Reconstructing Shredded Documents

Razvan Ranca

ranca.razvan@gmail.com

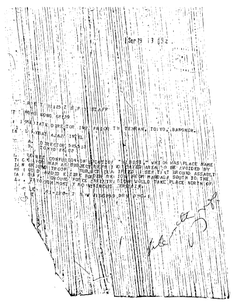
The University of Edinburgh

Importance

Despite the development of techniques permitting the purely electronic storage and transmittal of sensitive documents, many such documents are still printed and eventually shredded. Traditionally, the cost of reconstructing these documents was considered prohibitive, however with the development of methods that largely automate the process this situation is changing.

It is currently unclear what level of security the paper shredder still offers.

A few of the 16,000 bags holding the shredded archives of the East German secret police



A shredded document reconstructed during the Iran hostage crisis

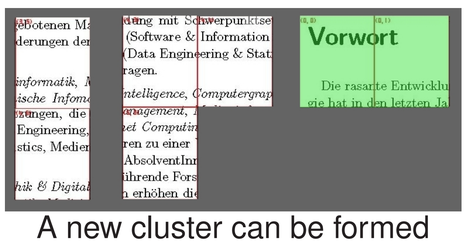
Probabilistic Score

procedure ProbScore(Et)
  Initialize pr
  for all $Ex \in Edges$ **do**
    for all $pixel \in Ex$ **do**
      $pr_{Ex} \leftarrow pr_{Ex} * \Pr(pixel | Neighbors_{Et})$
  Normalize pr
  Assert: $\sum_{Ex \in Edges} pr_{Ex} = 1$
  return $\max pr$, $\arg\max pr$

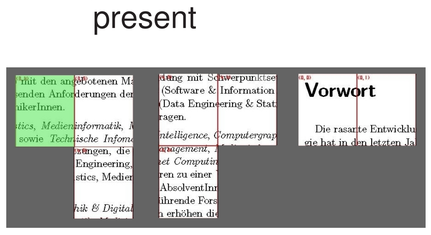
ProbScore returns the most likely match for the input edge Et, and the probability of that match being correct.

