

Economical Bimodal Classification of a Massive Heterogeneous Document Collection

Patrick Schone (patrickjohn.schone@familysearch.org) 24 February 2020



Overview

- Timelines (Lead-up)
- Description of the Collections
- Classification Goals for Automation
- Speed-focused System Architectures
- Performance and Outcomes



Timelines (Lead-up)

2015:

FamilySearch was able to auto-*index* 21M born-digital newspapers. Can auto-indexing work with born-paper? How about handwriting??

2016-2017:

FamilySearch & BYU collaborate on technologies to auto-transcribe HW.

2017-2018:

FamilySearch auto-*transcribed* about 33M newspaper stories and over 110M mostly-English handwritten & mixed documents with the goal of auto-*indexing* them.

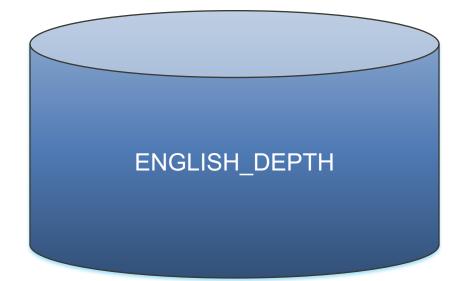
2019:

Newspaper going forward. But the massively-heterogeneous collection makes auto-*indexing* complex. Need to group & categorize documents, identify 'gotchas', and subdivide images.



Collections Of Interest

Two different, but related, kinds of corpora:



ENGLISH_BREADTH

163K Rolls of Film, every image [Abt 110M images] Represents **EVERY** *instance* of particular types of US Legal documents ~1M Rolls of Film, several ims/roll [Abt 3-4M images] Represents **EVERY** 'English' *roll*



Can We Classify After-the-Fact?

If we could **describe each image** of the Breadth/Depth corpora, we could target sub-collections for auto-indexing based on current capabilities & develop the capability for others.

Also, if we could **identify any anomalies**, that might help us do a better job handling them.

But we want to do this **quickly**! We want to finish in a week or so. But if we only took 1 sec/document (typical load time of a full image), it'd take [1.1 x 10⁸ images] x [1 sec/image] = 3.5 CPU years !



Classify: Semantic Categories

130+ Semantic Categories: What is the PURPOSE for the document?

NONE & COGNOME	NUM.	ANNO	REMITERO Registero	NOME IL COGNOMI	MUM. dulf Ufficio	ANNO	NUMB A English
Constitution of the	1.1	11.5	111	General Anto to		aus	- ce
Spinite autoriai		~	1.21			-	14
Contraction from al 20	1.00	-	101	Eposite Bernet	1000	-	10
Aplant Jakarthe	1000	-	115	Equal trigi		-	100
Ground Broger Frederickychicks			eer	Equili amiles		-	124
Grent Course		-	Rea	Syrand Martale	1253	-	10
Aprecia constra			24	Gund Herricanon			14
Apren Programin		-	200	Sprate Burden Manuel		-	e
Agreent finish Stream		~	1.24	Spread Break			120
ligned to be de	1.11	-	2.29	Agained and and the state	1	1905	1
Speech Branis		Nor	12	Egrand Jung is	1	-	11
Sprink for and		-	12	Topparist descension	1.50		1.12
Lation Man Star		1		Conte changes	11.2		-
Syrate Baliton	-		97	Const Grant			· ge
Eladel + transformer			_0,-	Grand James			1.00
Jana Aguine			en.	Elgens Burger		-	10
Superite formate			100	Const for constitution	-	-	100
Const. poplar			100	Syrand Bringen			14

Registration/Civil



General/Newspaper

STATE ON DAUG the No. 220 File I alian Distitut No. 74778 Regis errol in a heatidal or bathytice, give its water batend M Date Hoy 12 who End I Don't in Ma 793 Mayor of highry - Alacki Vital/Death/Legal **Probate/Will** Know all Men by these Presents Muoro all Mien by these Presents : a fill og lad on yn lin boudy of Bedown on a State of 8 Family/Pedigree Land/Deed



Classify: Layout Categories

~12 Layout Categories: What is the STRUCTURE of the document?

