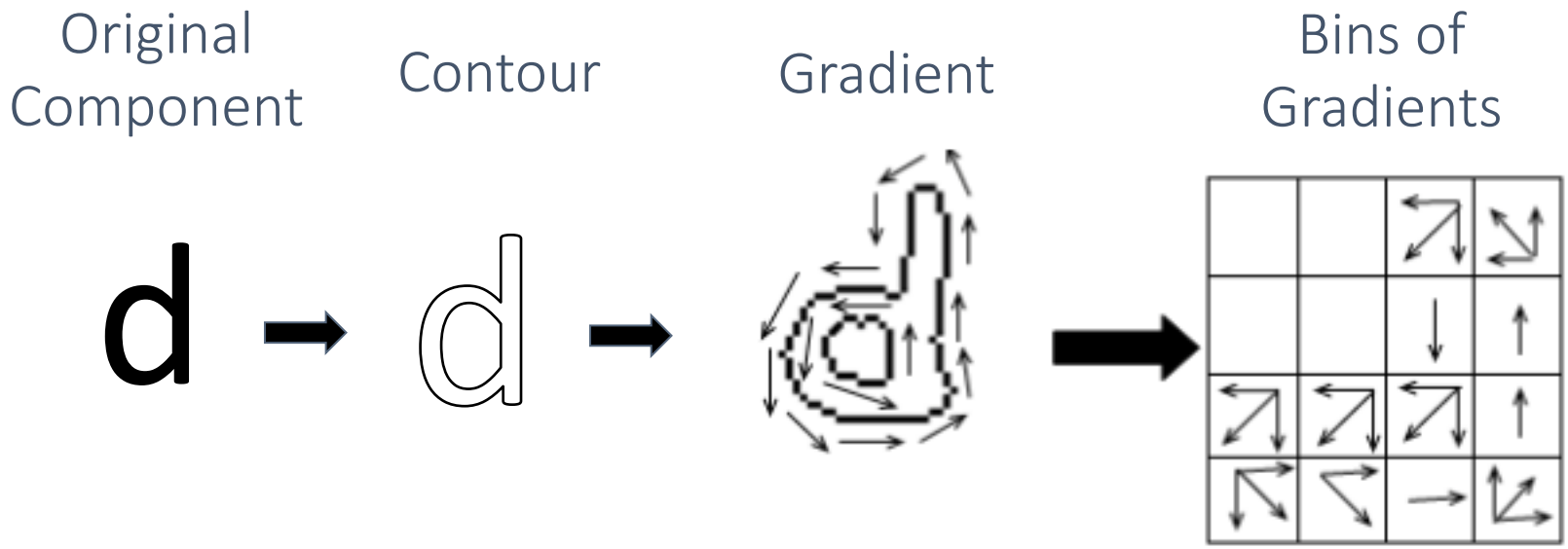
Connected Components

- Regions of contiguous marking
- Threshold by area, aspect ratio
- Letters, digraphs, words (50-300/page)

