The Familysearch Iris Evaluation

Goal: evaluate the technology for automatic information extraction from US Census 1930 handwritten forms.

1930 Census forms
- In total: 2.5 millions images, 123 millions records
- For the evaluation: 14639 images, 740 000 records

Information extraction from 10 fields
- Large vocabulary: Identity (surname, given name, middle initial and/or titles), person’s, father’s and mother’s birth place, immigration year
- Small vocabulary: relation to head (wife, son), sex, age (3/12, 46), race (W, B), marital status.

Table Layout Analysis

Registration with printed anchors
- Black border detection and removal
- Skew correction
- Noisy borders detection and removal
- Keywords detection

Cells extraction
- Table detection
- Column detection
- Row detection
- Cell extraction at column/row intersection
- Around 500 cells extracted per page

Handwriting Recognition

Small lexicon/character recognizer: Convolution neural networks
- 2-Dimensional Long Short-Term Memory units (LSTM)
- Training from sequences (CTC)
- Lexical/Syntactical constraints encoded with Weighted Finite State Transducers (WFST)
- Recognizer combination using ROVER

Large lexicon recognizer: recurrent neural networks

Identify Field: Lexical Resources
- US Census 1920 dictionaries are used for realistic coverage
- One dictionary per state, with frequencies
- Unfrequent names are filtered out

Recognition Results
- Identity field:
  - 26% WER at full automation
  - 70% automation at human level
  - WER (10%)
- Small vocabulary fields:
  - 1% WER at 90% automation
  - An automatic system can replace a keyer in a double keying setting
  - Fields can be automatically located for keyers