14th INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION
1-4 September
ICFHR 2014
Crete Island-Greece

Rejecting both segmentation and classification errors in handwritten form processing

Claudio De Stefano, Francesco Fontanella, Alessandra Scotto di Freca

Dipartimento di Ingegneria Elettrica e dell'Informazione (DIEI)
University of Cassino and Southern Lazio, Cassino (FR), Italy
{destefano, fontanella, a.scotto}@unicas.it

Angelo Marcelli, Antonio Parziale

Dipartimento di Ingegneria dell'Informazione, Ingegneria Elettrica
e Matematica Applicata (DIEM)
University of Salerno, Fisciano (SA), Italy
{amarcelli, anparziale}@unisa.it

Outline

Motivation

Rationale of the approach

Implementation

Experimental results

Conclusions and future works

Motivation (1/3)

handwritten form processing is a routine task performed daily in many organizations for capturing the data to be processed

- the core technology generally used is an OCR engine
- most of the systems commercially available for daily use within organizations consider a verification stage in which a human operator is needed to confirm or modify the output of the OCR engine
- to reduce the cost of the verification, a **reject option** for detecting the cases when the output of the OCR engine is not reliable is often introduced, so as to limit the call for verification to just these cases

Motivation (2/3)

for implementing a reject option an estimate of the reliability, or confidence, of the classifier decision is generally required

- on the basis of such measure, the ideal solution is the one that rejects only the specimen leading to recognition errors but none of the ones that can be correctly recognized
- if the conditional probability densities of each class are exactly known, the optimal error/rejection tradeoff can be found, but …
  - ✓ in real applications these densities are generally unknown
  - ✓ samples processed by the OCR engine do not necessarily correspond to isolated characters

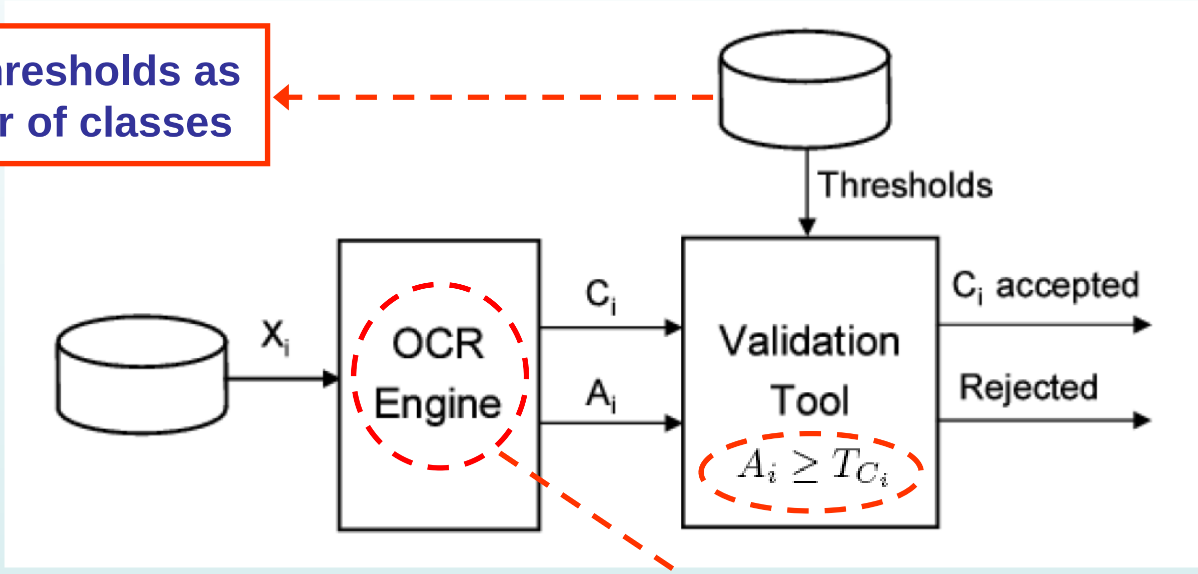
Motivation (3/3)

we have investigated to which extent the reliability provided by the OCR engine allow us to implement a reject option that deals simultaneously with **both** classification and segmentation errors