Kruskal-inspired Search

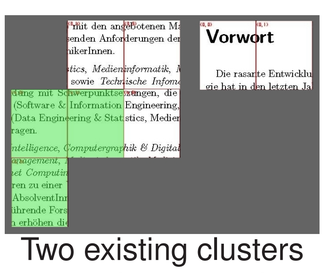
Initially, two clusters present



A new cluster can be formed



An existing cluster can be enlarged



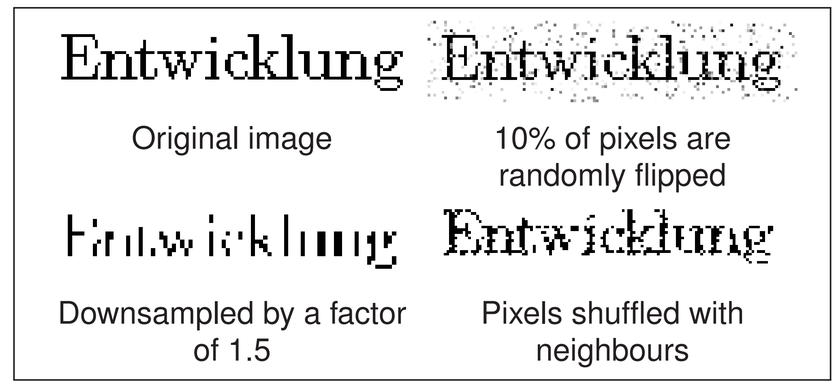
Two existing clusters can be merged

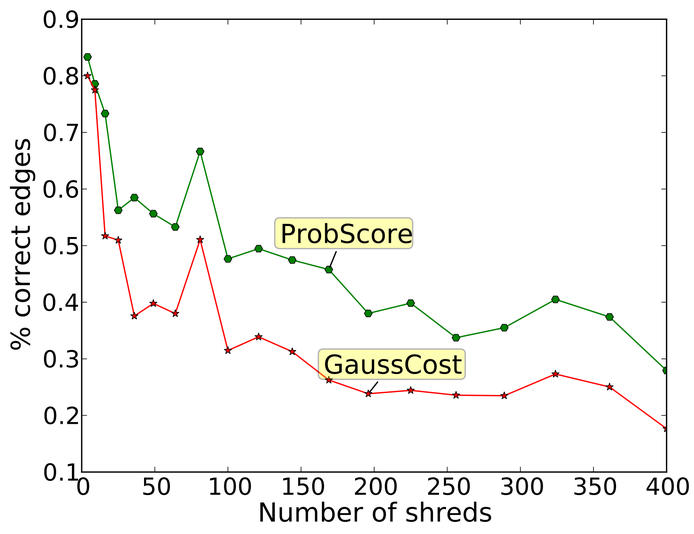
The "Kruskal" heuristic is based on the "ReconstructShreds" method presented in [1]. Analogous to the minimum spanning tree "Kruskal's Algorithm", this method tries to greedily unite the two best matching edges at every step.
The main difference from the basic algorithm is that the shreds need to have a specific position in 2D space and therefore uniting two pieces can cause an overlap. Any such overlap would lead to an illegal solution and so must be prevented.

Probabilistic Score - Evaluation

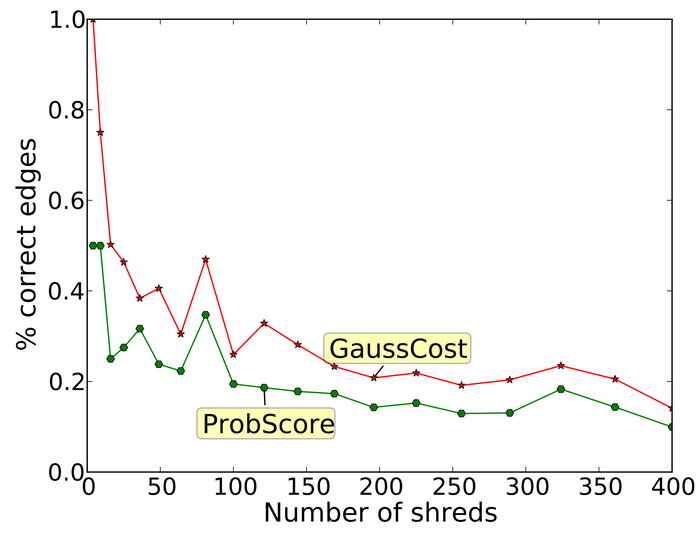
The probabilistic score function is evaluated against the best previously published cost function [1, 2]. The comparison looks at both noisy and noise-less documents.

ProbScore outperforms the Gaussian cost on three out of the four instances.