< 5px



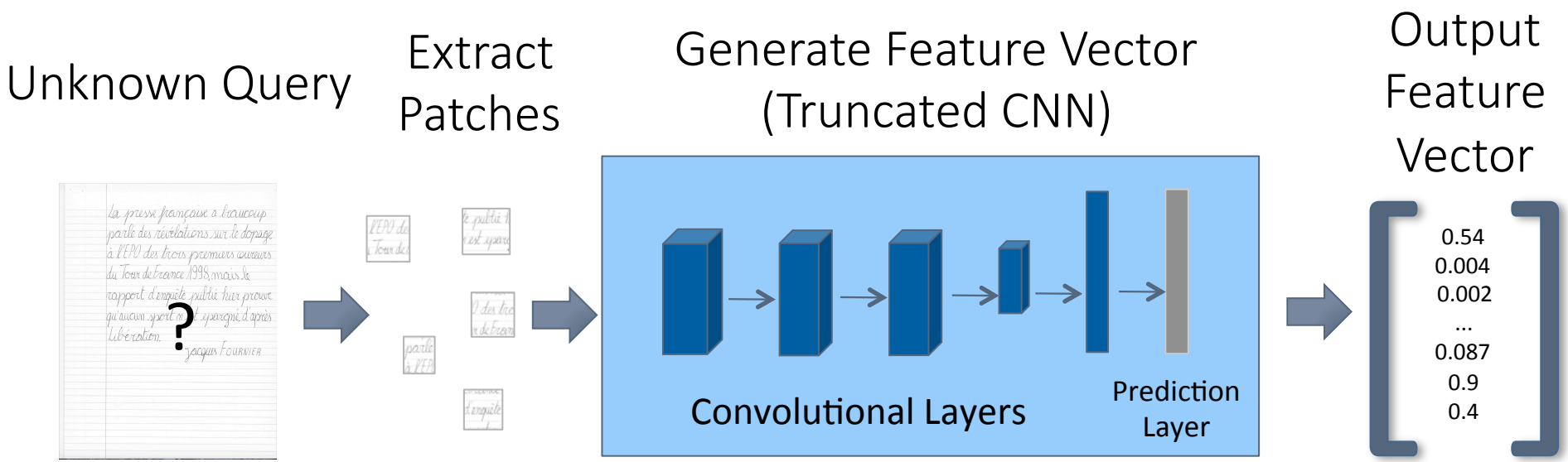
Features



Jain and Doermann 2013, Jain 2015

Patch CNNs

Use CNN to Process Query Document



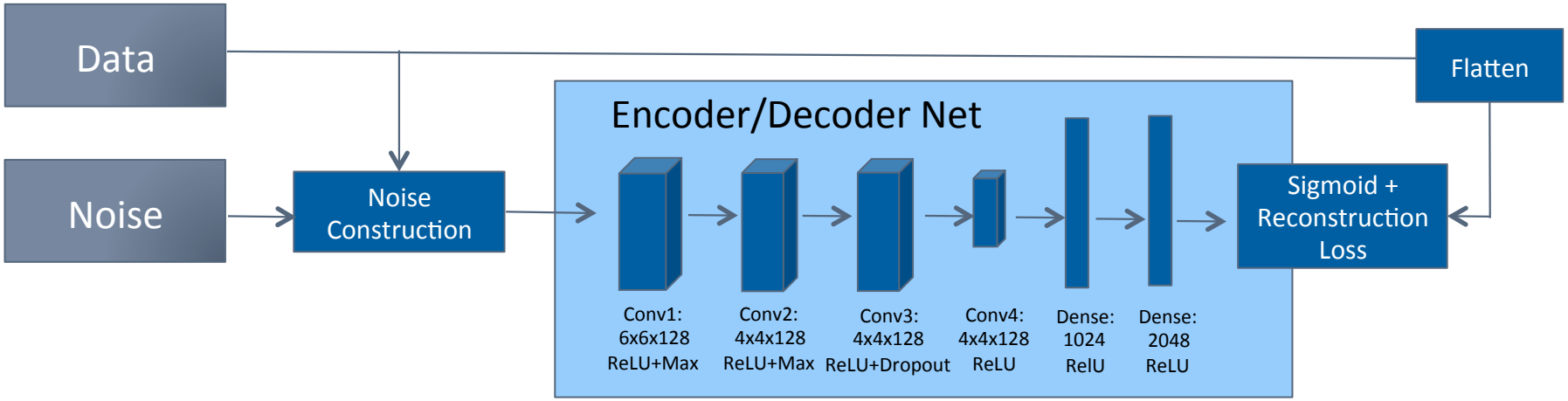
(➡ global classifier)

Agenda

- Background
- State of the Art
- Contributions
- Results

LiNet Denoiser

Denoising Binary Encoder-Decoder Convolutional Networks



Training LiNet

Training Steps:

