





Lightning-Fast Visual Language Identification

Patrick Schone & *Elder* Johnathan Schone

{patrickjohn.schone, johnathan.schone}@familysearch.org

4 March 2025

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS

FAMILYSEARCH

Background

Goal:
Build a visual language identifier capable of tagging the language(s) of every FamilySearch image using the image properties (and possible metadata). --- FIVE BILLION Images

Limitations:
Needs to be CHEAP. About 50K-100K images per $1.
Needs to be FAST. About 2K images/second to finish in 1m.
Needs to be GOOD. About 95+% accuracy/image.

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS
FAMILYSEARCH

Training Dataset Information



Amharic	Devana-gari	Kannada	Sinhala
Arabic	Empty	Katakana	Syriac
Armenian	Fraktur	Khmer	Tama-zight
Bengali	Georgian	Lao	Tamil
Burmese	Gothic	Latin	Thaana
Chinese-simplified	Greek	Malayalam	Thai
Chinese-traditional	Gujarati	Manchu	Tibetan
Cortesana	Gurmukhi	Norse	Tifinagh
Cuneiform	Hangul	Old-church-slavonic	Unknown
Cyrillic	Hebrew	Oriya	
Demotic	Hierogly-phics	Runes	
Deseret	Hiragana	Secretary-hand	

AFRIKAANS	GEORGIAN	LATVIAN	SANSKRIT
ALBANIAN	GERMAN	LETTER	SERBIAN
AMHARIC	GREEK	LINGALA	SHONA
ARABIC	GUARANI	LITHUANIAN	SINHALA
ARAMAIC	GUJARATI	LUXEMBOUR-GHISH	SLOVAK
ARMENIAN	GURMUKHI	MACEDONIAN	SLOVENIAN
AZERBAI-JANI	HAITIAN	MALAGASY	SPANISH
BASQUE	HAUSA	MALAY	SWAHILI
BELARU-SIAN	HAWAIIAN	MALAYALAM	SWEDISH
BENGALI	HEBREW	MANCHU	SYRIAC
BULGARIAN	HINDI	MAORI	TAGALOG
BURMESE	HUNGAR-IAN	MEKEO	TAHITIAN
CATALAN	ICELANDIC	MESOPOTA-MIAN	TAMIL
CEBUANO	IDO	MONGOLIAN	TATAR
CELTIC	IGBO	NAHUATL	THAANA
CHEROKEE	ILLEGIBLE	NAVAJO	THAI
CHINESE	ILOCANO	NORWEGIAN	TIBETAN
CROATIAN	INDONES-IAN	NUMERICS	TIFINAGH
CZECH	ITALIAN	ORIYA	TONGAN
DANISH	JAPANESE	PASHTO	TURKISH
DUTCH	JAVANESE	PERSIAN	UIGHUR
EGYPTIAN	KANNADA	POKAO	UKRAINIAN
EMPTY	KAZAKH	POLISH	UNKNOWN
ENGLISH	KHMER	PORTUGUESE	URDU
ESTONIAN	KOREAN	PUNCTUATION	VIETNAMESE
FINNISH	KUNI	ROMANIAN	WELSH
FRENCH	LAO	RUSSIAN	YIDDISH
GAELIC	LATIN	SAMOAN	YORUBA