- the system should be able to reject both images containing single characters not correctly recognized and images containing more than one character or even cursive handwriting
  - ✓ reduction of the operational cost

- the system designer can use an OCR engine trained with boxed **isolated characters** for processing also boxed **data fields**
  - ✓ reduction of the development cost
  - ✓ shorter deployment time

Rationale of the approach (1/2)



As many thresholds as the number of classes

Random Forest Classifier

$$C_i \in \{1, 2, \ldots, C\}$$

$$A_i = P(C_i|\mathbf{x_i})$$

Rationale of the approach (2/2)

Given a sample x_i to be classified, the conditional probabilities for each class are computed by averaging the conditional probabilities given by the trees making up the ensemble:

$$P(c|\mathrm{x_i}) = \frac{1}{L}\sum_{j=1}^{L} P(c|\mathrm{x_i}, T_j)$$

$$P(c|\mathrm{x_i}, T_j) = \frac{n_c}{n}$$

- n_c is the number of training samples belonging to the class c that have been classified by a leaf node of T_j,
- n is the total number of training samples classified by that leaf node of T_j.

Both values are obtained at the end of the learning procedure of RF.

Implementation (1/4)

- we considered 52 class-related reject thresholds in our reject rule
- each threshold value is encoded by means of a string of 10 bits, so as to obtain a reasonable precision for representing a real value in the range [0.0, 1.0]
- thus, a candidate solution is composed of a bit string of 520 bits
- the optimization algorithm should find the optimal solution in a search space whose cardinality is 2^{520}

very complex optimization problem!

Implementation (2/4)

We have reformulated the problem of finding the whole set of C thresholds in C sub-problems in which a single threshold must be found:

- given a dataset D the optimal value of each class-related threshold T_{C_i} is searched by considering only the responses of the OCR module providing as output the class C_i
- for each response, both the associated probability and the knowledge about the “true class” of the corresponding sample are used to optimize the objective function

Implementation (3/4)

$$f_o(C_i, T_{C_i}, D) = \frac{N_{cr} + N_{ea} + N_{sa}}{N_{C_i}}$$

- N_{Ci}: the number of samples of D assigned to the class C_i
- N_{cr}: the number of samples of D correctly assigned to the class C_i
- N_{ea}: the number of samples of D erroneously assigned to the class C_i with $A_i \geq T_{Ci}$
- N_{sa}: the number of samples of D corresponding to segmentation error assigned to the class C_i with $A_i \geq T_{Ci}$

Implementation (4/4)

- it guarantees to find the global optimum: since each threshold T_{C_i} has effect only on the responses of the OCR module providing as output the class C_i, its optimal value can be searched without considering neither the other responses, nor the other threshold values
- it allows us to exhaustively search the optimal solution for each threshold: in this case, in fact, each candidate solution is composed of a bit string of 10 bits and the cardinality of the search space is only 2^{10}: for each class C_i, find the value T_{C_i} corresponding to