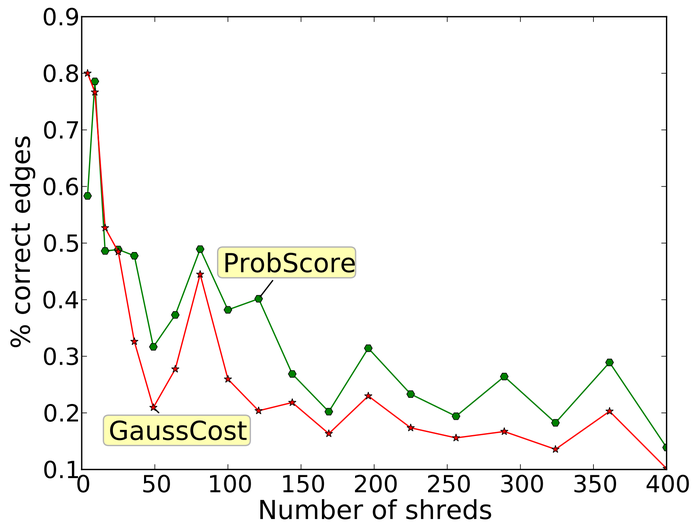




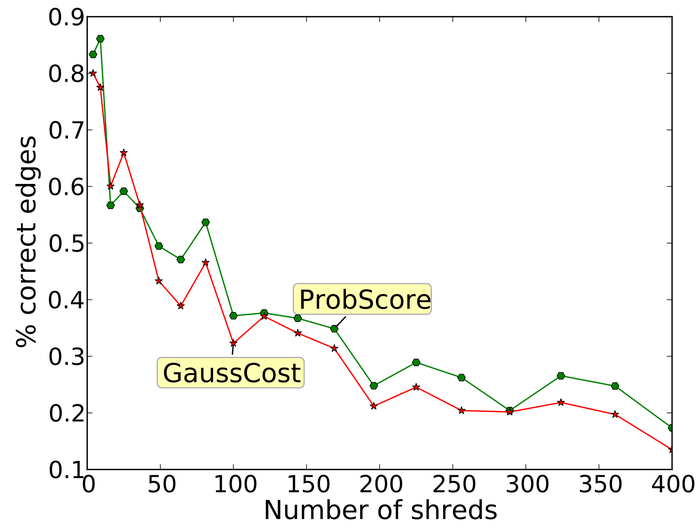
Original image



10% of pixels are randomly flipped

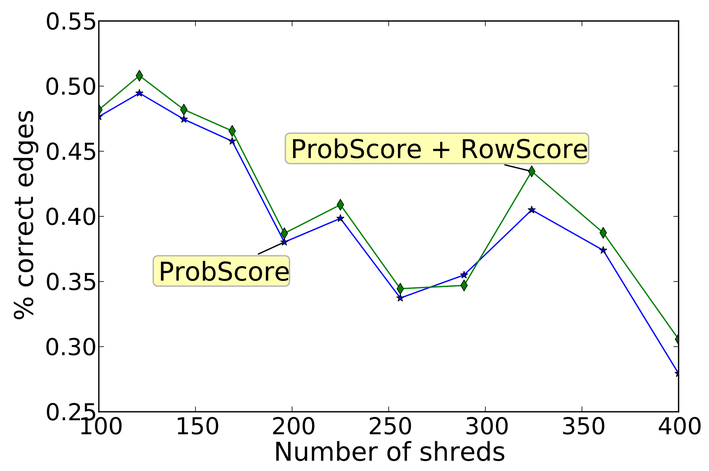


Downsampled by a factor of 1.5



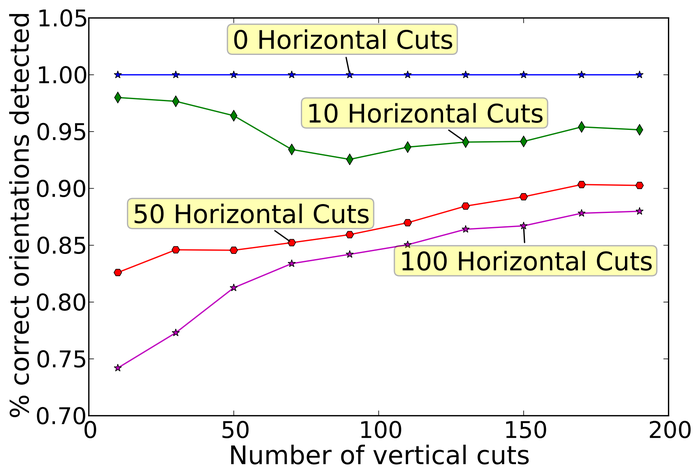
Pixels randomly shuffled with neighbours

Row detection

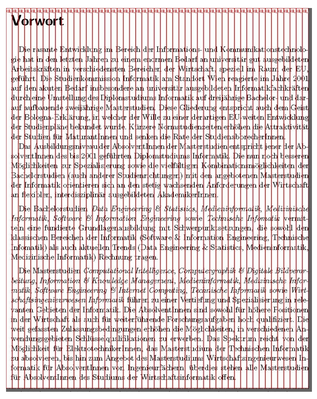


Simple Gaussian model on distance between matching rows improves accuracy slightly

Orientation detection can be performed by counting the pixels in the upper and lower regions and predicting that more black pixels will be on top



Results



Strip shredder

Cross-cut shredder

Both images are of documents cut into 49 shreds. While the strip shredded variant is perfectly reconstructed, only certain sections of the cross-cut document can be recovered with any certainty.

Even if the total number of shreds is the same, the cross-cut case is *much* harder to solve.

References

[1] Sleit, A., Massad, Y. and Musaddaq M. An alternative clustering approach for reconstructing cross cut shredded text documents In *Telecommunication Systems '11*

[2] Prandtstetter, M. and Raidl, G. Combining Forces to Reconstruct Strip Shredded Text Documents In *Hybrid metaheuristics '08*