Images are tagged by LANGUAGE_Script pairs

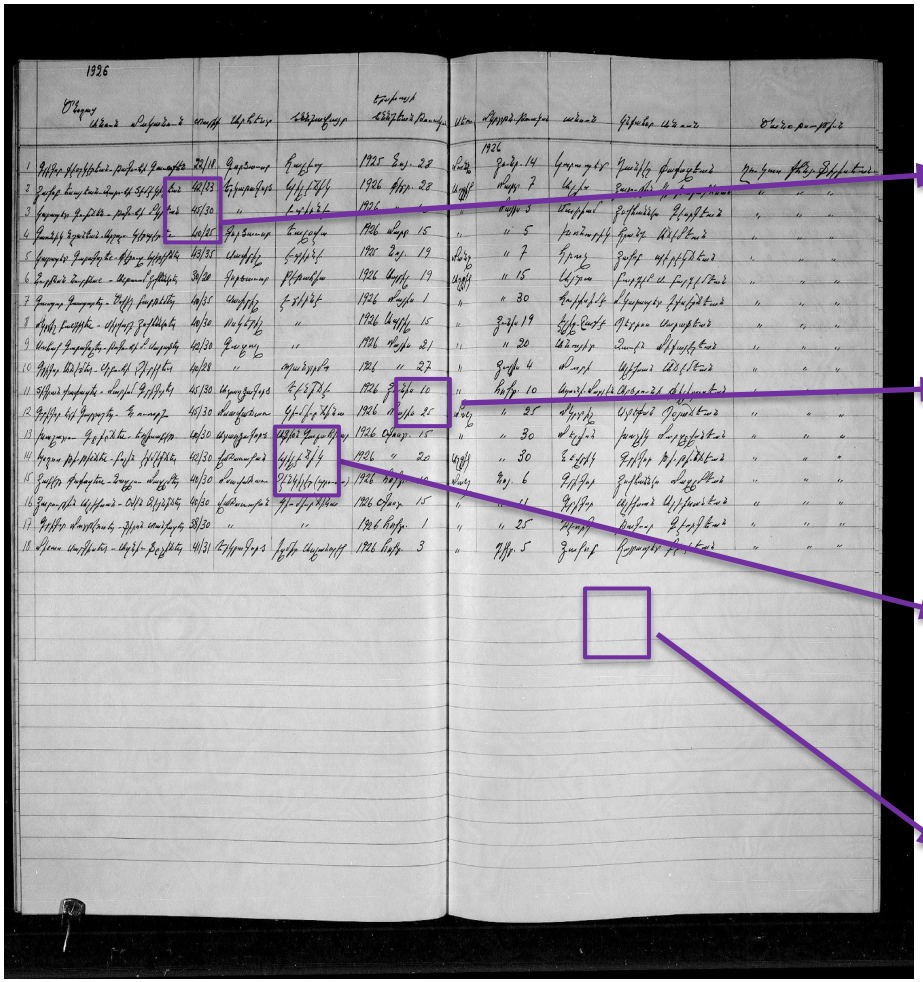
And secondary pairs are also tagged.

389 possible combos

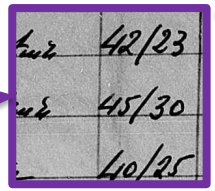
CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS
FAMILYSEARCH

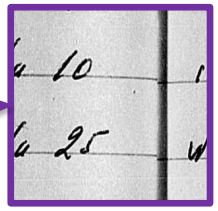
Initial Approaches

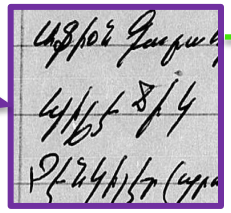
1. Rapidly Decode Image



2. Rand Select sub-cells









3. 93.3% Coarse categories (+sz/scr?)