1. Binary training data (IAM dataset)

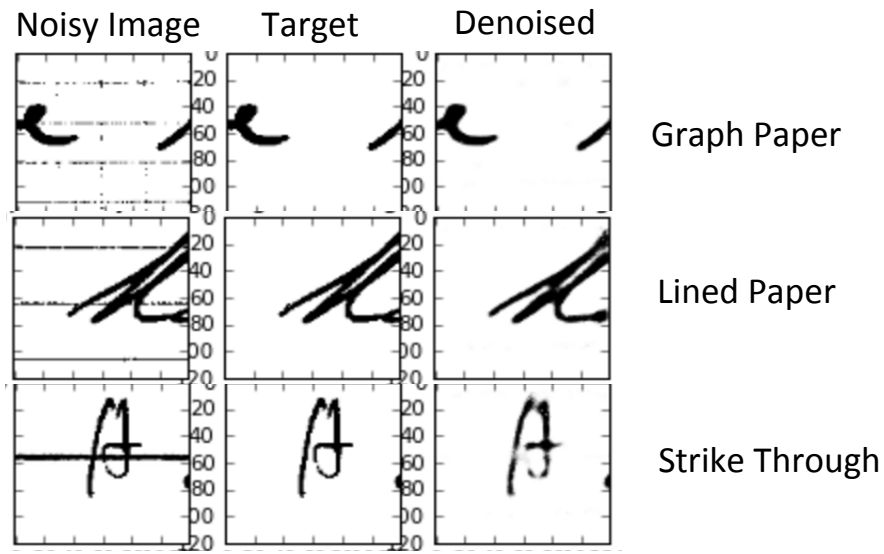
Random 56x56 patch	Random 120x120 patch

2. Assign random intensity

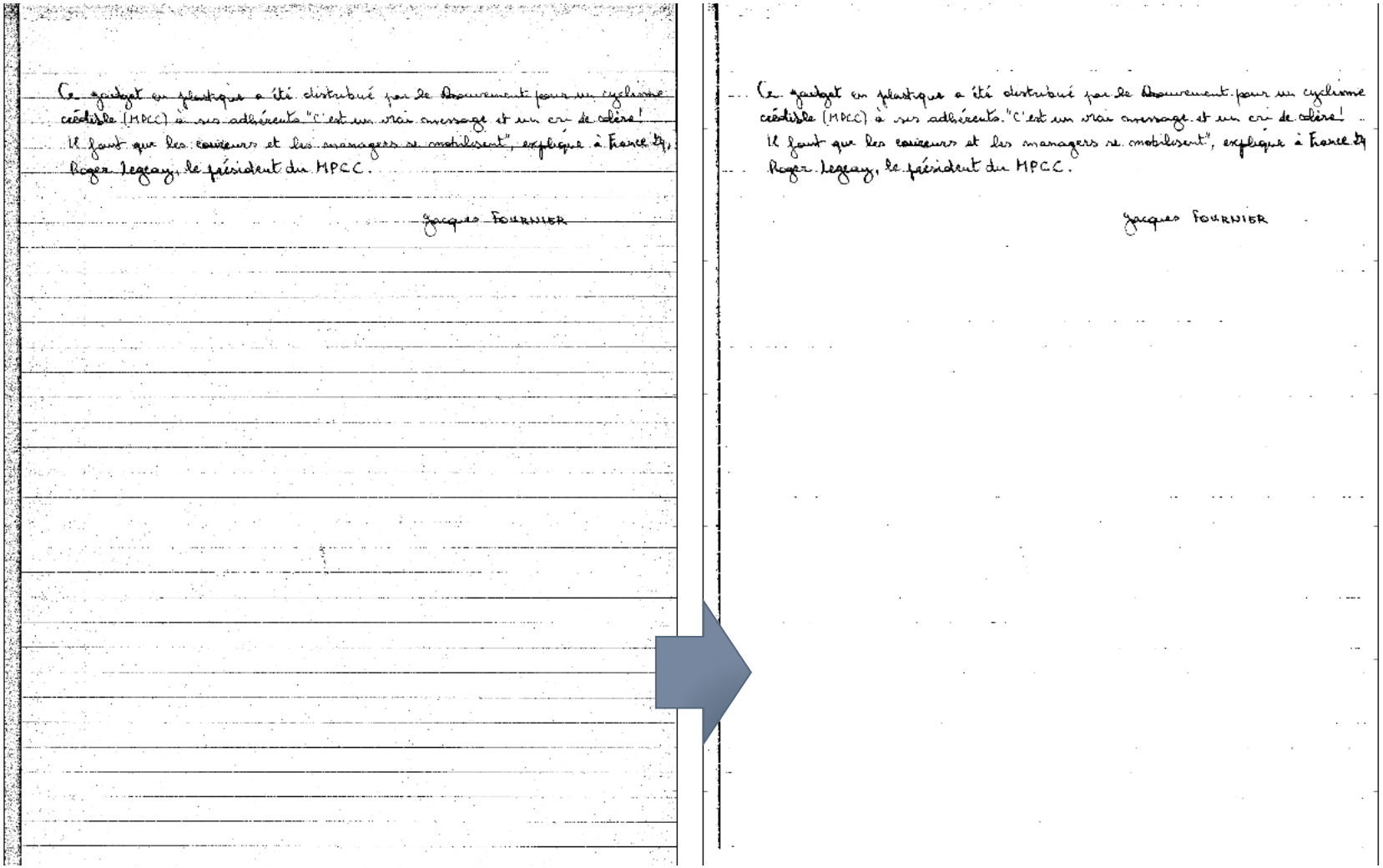
3. Add random noise

Graph paper	Lined paper	Coffee stains	Stamps	Etc.

4. Use binary data as a target



Denoised Images



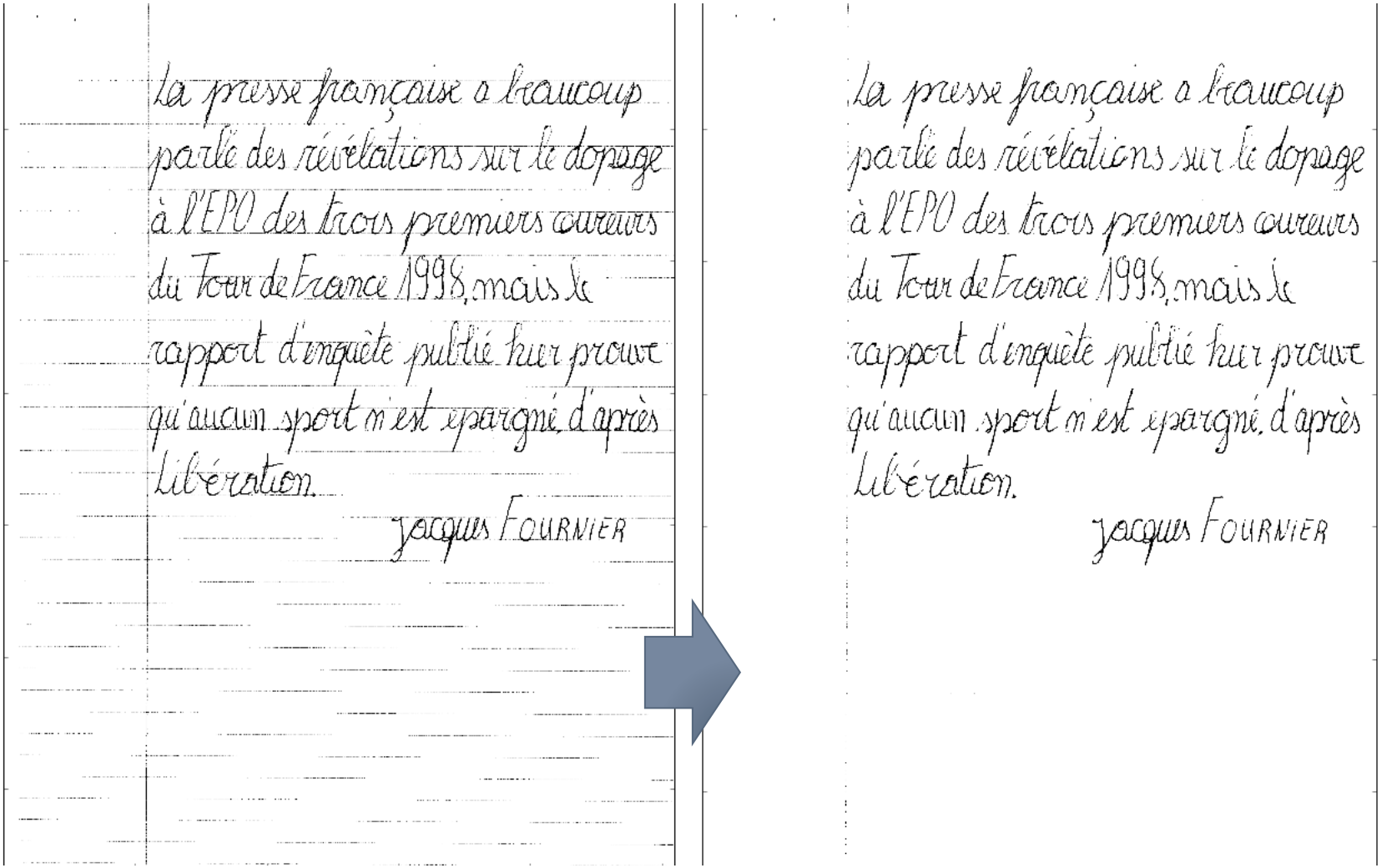
Ce gadget en plastique a été distribué par le Mouvement pour un cyclisme crédible (MPCC) à ses adhérents. "C'est un vrai message et un cri de colère! Il faut que les coureurs et les managers se mobilisent", explique à France 24, Roger Legeay, le président du MPCC.