$$\min_{0 \le T_{C_i} \le 1} f_o(C_i, T_{C_i}, \mathcal{D})$$

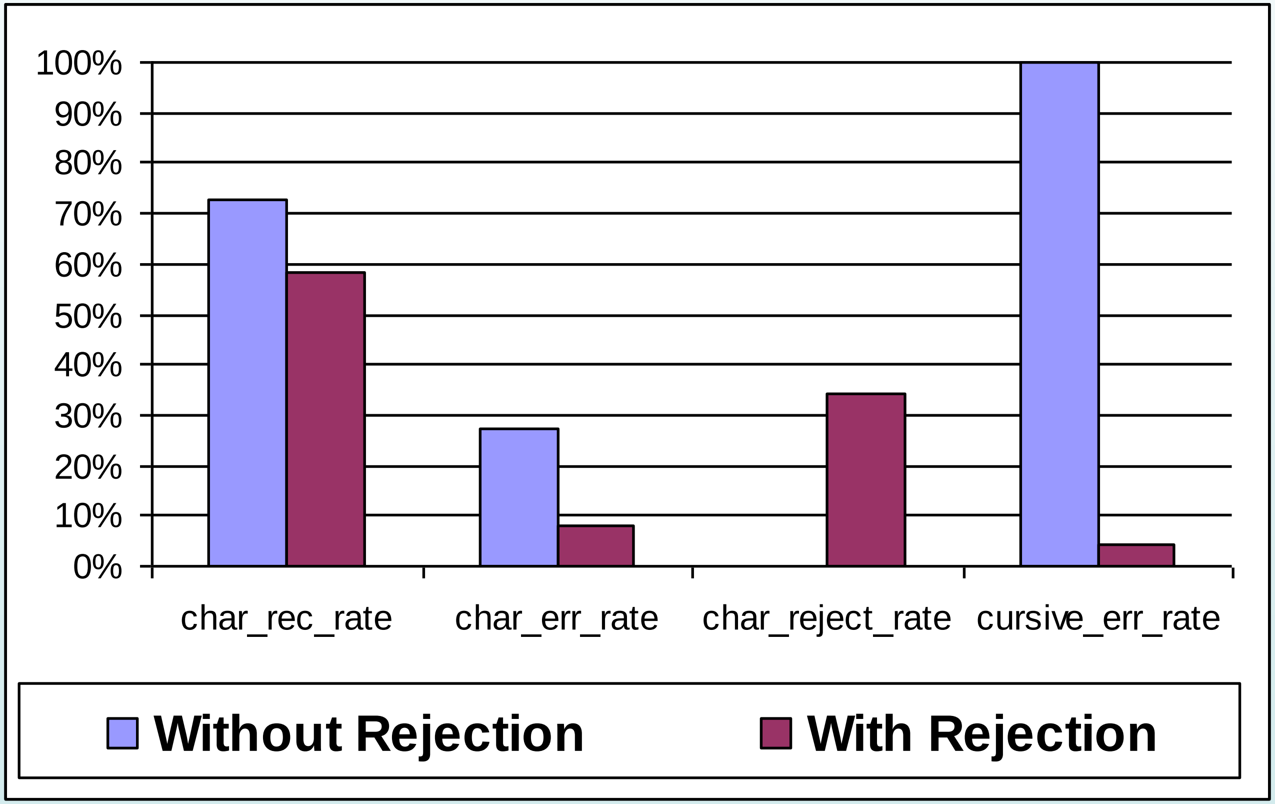
Experimental Results (1/5)

- we used a data set of 5,000 forms currently in use by an organization
- after the digitalization, a segmentation algorithm was applied to each image to extract sub-images containing single characters.
- we have manually labeled the samples adding a further class for representing segmentation errors, i.e. images that correspond to pieces of ink containing more than one characters or cursive handwriting
- thus, we have 52 classes to be discriminated (uppercase and lowercase letters) and a further class (denoted as *cursive*) which corresponds to segmentation errors

Experimental Results (2/5)

- the samples were arranged into three statistically independent datasets:
    - ✓ the dataset TR1 contains 5,369 samples of isolated characters and was used to train the RF classifier
    - ✓ the dataset TR2 contains 10,355 samples of both isolated characters and cursive (5,210 characters and 5,145 segmentation errors) and was used for finding the threshold values
    - ✓ the dataset TS contains 10,358 samples of both isolated characters and cursive (5,212 characters and 5,146 segmentation errors) and was used for testing the system
- the data are strongly unbalanced, with classes having hundreds of samples and classes having just few tens

Experimental Results (3/5)