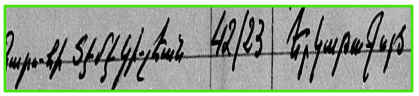
Text/ MedSz

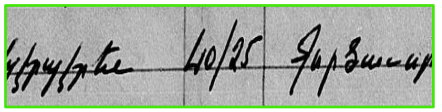
Text/ Num

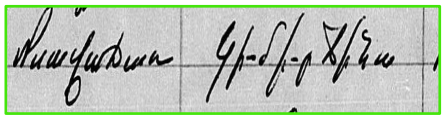
Text/ MedSz

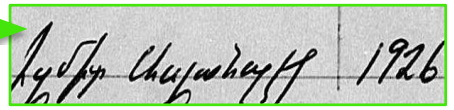
None

4. Refine & Lengthen







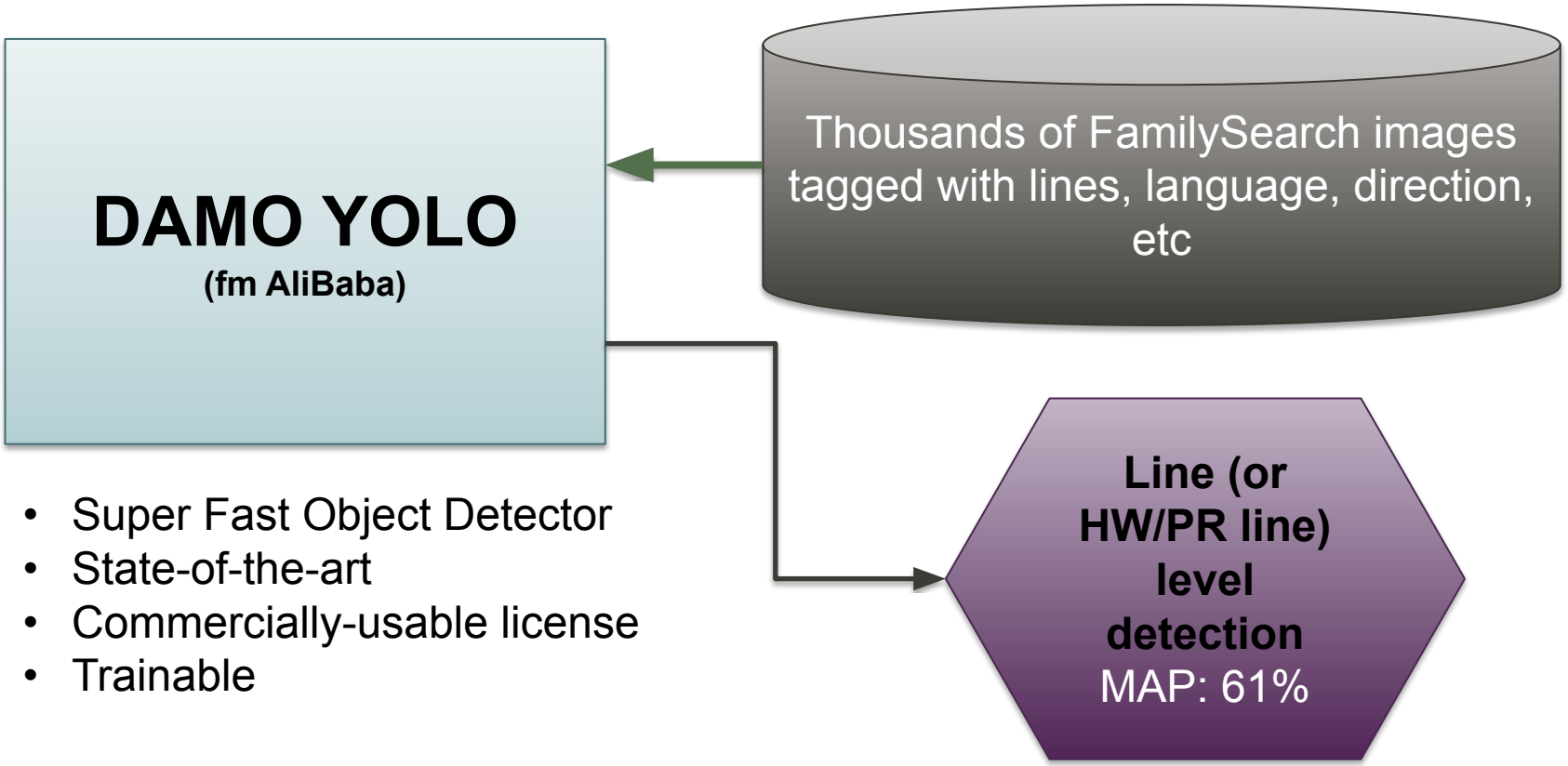


5. Catgz & Vote

Fast, but various complexities

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS | FAMILYSEARCH