Jacques FOURNIER

Ce gadget en plastique a été distribué par le Mouvement pour un cyclisme crédible (MPCC) à ses adhérents. "C'est un vrai message et un cri de colère! Il faut que les coureurs et les managers se mobilisent", explique à France 24, Roger Legeay, le président du MPCC.

Jacques FOURNIER

Denoised Images



Agenda

- Background
- State of the Art
- Contributions
- Results

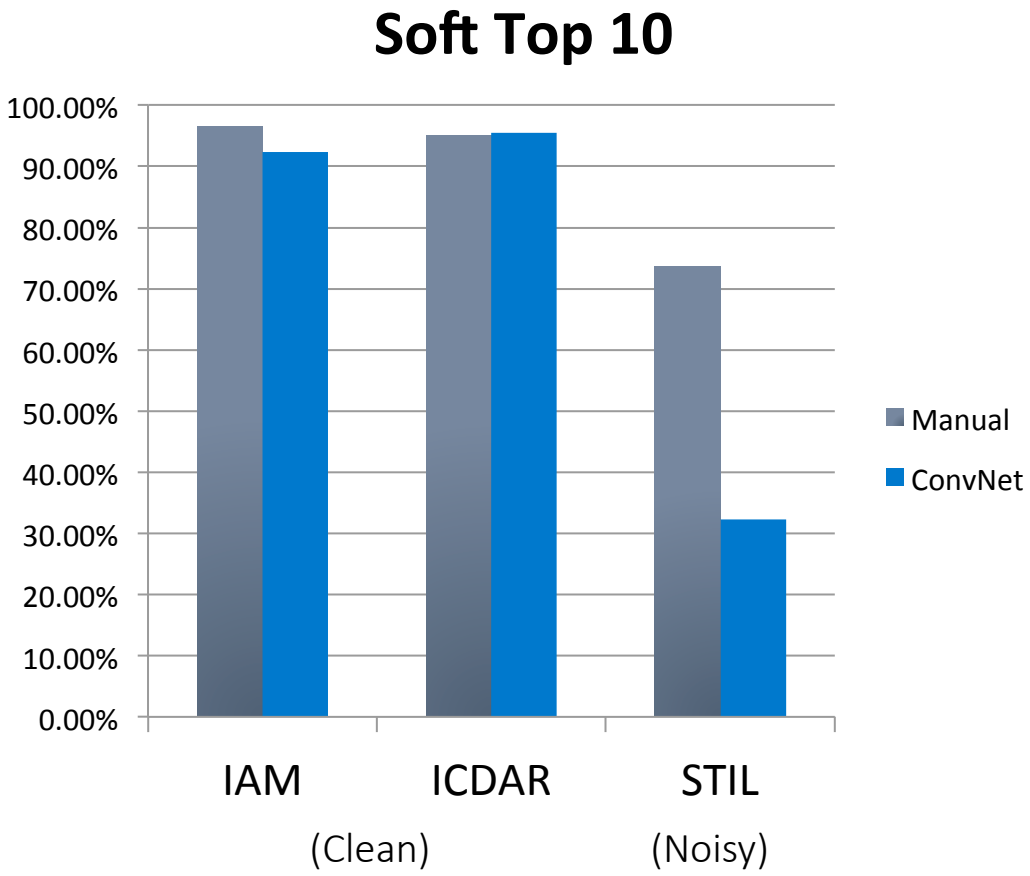
Datasets Used

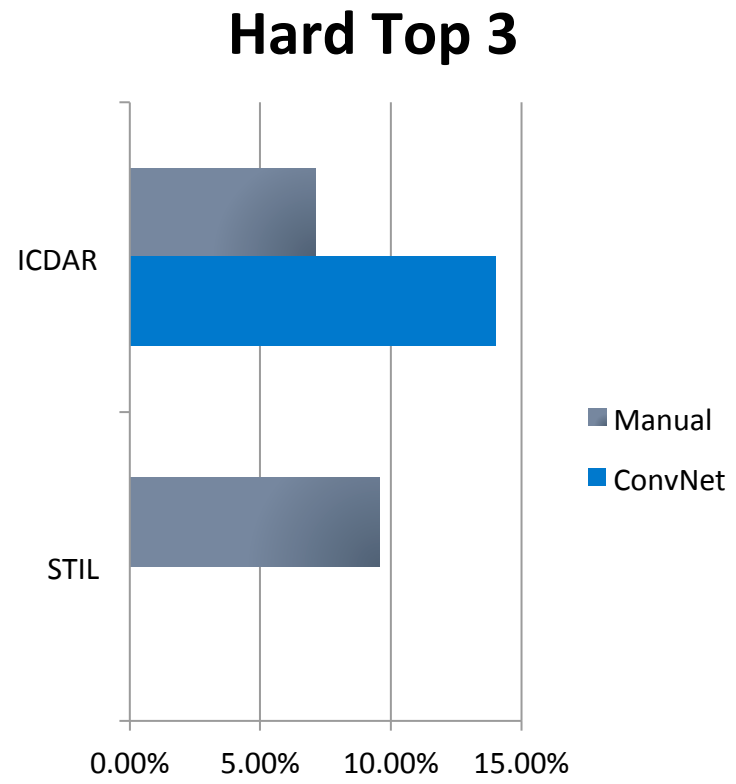
	IAM 2011 Writer ID	ICDAR 2013	NMEC Noisy Data
LANGUAGES	English	English/Greek	French/Arabic
DOCUMENTS	1539	Benchmarking: 1000 Experimental: 400	800 (French)
AUTHORS	657 (301 with multiple docs)	350 (4 Each)	100 (8 Each)
SCALE	Large	Mid-sized	Mid-Sized
BALANCED DATA		✓	✓
SEGMENTED	✓		
NOISY			✓

Metrics

- ICDAR 2013 evaluation metrics
  - Soft criterion
