Page Segmentation using Fully Convolutional Networks

Seth Stewart
www.linkedin.com/in/sethstewart

Bio

Seth Stewart works as an MS student with Dr. Bill Barrett at Brigham Young University, joining his lab in the Fall of 2015. He began his machine learning research as an intern at FamilySearch working on digital newsprint transcription. He loves to find and preserve family history gems for posterity.

Problem Statement

Many historical documents are forms such as those shown in Figure 1a and 1c, and include birth, death and marriage certificates, census records, etc. While layouts vary greatly, they typically consist of machine print, handwriting, dotted lines, solid lines, and stamps. In order to perform meaningful tasks such as handwriting recognition, OCR, and partitioning a form into zones, rows or columns to associate information, it is necessary to understand which pixels represent handwriting, which are machine-printed text, which are form lines and which are unrelated content.

Methods

We use a fully convolutional neural network (FCN) to peel apart a given document image into semantically distinct layers (i.e. machine print, handwriting, solid lines, dotted lines, stamps, etc.), at the pixel level, preparing it for subsequent intelligent recognition and content association. The process of classifying document components is referred to in literature as *page segmentation* [3]. We show that an FCN can accurately segment new document images when trained on a single representative pixel-labeled document image, even when layouts differ dramatically (see Figure 1).

Many previous methods in document page segmentation assumed large, contiguous rectangular or polygonal content regions without occlusions or overlapping content [1], [2]. These assumptions do not work well on form images. The tight, mixed layouts and frequent stroke overlap in these documents require a new paradigm of document content segmentation. Fortunately, FCNs naturally support pixel-level prediction. Also, in contrast to many existing segmentation approaches [4], we allow multiple labels to be predicted per pixel location, which allows for recovery of overlapped content; see Figure 2. Precise pixel labels enable layerwise content reconstruction at unprecedented levels of detail; see Figure 3. Given our pixel-level content labeling, we show potential for improvements in OCR and handwritten text recognition, and for associating semantically related document components (e.g. machine print and related handwriting) for contextually constrained recognition.

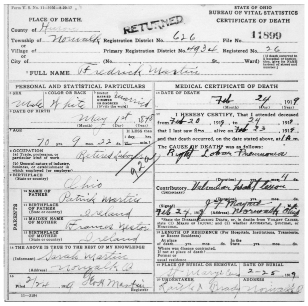
References

[1] D. Silva, L. F., Conci, A., and Sanchez, A. Automatic discrimination between printed and handwritten text in documents. In *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing* (Oct 2009), pp. 261–267.

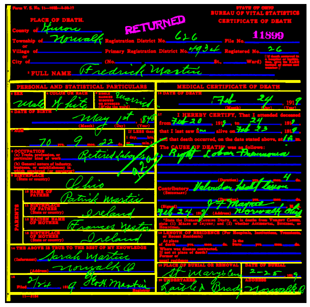
[2] Mehri, M., Héroux, P., Gomez-Krämer, P., and Mullot, R. Texture feature benchmarking and evaluation for historical document image analysis. *International Journal on Document Analysis and Recognition (IJDAR) 20*, 1 (2017), 1–35.

[3] Shafait, F., Keysers, D., and Breuel, T. Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 6 (June 2008), 941–954.

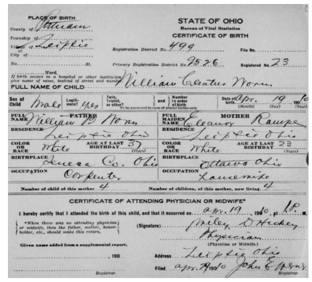
[4] Thoma, M. A survey of semantic segmentation. *CoRR abs/1602.06541* (2016).



(a) Ohio death record



(b) Ground truth labels for (a)

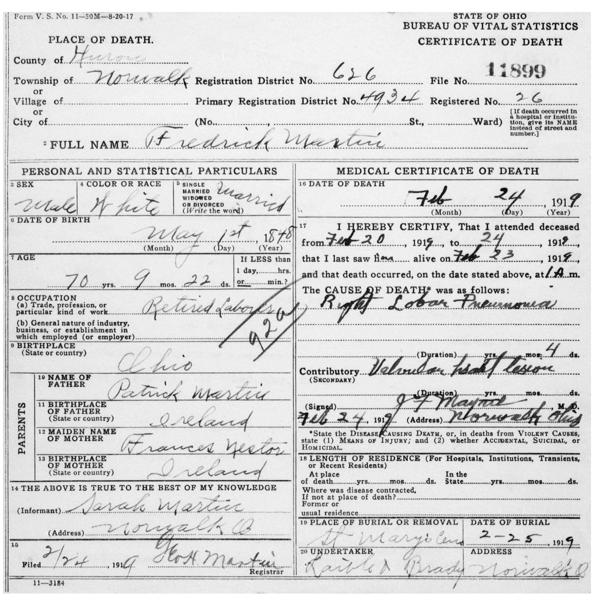


(c) Birth certificate

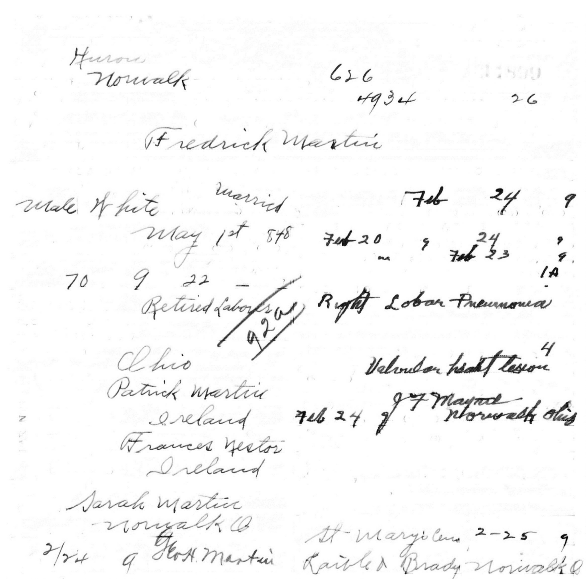
(d) Classification results on birth certificate in (c)

Figure 1: A CNN trained on a single hand-annotated form image (a),(b) generalizes well to recognizing pixel-level content on a never-seen-before type of document image (c), (d)

Figure 2: Overlap of content types is extremely common. Traditional classification approaches would segment this content destructively by assigning each overlapped region to only one class, making OCR and handwriting recognition very difficult.



(a) Form Image



(b) Handwriting masked using predicted handwriting pixels by FCN

Figure 3: An FCN enables automatic pixel-level isolation of content types, which will facilitate tasks like handwriting recognition.