NOME & COSNOMR	NUM.	ANNO	Realized Register	NOME IL COGNOMI	MUM. dulf Ufficio	ANNO	NUMB A English
Specifi Goodda	22	cos.	141	General Anton to		aus	- 00
Genela quinai		~	121	Equip demains		-	14
Equipe frontes			101	Eponto - Bernardo		-	110
Carrier Jackwook			115	Equal danci		-	100
Ground Broger Frederic Alleria			er	Equal anides		-	104
Growthe Journal -		-	Rea	Syrand Markels		- 3	111
Agreed Courtes	1	~	24	Gund Herrinane.		-	14
Apress Provences		-	200	Lands durates Simon			e
Agreent former to		-	1.04	Square Hickory -			120
legans de ter Conta		~	Bar	Agende la de mais		1900	1
Specific descrite		de	12	Equard Jung is	1	-	11
Aquesta fire and			12	Elfant antenio	1.50		1.2
Later Landton			. es	Conte Stranger	11.2		-
Syrate Doctor .			40	Const Grant			12
Aladah a lang ian ti			-01	Guard Japan		-	100
land America	-		111	Elgenia Burger		-	110
Squart James .			100		-	-	100
Construction of the second			1.00	Agreech Bringen		~	1.90

Table/1 Line Per Row



IVIUITICOIUMN



Fill in the Blank

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS



Lenarie on Daug .

Graphical



7

Classify: Story Count

~12 Story Classification: How many unique 'stories' are in the document?

NONE & CORNONE	SUM.	ANDO	aconteo a Nagotre	NOME II COGNOMI	MUM. dali URMA	ANNO	MUMIN A English
Spirit Gooden	22	11.5	111	General Anto to		cere	ce
Squale quina?		~	1.21	Equile Security		-	ia.
Course from 12.			101	Eponto Derrorato		-	110
Calut Jackwook			115	Equal daily:		-	
Correction Broger Friend adjoint			eer	Equili amiles		-	dan
Great Junior		-	. Rea	Syrand Martale		- 3	
Apresit Courtes		~	24	Course Herrison			14
Apress Decorrise		-	200	Land and and	1		e
Agreed for and		~	1.04	Constant discovery -	1.00		i.
lande de Corte		~	2.29	Agained and and the state		-198	1
Specific Marine		de	12	Egrand Jung is			11
Sprink for and		-	12	Effect descurie			1.2
A plante allow Star		-		Gunt than go			-
Sprate Colling			er.	Constraint of the second			12
Spitel Mary Card			_0,-	Equal Stranger			1.00
Sterner Barris	-	-	111	Elguna Constantino	1.1	-	10
Squart formates			100	Comprender for security	-	-	120
Const. profile and			1.00	Syrand Harman			14

Story=1n



Story=many

" In The Roberts Cant - 2. Hove That m the 17 day of July 1908 . La Tilly of Staffing Cant of a State frank & P. Com I ging and leg-My children T. C. Terrenter and Warres & Sher and my estate, after finging my soil of Story=E&S Know all Mien by these Presents : Know all Men by these Presents : artenary of and fifty and and Small when hand norder . also all and the same hand that late barbard alicher B. Welch by Hours Verbith, said and is

State of Streams,

Story=2

with my adult office more the come

Hun had at a press mille

) a Similar no 9° and defel

DEPARTMENT OF CON DEP ACE OF DEATH red No. 15.6 Her long in U. S., If if beins bids? Did Dacrassed Set U. S. Navy or NAME ARFRED SCHWAKE ence. No.... (Used place of about RETIRED FORMER En ILM CHAGEN IN No. 793 Janget HOUER IN MARLOR Story=1 水禄元年 猫 益候

Lenarie on Daug .

Story=0p





Classify: Language Info

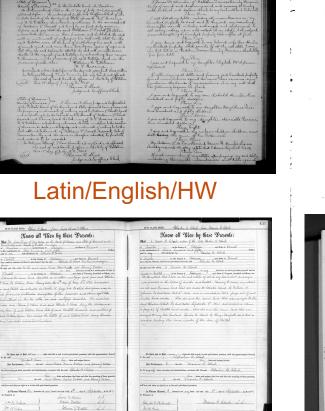
Linguistics: What are the Unicode scripts, language, countries, writing style?