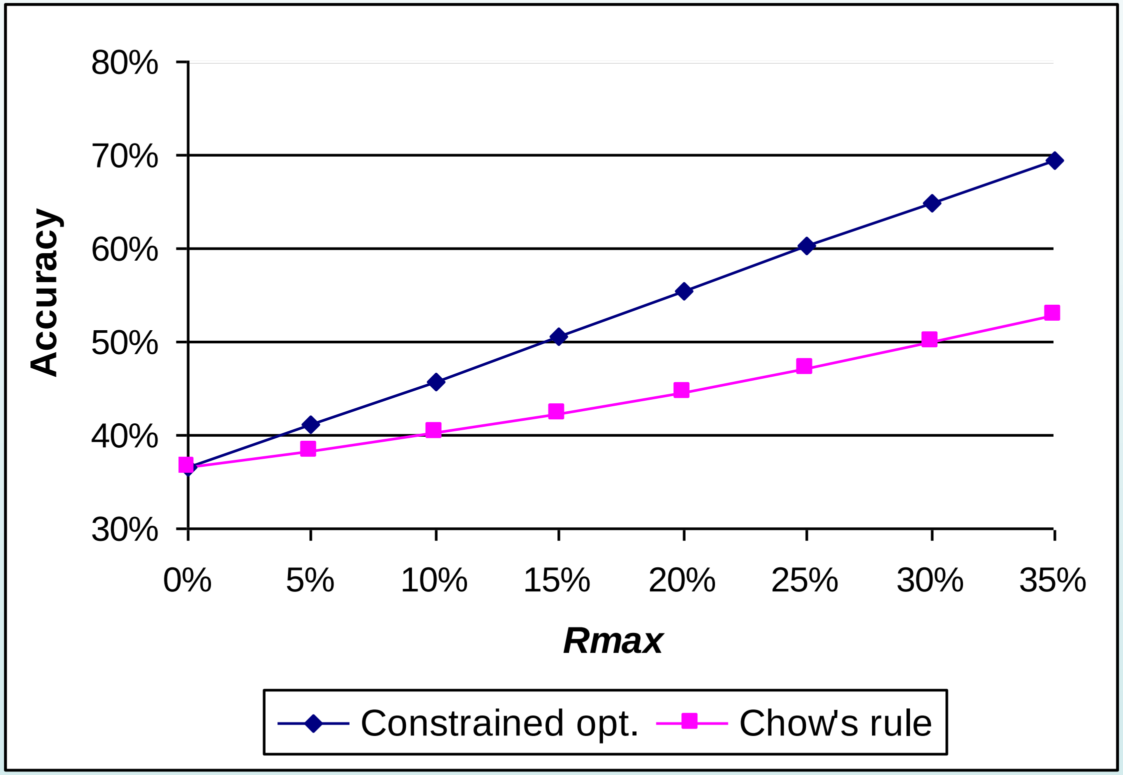
Experimental Results (4/5)

for better evaluating the effectiveness of our approach, we have also applied the Chow's reject rule, assuming that the reliability measures provided by RF are a good estimate of the a posteriori probabilities of each class

- we used the same dataset (TR2) to estimate both the values of Chow's reject threshold, and those of our thresholds
- the comparison with the Chow's reject rule implies the introduction of a constraint in our optimization algorithm: thus the optimization problem has been reformulated as follows

$$\begin{cases} \min f_o(C_i, T_{C_i}, \mathcal{D}) \\ R(C_i, T_{C_i}, \mathcal{D}) < R_{max} \end{cases}$$

Experimental Results (5/5)