    - Retrieve top *k* most similar to query document
    - Any of the top *k* are the correct author?
  - Hard criterion
    - Are all of the top *k* from the same author?
    - *k* typically smaller than for soft criterion
- STIL evaluation metric
  - Top 1 accuracy
    - How often is the highest-ranking author the correct one?

Preliminary Results





CNNs do better on clean data (IAM/ICDAR)

Manual features do better on noisy data (STIL)

Results Preview

Top 1 Accuracy Rates by Guise

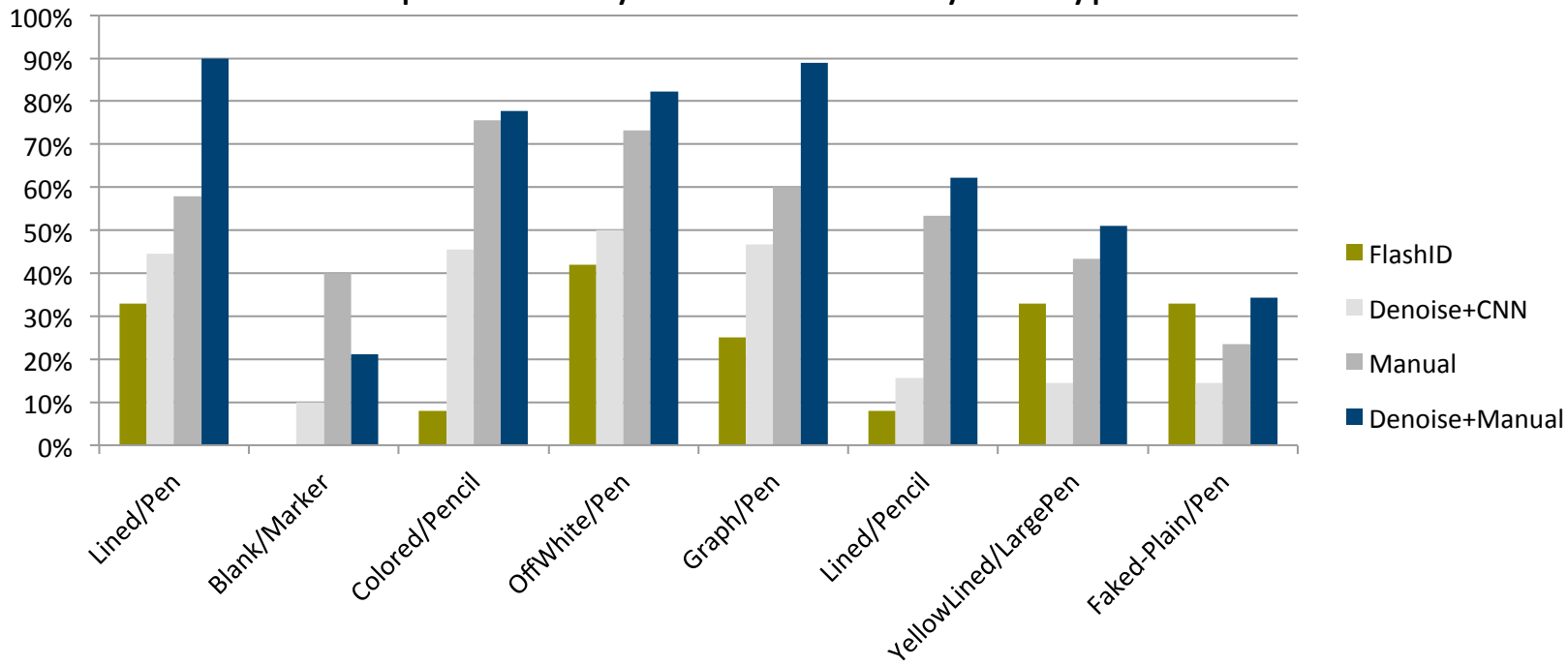
	Lined Pen	Blank Marker	Color Pencil	Off-white Pen	Graph Pen	Lined Pencil	Yellow Lined Pen	Fake Blank Pen
FlashID	33.0%	0.0%	8.0%	42.0%	25.0%	8.0%	33.0%	33.0%
CNN	44.4%	10.0%	45.6%	50.0%	46.7%	15.6%	14.4%	14.4%
Manual	51.1%	37.8%	72.2%	71.1%	62.2%	54.4%	38.9%	22.4%
Cleaned	90.0%	21.1%	77.8%	82.2%	88.9%	62.2%	51.1%	34.4%