Latin/Italian/MX



Latin/Spanish/PR



Latin/English/MX

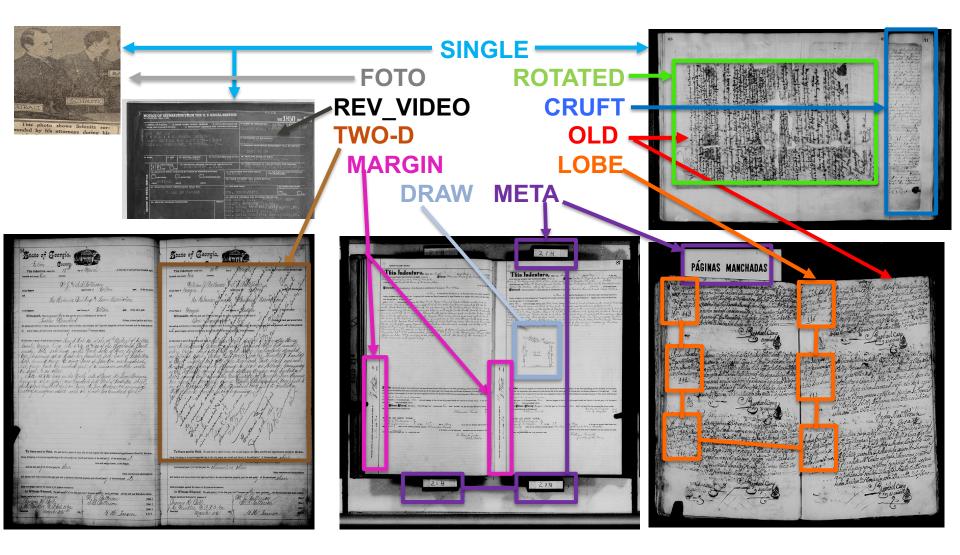
	STATE OF BUOK ST
a a dimensional sur-	
N 2	
	U. S. DEPARTMENT OF COMMENT REPAIL OF COMMENT STATE OF OHIO
	DEPARTMENT OF HEALTH
and a state of the	1 PLACE OF DEATH CERTIFICATE OF DEATH COLLECTED FILE No. (5711)
1:00	Township APR M3 Primary Registeration Distifut No. 1478 Registered No. 15
164 8 4 12	or Villago
1. A final statistical stat	as City of
PRYSECULO PRYSECULOS RECORD. En TRECORD. En	Length of millions in othy or Each when faith entered
13","TZKAKKERT" MEGG ad EXACTLT, PRYNICLA IN PERMANENT RECORD. EXACTLN, PRYNCLANS EXACTLN, PRYNCLANS EXACTLN, PRYNCLANS	RE Ward
SE 22 22 10	(Out dat a cost)
18 C C C S	2. BAK 14. COLOR. 3. BURGLE, MARRIED, Will the word 11 DATE OF DEATH (must ber, and part) How 7
FERMANDER	
1933111	The Wanted Without a Diame
A REAL	
A Part of the A	7. AOE (reat) Mantha Days 1 LASS than 1 day hrs. The PEDROTAL CAURE OF DEATH and related enters of large
(* arkivano) interministis 15. Y Yakukuka negalak AKB akada te mari EXUCTUT. Englank AKB akada te mari EXUCTUT. Englank AKB akada te mari EXUCTUT. Fa fact. AKB akada te mari ExUCTUT. Fa fact. AKB akada te mari Exuctur. Fa ar a negatar akama te mari ekster aka aka te negatar aka te mari ekster aka te mari ekster aka aka te negatar aka te mari ekster aka te mari ekster aka aka te negatar aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te mari ekster aka te te te mari ekster aka te mari ekster aka te	
10 INN-11 18 Abdd b 18 Abdd b 18 Abdr b 18 Abdr b 18 Abdr	Control of the design of the second s
10 10 10 10 10 10 10 10 10 10 10 10 10 1	. totation a battom in the Course p Alomach
20 A 4 4 5	A Contraction of the set of
A OB A OB A OB	0 THE STATE
ar university" r regelet. AGE a Distancial AGE a patient. AGE area patient. AGE area best a new be p	19. permitting (in a was) Germany Clearning of
191	
27- wited "uthation" Internation controls anglish AGE wheel a with unservice and the second to whether angles AGE wheel a we when a the angle and the second to them, a the angle at reviews.	To manife an a month and a mon
ALL	11. MAIDEN HAME Loglia muger 13. If doub on due to estave runne (visione) at to also
a diraction of the second of t	15. MATCHIN HAME & Office Multiple 16. MATCHIN HAME & Office Multiple 17. Matchine & South Control Forder 18. Matchine & Southal Forder 19. Matchine &
about the second	In the marie marie Canoning
1.5 1.5 1.5	
A DEALE STATE	The BURLY CREATING OF BRIDGER NOV IF who have a body
1 S 18.	The provide the state of the second state of t
1 A	MAPOLEON OHIO IN MAPLE
3 ×	11. PRED 11-11. 11/2 Anthregadaren ton 11/1 1 110 cum Redulation
13.2	Concerning out provident in an and particular second second second second
2423	
	Latin/English/MX

Chinese/Japanese/HP

FAMILYSEARCH



Anomalies: Binary Properties



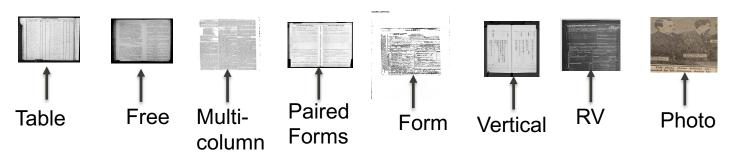


Speedy Classification?