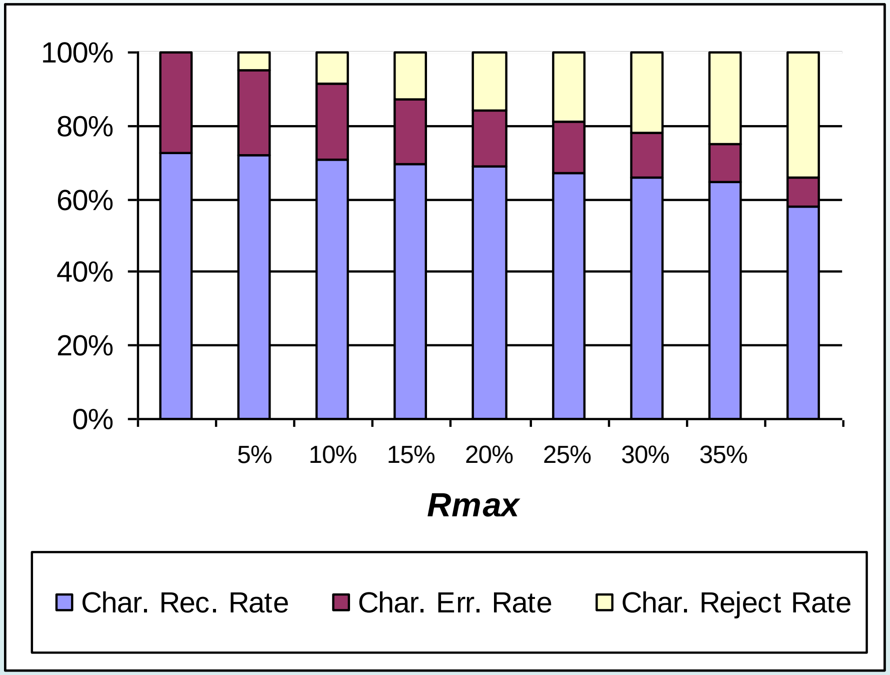


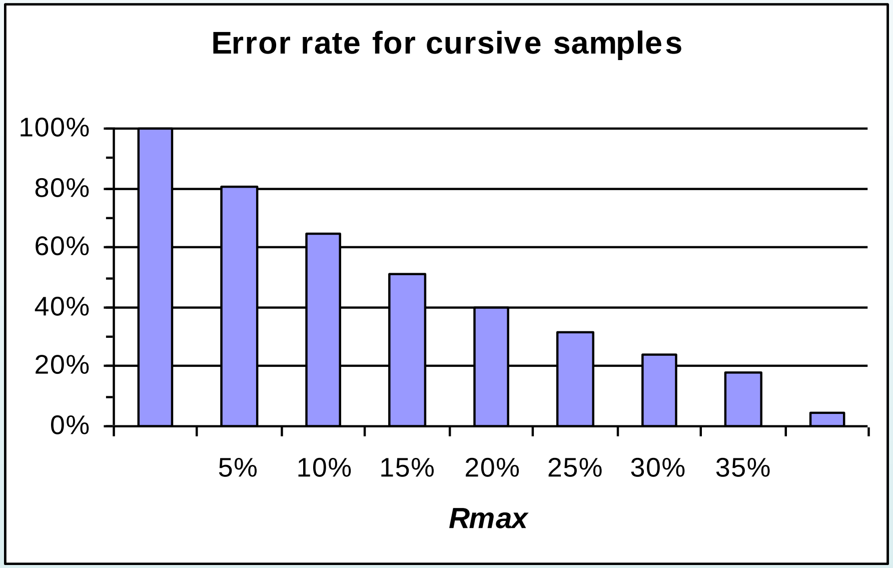
Conclusions and future works

- we have investigated to which extent the reliability provided by an OCR engine, designed to deal with boxed isolated characters, can be used to implement a reject option able to detect both segmentation and classification errors.
- these results are particularly meaningful considering that in our experiments we have included a number of samples representing segmentation errors almost equal to that of samples corresponding to isolated characters
- in future works we will attempt to characterize the behavior of our system as the number of segmentation error varies, as well as toreduce the computational cost by using 8 bits to encode the threshold values

Thank you

Experimental Results




Error rate for cursive samples