Improving Features Extraction for Supervised Invoice Classification

Alberto Bartoli, Giorgio Davanzo, *Eric Medvet*, Enrico Sorio

Deei - University of Trieste
Via Valerio 10, Trieste, Italy

AIA 2010

Talk's outline

Document understanding: what?

Document understanding $\equiv$ automatic extraction of information

Document understanding: what?

Document understanding $\equiv$ automatic extraction of information

- scientific paper $\rightarrow$ title, authors, DOI, ...
- form $\rightarrow$ inputed infos
- invoices $\rightarrow$ date, number, total amount, ...

Document understanding: why?

Automatic extraction of information

Document understanding: why?

Automatic extraction of information

- huge volume ← 2 billion paper forms by Japanese public administration

Document understanding: why?

Automatic extraction of information

- huge volume ← 2 billion paper forms by Japanese public administration
- high cost ← processing cost 13$ per unit for invoices

Document understanding: why?

Automatic extraction of information

- huge volume ← 2 billion paper forms by Japanese public administration
- high cost ← processing cost 13$ per unit for invoices

⇒ *Manual* processing is unpractical

Document understanding: typical work-flow

1. acquire the document
2. classify the document ("which model should we use?")
3. extract information (apply the model)

Document understanding: typical work-flow

1. acquire the document
2. classify the document ("which model should we use?")
3. extract information (apply the model)

Document classification: problem statement

We consider *invoice* classification...

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

- only visual features available (pixels)

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

- only visual features available (pixels)
- no structural features, no text features

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

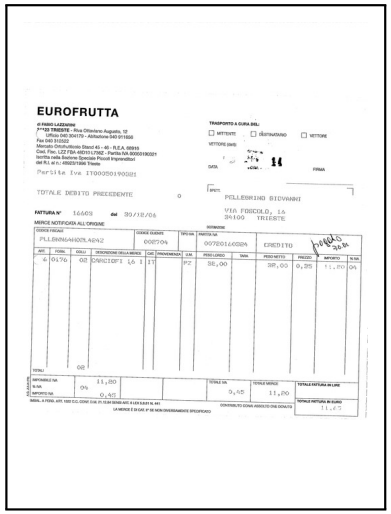
...before OCR

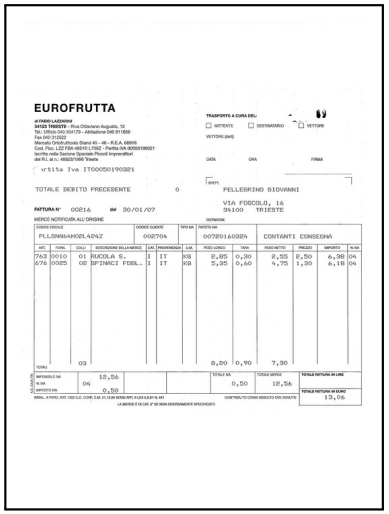
- only visual features available (pixels)
- no structural features, no text features

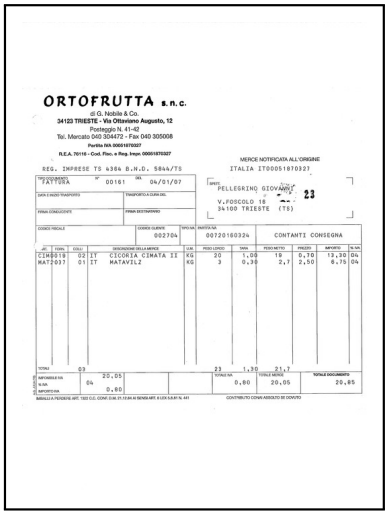
Goal

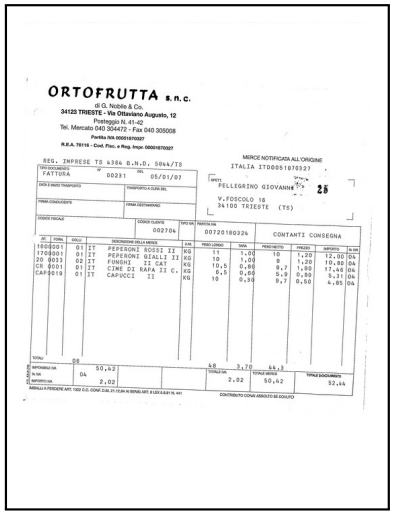
Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets

Classes with strong visual similarities



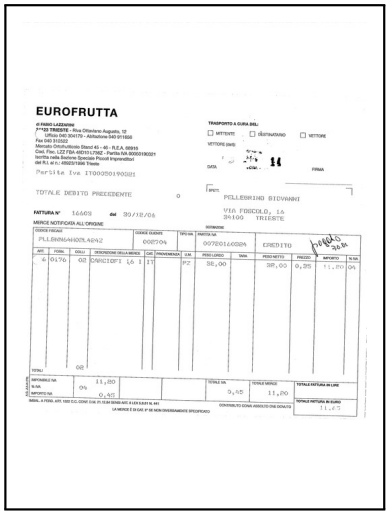


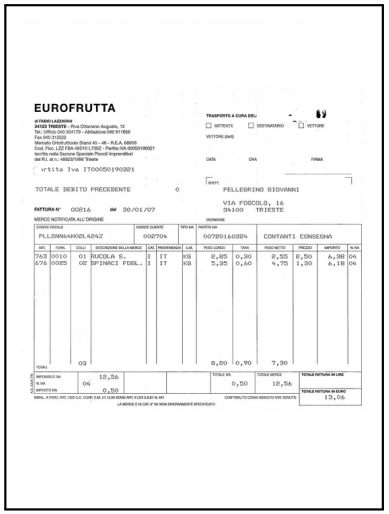




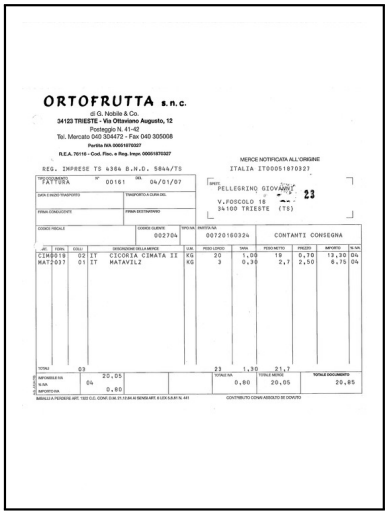
Classes with strong visual similarities

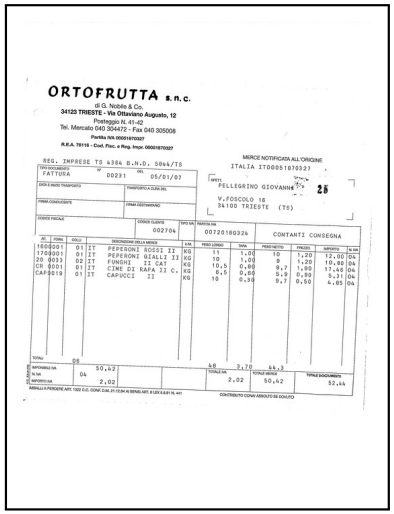
Class A





Class B





Additional challenge

Real-world scanned documents:

Additional challenge

Real-world scanned documents:

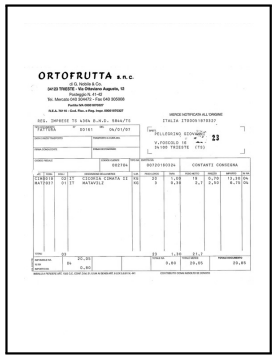
- positioning errors
- non standard paper size
- cut documents