One Option: Use **thumbnail images** and do image-level classification.

Definite 'Wins':

- FamilySearch automatically stores 200x200 thumbnails of each image.
- Thumbnails for an entire roll of film (1000 images) occupy about the same storage space as 3 images [so, over 99% compression].
- Since these are small, load time and subsequent processing time is short.
- Can see color, periphery, two-up-ness, photos, & line patterns



Drawbacks:

 Their amount of detail is limited, so it's hard to assess the true semantics. Have to guess the semantics based on 'this is a paired form, and that's what deeds look like, so I'll guess it's a deed."



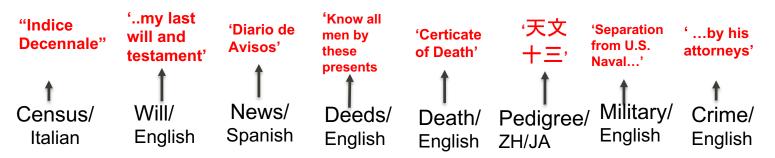


Speedy Classification?

<u>Another Option</u>: Use transcripts with bounding boxes & do text-level classification.

Definite Wins:

- Processing transcript is orders of magnitude faster than thumbnails.
- Semantic information is often very clear at the textual level.
- Language, script, country, writing style should all be straightforward to note.



Serious Drawbacks:

- Color is gone; borders are likely gone; photos are gone. How can one even tell if an image was reverse video if all you have is the transcript? How can you tell if it was complicated form or if it was nicely laid out?
- One needs to have the transcripts already.



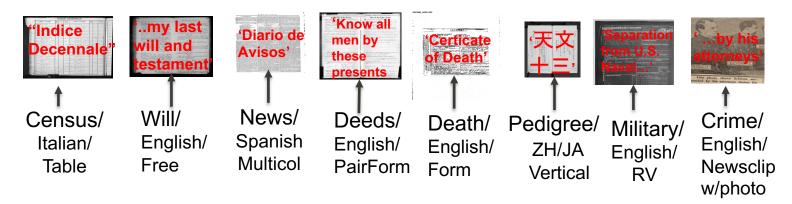


Speedy Classification?

BEST Option: Use **BOTH** snapshots AND transcripts+bounding boxes.

Definite Wins:

- Get the best of both worlds: semantics from text, visuals from thumbnail.
- Not much more expensive than JUST thumbnails when using both.
- Can toggle and use text-based or image-based models if that's all one has.

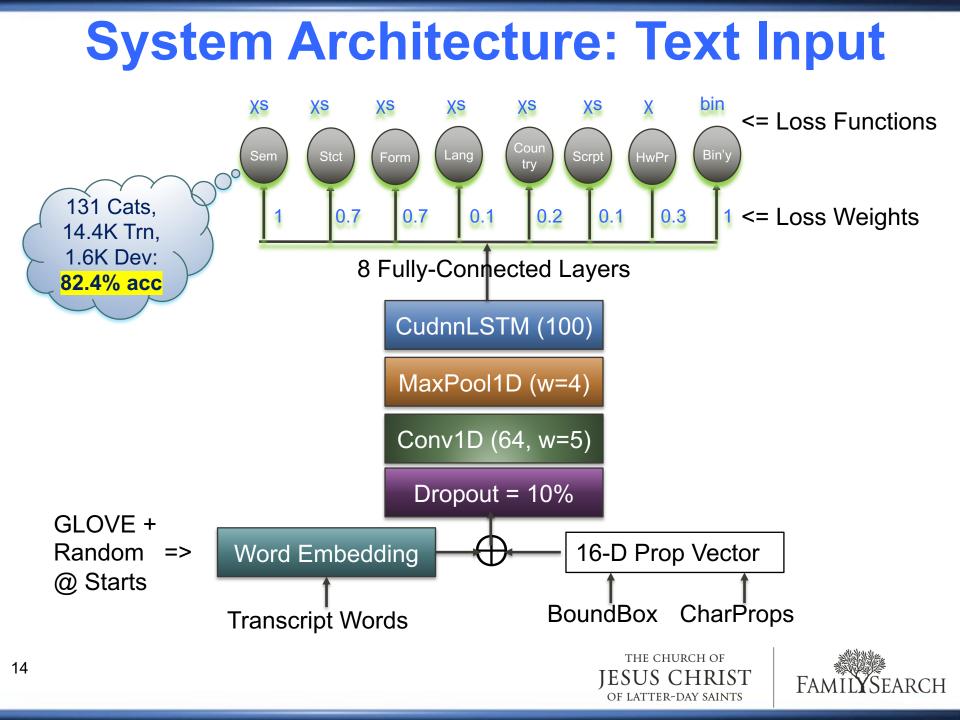


Drawbacks:

• Model management is slightly more complex.