Top 1 Accuracy

Aggregate top 1 accuracy

	Sciometrics FlashID	Denoised + CNN	Manual	Denoised + Manual
Top 1	**22.7%**	30.1%	53.3%	**63.5%**

Top 1 accuracy: broken down by doc type



(U) Denoising achieves considerable improvement except w/markers

An IQT Lab. Private—Do not disseminate without approval of Lab41 or In-Q-Tel

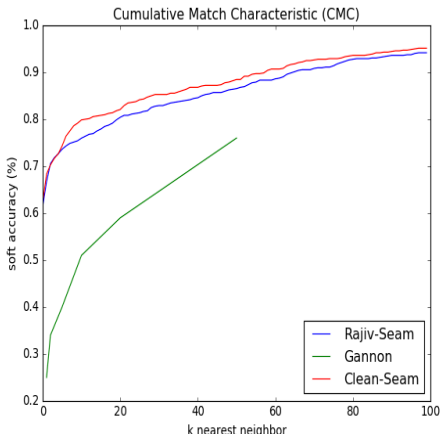
Questions?

Appendix

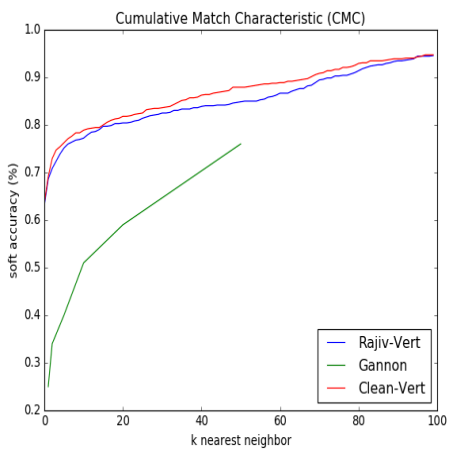
- Varying Top k Accuracy
- Categorized Top 1 Accuracy (All)
- Soft and Hard Top 10 (All)
- Analysis of Denoised Marker