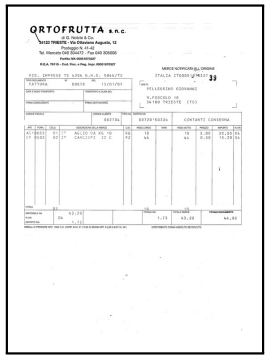
Additional challenge

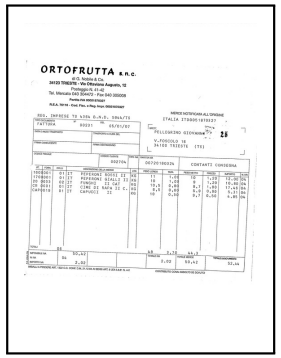
Real-world scanned documents:

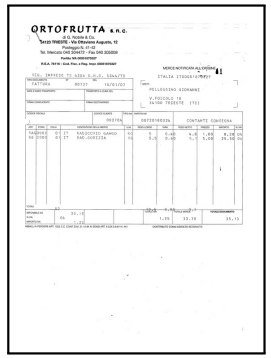
- positioning errors
- non standard paper size
- cut documents

Documents of the same class









System overview

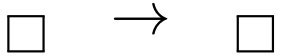
Document classification

1. preprocessing
2. feature extraction
3. actual classification

System overview

Document classification

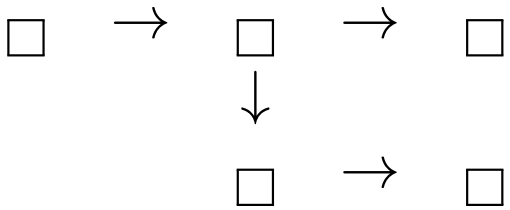
1. preprocessing image $\rightarrow$ image
2. feature extraction
3. actual classification



System overview

Document classification

1. preprocessing image $\rightarrow$ image
2. feature extraction image $\rightarrow$ image
3. actual classification



System overview

Document classification

1. preprocessing image $\rightarrow$ image
2. feature extraction image $\rightarrow$ image $\rightarrow$ numeric vector
3. actual classification

$$\square \rightarrow \square \rightarrow \square \rightarrow v^b$$
$$\square \rightarrow \square \rightarrow v^e$$
$$\rightarrow v = v^b \oplus v^e$$

System overview

Document classification

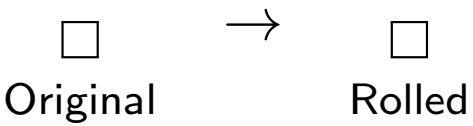
1. preprocessing image $\rightarrow$ image
2. feature extraction image $\rightarrow$ image $\rightarrow$ numeric vector
3. actual classification numeric vector $\rightarrow$ class

$$\square \rightarrow \square \rightarrow \square \rightarrow v^b$$
$$\downarrow$$
$$\square \rightarrow \square \rightarrow v^e$$
$$\rightarrow v = v^b \oplus v^e$$

Preprocessing

Why? See “additional challenge”
How:

1. scale image (300dpi $\rightarrow$ 50dpi)
2. apply edge detector
3. find uppermost relevant pixel
4. “roll” image



Feature extraction: black density

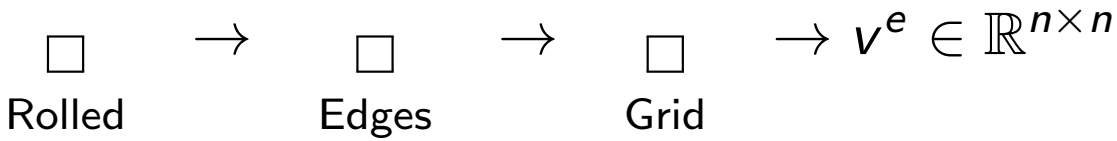
From rolled image:

1. $n \times n$ grid (best n is 16):
2. for each cell, count the number of black pixel

Feature extraction: edge density

From rolled image:

1. build edges image
2. $n \times n$ grid (same n as before):
3. for each cell, count the number of black pixel



Test bed

We want to assess effectiveness of:

- our proposal (grayscale + edge density)
- baseline (SVM, only grayscale)

In all cases, PCA

Test bed

We want to assess effectiveness of:

- our proposal (grayscale + edge density)
  - SVM (*page-svm*)
  - header, SVM (*header-svm*)
  - header&footer, SVM (*header&footer-svm*)
  - distance based (*page-distance*)
- baseline (SVM, only grayscale)

In all cases, PCA

Test bed

We want to assess effectiveness of:

- our proposal (grayscale + edge density)
  - SVM (*page-svm*)
  - header, SVM (*header-svm*) ← faster
  - header&footer, SVM (*header&footer-svm*) ← faster
  - distance based (*page-distance*)
- baseline (SVM, only grayscale)

In all cases, PCA

Dataset

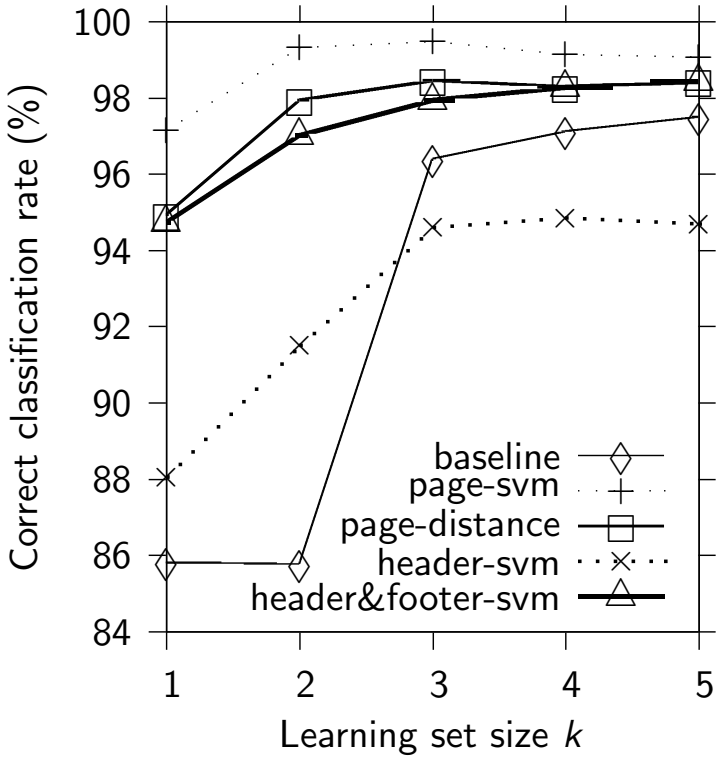
Real-world paper invoices:

- 68 issuers (issuer $\equiv$ class)
- 562 invoices

For each class, k invoices for learning, remaining for testing

Results

k = learning set size



k	Invoices	Errors	Error %
1	494	14	97.17
2	436	3	99.31
3	389	2	99.49
4	350	3	99.14
5	321	3	99.07

Table: Results for *page-svm*

Conclusions

Goal

Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets

Conclusions

Goal

Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets

Key findings:

- adding edges gives better classification

Conclusions

Goal

Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets

Key findings:

- adding edges gives better classification
- in particular with small training sets ($+10\%$ for $k = 1, 2$)

Conclusions

Goal

Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets

Key findings:

- adding edges gives better classification
- in particular with small training sets ($+10\%$ for $k = 1, 2$)
- faster *header&footer-svm* version performs well

Thanks

Thanks for the attention!

Document classification: problem statement

We consider *invoice* classification...

- large number of classes

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

- only visual features available (pixels)

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

- only visual features available (pixels)
- no structural features, no text features

Document classification: problem statement

We consider *invoice* classification...

- large number of classes, many with few documents
- classes with strong visual similarities

...before OCR

- only visual features available (pixels)
- no structural features, no text features

Goal

Finding **"better" features** in the proposed scenario to grant precise classification also with small training sets