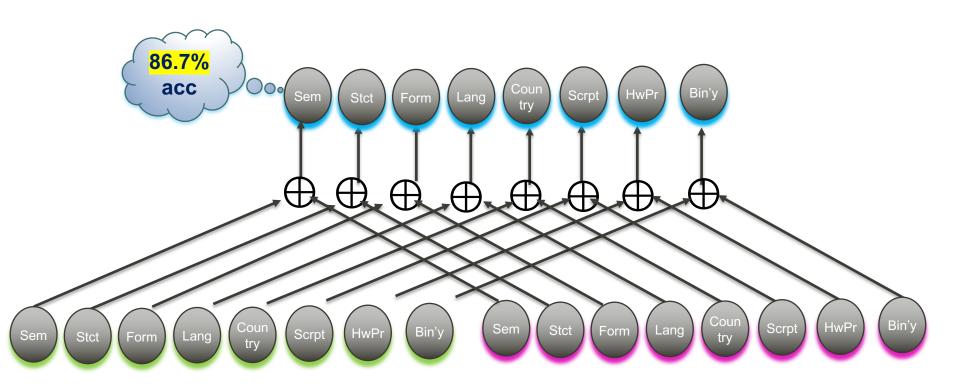


System Design: Image Input bin χs χs χs χs χs χs 82.1% Coun acc Lang Sem Stct HwPr Bin'y Form Scrpt trv 0.7 0.7 0.2 0.3 0.1 0.1 EfficientNet [M. Tan, Q. Le, 2019] 8 Fully-Connected Layers #Param #Flops Net **xVersus** Flatten 9% (ResNet50) 5.3M 0.39B **B**0 Dropout (20%) 12% (Incpt'nV3) 0.70B 7.8M **B1** 7x7 2D MaxPool 1.0 B 7.6% (Incpt'nV4) **B**2 9.2M **B**3 12M 1.8 B 5.6% (ResNxt50) 19M **B4** 4.2 B 18%(AmoebaNtA) **Top-Removed** EfficientNet/B1 **B**5 30M 9.9 B 24%(AmoebaNtC) **B6** 43M 19 B 224 x 200 x **B7** 66M 37 B 200 224 Results reported by Tan&Le. THE CHURCH OF **IESUS CHRIST**

FAMILYSEARCH

OF LATTER-DAY SAINTS

System Design: Fused Input



For fully-connected weights at start, assume near-50% weights for class C from text(or image) going to class C in final, and near-zero weights for all other connections.



Outcomes: Timings 115,973,482 Images

Ran TWO trials. First was TEXT ONLY, second was FULL.

TextOnly:

Ran on one box (Dual-Gpu System). Three jobs/Gpu (but lock around Gpu process) Took 3.5 days.

FullSystem:

Re-Ran on 3 diff't machines, with variable number of Gpus. But would have taken ~20 days on system of 'TextOnly' (with bulk of the additional cost going to thumbnail processing).



Outcomes: Results 115,973,482 Images

Semantics	%
Deeds	52.6
Land Index	11.6
Gen.Legal	8.3
Gen.Probate	5.6
Will	4.0
Inventory	3.4
Recpt/Check	1.1

Layouts	%	Recording	%
Freeform	68.1	Handwrit'n	59.1
Fill-in	18.2	Mixed	22.0
Table/1line	10.4	PrintOnly	18.3
Form	1.7	Blank	0.7

Anomalies	%
One-ups	52.4
Old (<1800)	3.7
HasMeta	2.0
HasLobes	1.5
ReverseVid	0.6
BleedThru	0.5

#Stories	%
Exactly 1	35.0
EndOrStrt	19.3
>1, but <2	9.3
End&Start	8.4
1-∞ Index	7.7
Exactly 2	7.2
Many	7.0





Summary

- Identified deep neural networks to mine text and image content, with sparse network combiner
- 86.7% acc on 131 category determination, plus generates multiple other kinds of classifications simultaneously
- Demonstrated result on large collection of >110 images

QUESTIONS?