Varying Top *k* Accuracy

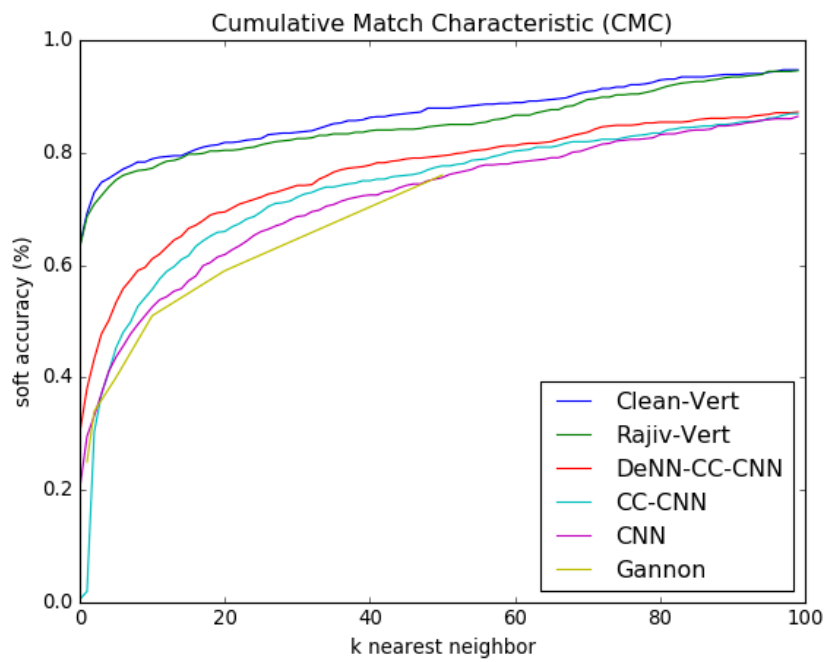
Seam Features



Vertical Features



Manual vs DNN vs Gannon

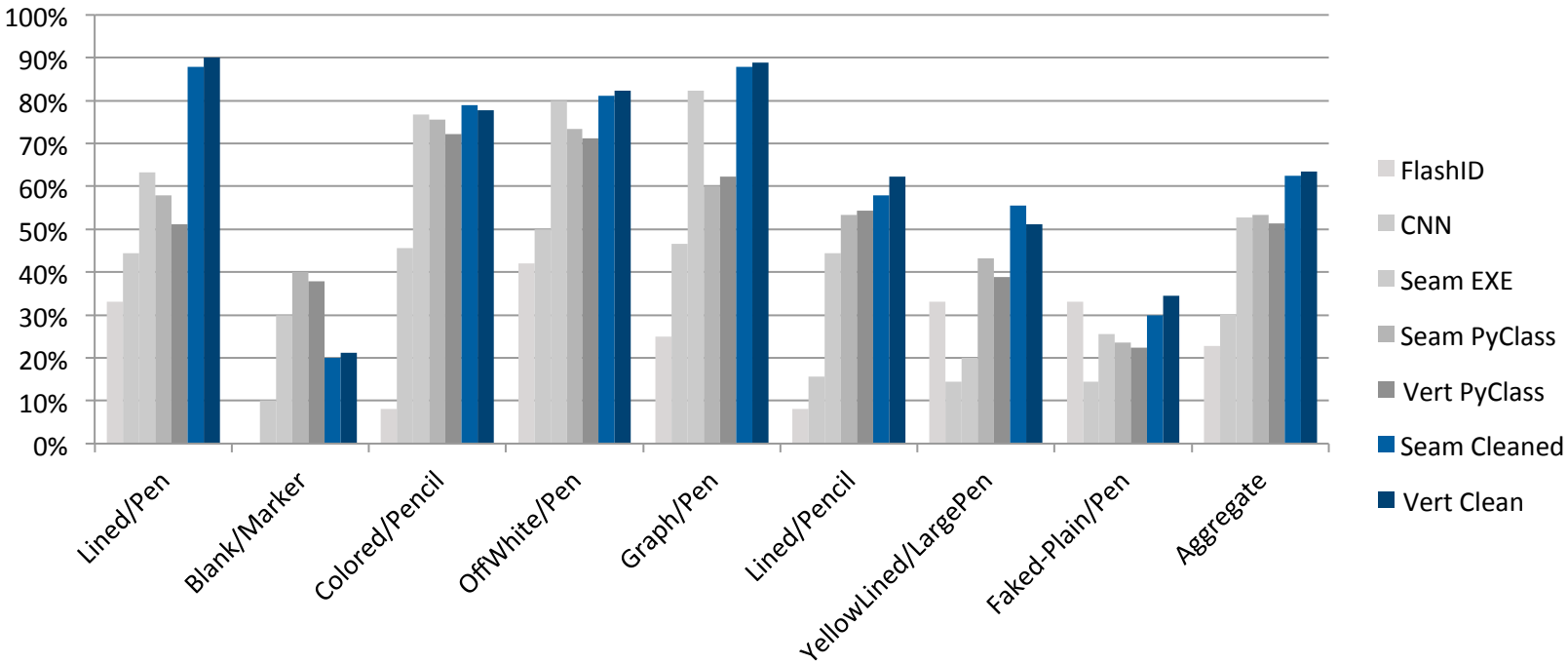


Logarithmic improvement over k

Performance exceeds that of current "state of the art"

An IQT Lab. Private—Do not disseminate without approval of Lab41 or In-Q-Tel

Categorized Top 1 Accuracy



	001	002	003	004	005	006	007	008
FlashID	33.%	0.0%	8%	42%	25%	8%	33%	33%
CNN	44.4%	10.0%	45.6%	50.0%	46.7%	15.6%	14.4%	14.4%
JainSC	63.3%	30.0%	76.7%	80%	82.2%	44.4%	20.0%	25.6%
JainS	57.8%	40.0%	75.6%	73.3%	60.0%	53.3%	43.3%	23.6%
CleanS	87.8%	20%	78.9%	81.1%	87.8%	57.8%	55.6%	30%
JainV	51.1%	37.8%	72.2%	71.1%	62.2%	54.4%	38.9%	22.4%
CleanV	90.0%	21.1%	77.8%	82.2%	88.9%	62.2%	51.1%	34.4%

Denoising achieves considerable improvement except w/markers

An IQT Lab. Private—Do not disseminate without approval of Lab41 or In-Q-Tel

Final Test Matrix

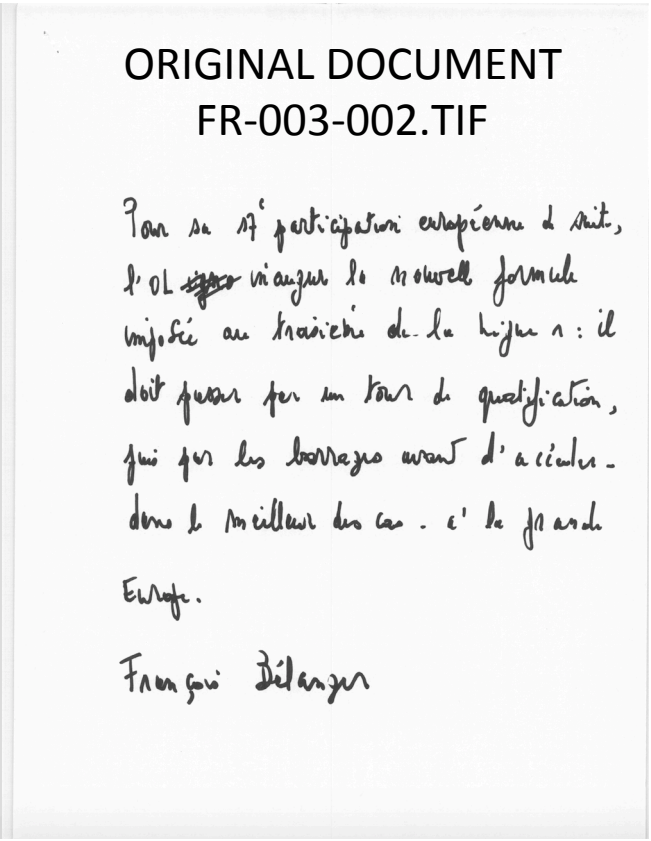
		IAM	ICDAR-BE		ICDAR-EX		NMEC	
		Top 10	Top 10	HARD	Top 10	HARD	Top 10	HARD
Manual Features	CC+Vert-Feats (Jain et al)	96.5	99.1	7.1	99.25	6.75	76.94	9.58
	CC+Seam-Feats (Jain et al)	95.4	98.6	8.4	99.25	7.75	75.42	11.6
	CC+TAS-Feats (Jain et al)	91.8	95.0	5.7	96.75	4.5	73.6	6.71
Deep Features	CNN (Modified Fiel et al)	92.4	95.5	13.8	96.5	13.7	32.3	0.0
	VarMean+CNN	Did Not Test					36.25	0.7
	CC+CNN						54.2	0.0
	DeNN+CC+Seam						79.2	40.4
	DeNN+CNN						53.1	2.4
	CC+DeNN-CNN						59.6	2.7