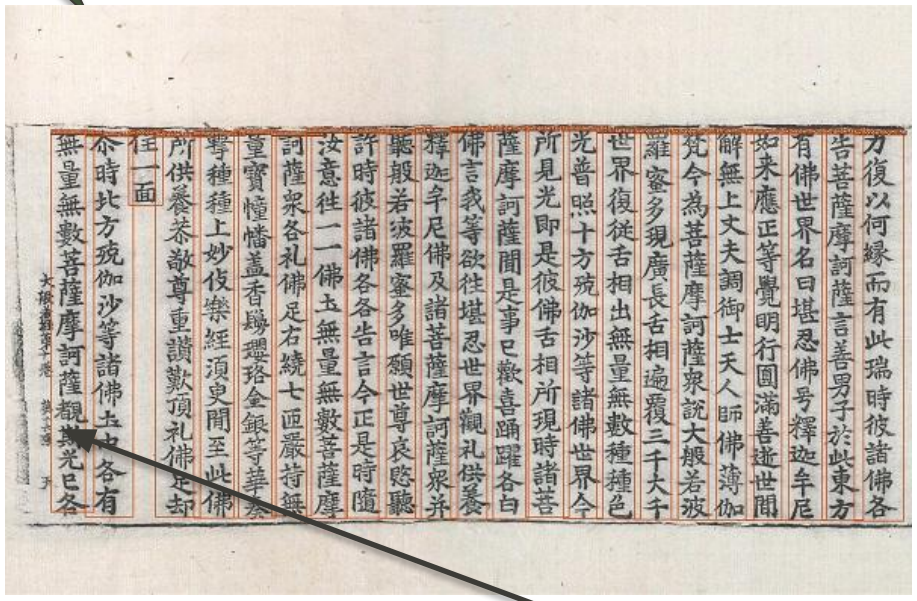
Alternative: Build Speedy Line Detector

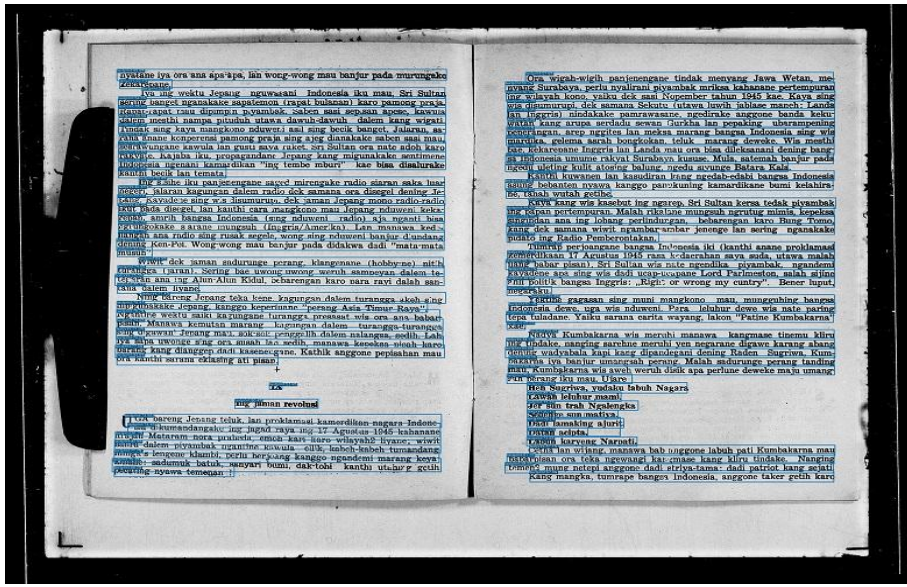


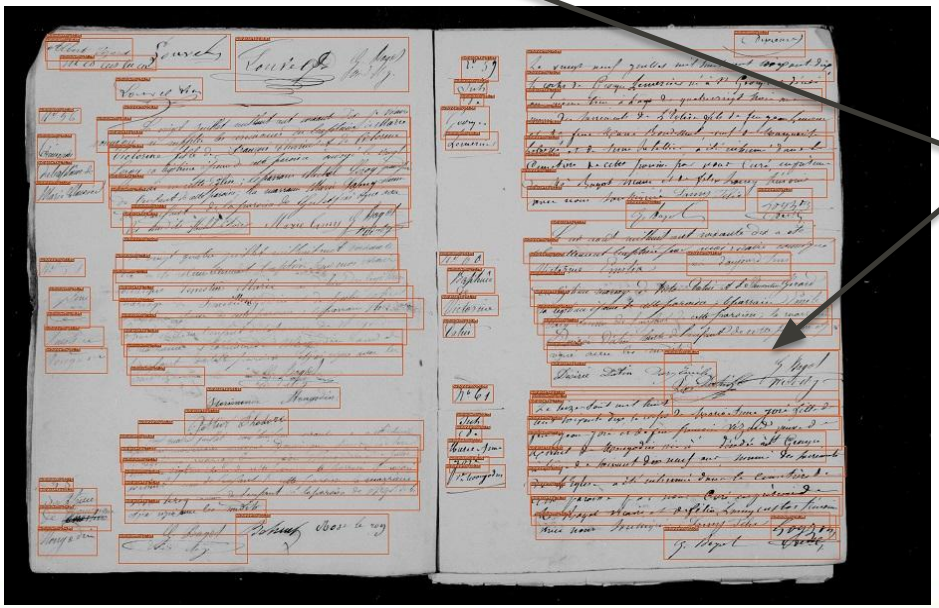
- Super Fast Object Detector
- State-of-the-art
- Commercially-usable license
- Trainable

Calling in "DYFS" for this paper

DYFS: Rapid Text Line Detection







DYFS:
Lines missed, but still excellent
5.5 Full-sized images/second, or
30 900x990 per second.

