

The Human Element in Document Analysis Systems

Daniel Lopresti [0000-0003-2129-4223]

Lehigh University, Bethlehem PA 18015, USA

lopresti@cse.lehigh.edu



Summary

- Humans are a vital component in document analysis systems. A more intentional inclusion of the human element would yield better outcomes for those impacted by what we build, and also lead to interesting research questions.

Motivation

- The “S” in “DAS” refers to systems.
- This definition should include humans at all levels, including those who design, implement, and use document analysis systems.
- It should also include those who are indirectly impacted by the systems we build.
- To date, little attention has been paid to this, beyond attempts to facilitate ground-truthing and a few applications to help the disabled.
- Elsewhere, the broader field of AI is confronting serious issues of bias and fairness (e.g., facial recognition, large language models).
- Is the document analysis community immune to such concerns?

Three Illustrative Cases

Where might we start looking to see whether our systems and our policies negatively impact segments of the population?

Reliably Reading Hand-Marked Paper Ballots

- Optical Mark Reading has a long history and appears to be a solved problem.
- So simple it does not attract much attention.
- When examining a real population of voters, however, representing all demographics from across society, the problem becomes much more challenging.
- Voter intent is the overriding consideration.
- A system that fails to count certain voters’ ballots accurately is a serious problem. A ballot-reading system that looks like it is doing a good job on average may still be unfairly disenfranchising certain groups of voters.
- Images below are from ballots challenged during the 2008 U.S. presidential election.
- Do we know whose votes are not being counted correctly? Should we worry?

Robust Signature Verification for Elections

- Voters are sometimes required to sign their name as proof of identity.
- This signature is compared to one collected when the voter first registered, which could be decades earlier.
- These comparisons are normally made by election officials, but if there is a push to automate this process, who will be hurt?
- Voters with a name-change due to marriage, a hand injury, a stroke, or forgetfulness about how they signed so many years ago?

Document Analysis for Under-Resourced Languages

- Thousands of languages in the world today, but most research reported at events like DAS reflects only a small percentage of these.
- Often, submissions applying known techniques to a new language will be rejected for “lack of contribution.”
- Existing methods are often easily adapted, but why should first language “win the race”?
- What do we lose by erecting such barriers?