An IQT Lab. Private—Do not disseminate without approval of Lab41 or In-Q-Tel

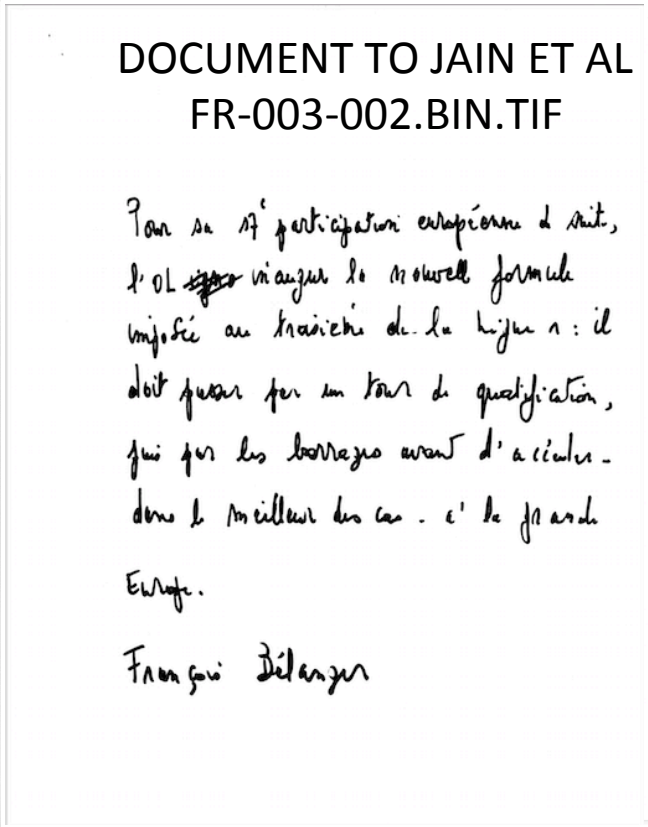
Analysis - Sharpies

- Documents 002:
  - Style: Normal, Size: Normal,
  - Paper: White plain, Instrument: Marker

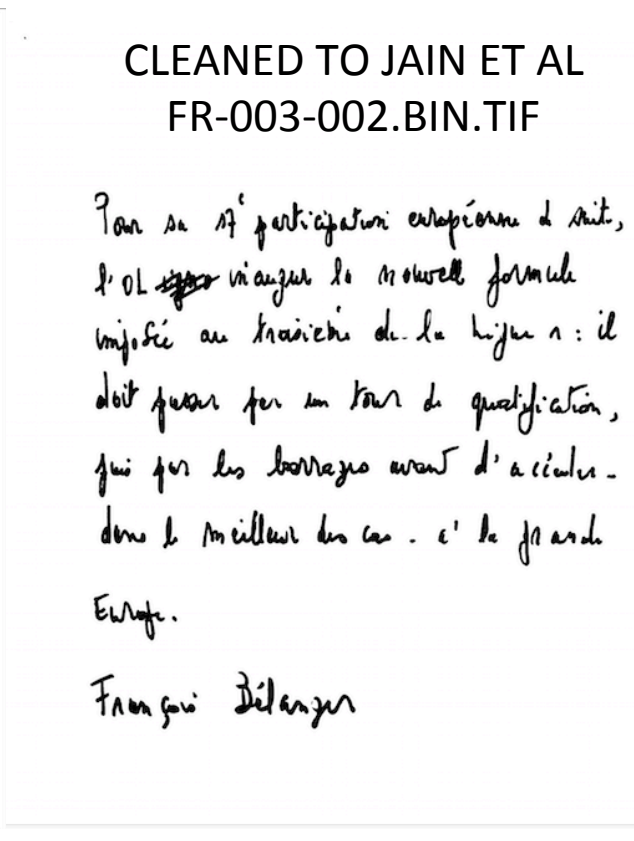
ORIGINAL DOCUMENT
FR-003-002.TIF



DOCUMENT TO JAIN ET AL
FR-003-002.BIN.TIF



CLEANED TO JAIN ET AL
FR-003-002.BIN.TIF



Performance drop mechanism unclear: require further investigation