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS
FAMILYSEARCH

Snippets Challenges

PUNCTUATION



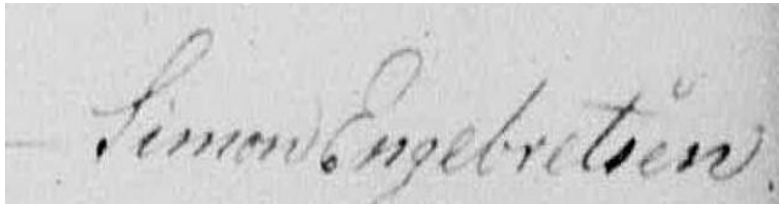
NUMERICS





Treat as “Specialized Languages”

NAMES



COMMON PHRASES



FRAGMENTS



Treat Normally

Snippets to Local Predictions

CNN 3x3 1->16, pool-by-2, RELU

Dropout 1/8

CNN 3x3 16->24, pool by 2, RELU

CNN 1x1 24->24, RELU

CNN 3x3 24->36, pool by 2, RELU

CNN 1x1 36->36, RELU

Dropout 1/8

CNN 3x3 36->54, pool by 2, RELU

CNN 3x3 54->81, RELU

Our Best Network So Far

Built a fast neural network
Trained with 1024-elem batches
Each batch filled using sqrt choice

	Language+ Script Snippet Accuracy	Script Snippet Accuracy
Uniform Baseline	**Random: 1/165 = 0.61%**	**Random: 1/56 = 21.8%**
Weighted Baseline	**Choose Best: 17.9%**	**Choose Best: 69.3%**
Ours	**63.4%**	**87.2%**

Snippet Confusions

LANGUAGE-Script Pair	Most-Frequent Confusion	Second Most-Frequent Confusion
AFRIKAANS_Latin	DUTCH_Latin	LATIN_Latin
ARABIC_Arabic	PASHTO_Arabic	TURKISH_Arabic
CATALAN_Latin	SPANISH_Latin	LATIN_Latin
CHINESE_Chinese-simplified	CHINESE_Chinese-traditional	MANCHU_Manchu
DANISH_Latin	DANISH_Fraktur	LATIN_Latin
ENGLISH_Latin	SPANISH_Latin	FRENCH_Latin
GEORGIAN_Georgian	ARMENIAN_Armenian	RUSSIAN_Cyrilic
JAPANESE_Hiragana	JAPANESE_Katakana	LETTER_Chinesetrad
SLOVENIAN_Latin	CROATIAN_Latin	SLOVAK_Latin
UKRAINIAN_ Cyrillic	BULGARIAN_ Cyrillic	RUSSIAN_ Cyrillic

Language-script confusions seem to match those (colored) that a human might expect to make. **Suggests system is learning language & not format.**

From Snippets to Whole Image

OPTION 1: Probabilistic $\Pr(X|D) = 1 - [\prod (1 - \Pr_i(X))]^{1/N} \{\forall i \in D\}$

	Scripts Only	Languages Only	Languages +Scripts
Best case Accuracy	95.0%	84.9%	**82.6%**

From Snippets to Whole Image

OPTION 1: Probabilistic $\Pr(X|D) = 1 - [\prod (1 - \Pr_i(X))]^{1/N} \{\forall i \in D\}$

	Scripts Only	Languages Only	Languages +Scripts
Best case Accuracy	95.0%	84.9%	**82.6%**

OPTION 2: Secondary FUSING network that consolidates snippet outputs

	Scripts Only	Languages Only	Languages +Scripts
Multi-case Accuracy	90.1%	83.3%	**79.7%**
Best case Accuracy	95.5%	88.1%	**86.4%**

16.0 images/second

Futures/Conclusions

Demonstrated data and technology that can quickly perform high-quality language/script detection.

We are JUST ABOUT THERE!

NEAR FUTURE REFINEMENTS

- Couple DYFS with snippet-level tagger for added speed
- Couple with whole-image/metadata decisions into fusion network
- Do million-image test
- Go for the gusto!!

Any Questions?