# A study on the Classification of Layout Components for Newspapers

Stefano Ferilli[1] Floriana Esposito[1]
Domenico Redavid[2]

[1] Dipartimento di Informatica – Università di Bari
*name*.*surname*@uniba.it

[2] Artificial Brain S.r.l.
redavid@abrain.it

12th Italian Research Conference
on Digital Libraries - IRCDL 2016
Firenze, Italy, 4-5 February, 2016

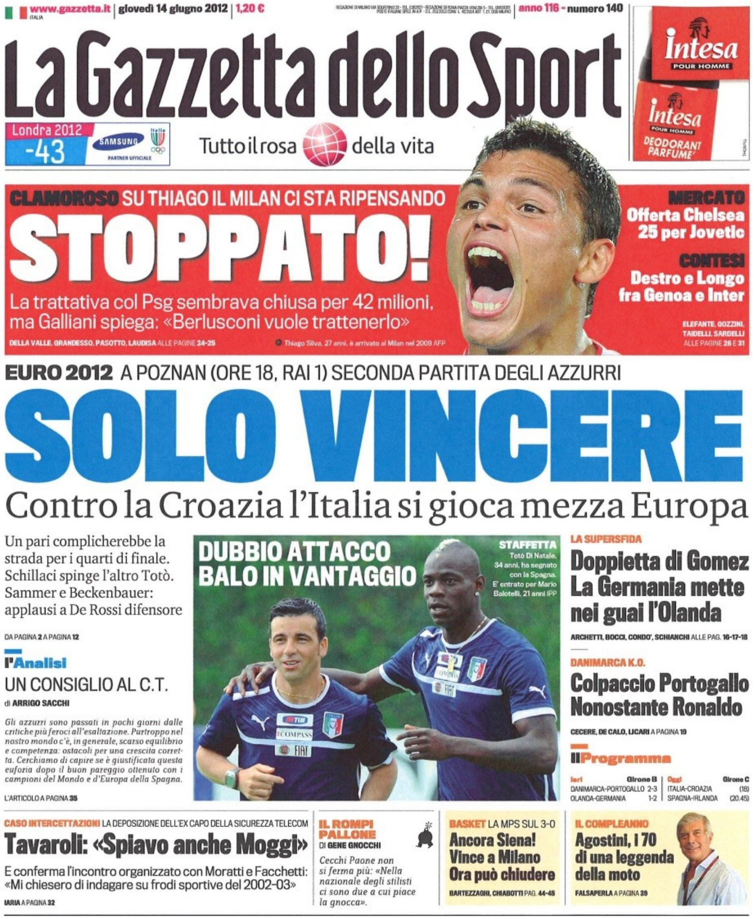
# Summary

- Introduction & Motivation
- Layout Analysis of Newspapers
- Proposed Modifications
- Experimental results
- Conclusions & Future Work

# Introduction

- Legacy newspapers available in printed form
  - Digitization → no explicit organization into meaningful higher-level components
    - Needed for automatically extracting useful information indexing
  - Approaches for automatic layout analysis often ineffective on newspapers
    - Much more complex layout
  - Objective: classification of layout blocks according to their content type.
    - adaptation of an existing approach, working on the description features and set of classes

# Objectives



www.gazzetta.it | giovedì 14 giugno 2012 | 1,20 €
ITALIA

REDAZIONE DI MILANO VIA SOLFERINO 28 [illegible] REDAZIONE DI ROMA PIAZZA VENEZIA 5 [illegible]

anno 116 - numero 140

La Gazzetta dello Sport

Londra 2012 -43 SAMSUNG PARTNER UFFICIALE

Tutto il rosa della vita

Intesa POUR HOMME
Intesa POUR HOMME DEODORANT PARFUME'

CLAMOROSO SU THIAGO IL MILAN CI STA RIPENSANDO

STOPPATO!

La trattativa col Psg sembrava chiusa per 42 milioni, ma Galliani spiega: «Berlusconi vuole trattenerlo»

DELLA VALLE, GRANDESSO, PASOTTO, LAUDISA ALLE PAGINE 24-25

Thiago Silva, 27 anni, è arrivato al Milan nel 2009 AFP

MERCATO
Offerta Chelsea 25 per Jovetic

CONTESI
Destro e Longo fra Genoa e Inter

ELEFANTE, GOZZINI, TAIDELLI, SARDELLI ALLE PAGINE 26 E 31

EURO 2012 A POZNAN (ORE 18, RAI 1) SECONDA PARTITA DEGLI AZZURRI

SOLO VINCERE

Contro la Croazia l'Italia si gioca mezza Europa

Un pari complicherebbe la strada per i quarti di finale. Schillaci spinge l'altro Totò. Sammer e Beckenbauer: applausi a De Rossi difensore

DA PAGINA 2 A PAGINA 12

l'Analisi

UN CONSIGLIO AL C.T.

di ARRIGO SACCHI

*Gli azzurri sono passati in pochi giorni dalle critiche più feroci all'esaltazione. Purtroppo nel nostro mondo c'è, in generale, scarso equilibrio e competenza: ostacoli per una crescita corretta. Cerchiamo di capire se è giustificata questa euforia dopo il buon pareggio ottenuto con i campioni del Mondo e d'Europa della Spagna.*

L'ARTICOLO A PAGINA 35

DUBBIO ATTACCO BALO IN VANTAGGIO

STAFFETTA
Totò Di Natale, 34 anni, ha segnato con la Spagna. E' entrato per Mario Balotelli, 21 anni IPP

LA SUPERSFIDA

Doppietta di Gomez La Germania mette nei guai l'Olanda

ARCHETTI, BOCCI, CONDO', SCHIANCHI ALLE PAG. 16-17-18

DANIMARCA K.O.

Colpaccio Portogallo Nonostante Ronaldo

CECERE, DE CALO, LICARI A PAGINA 19

Il Programma

| Ieri | Girone B | Oggi | Girone C |
|---|---|---|---|
| DANIMARCA-PORTOGALLO | 2-3 | ITALIA-CROAZIA | (18) |
| OLANDA-GERMANIA | 1-2 | SPAGNA-IRLANDA | (20.45) |

CASO INTERCETTAZIONI LA DEPOSIZIONE DELL'EX CAPO DELLA SICUREZZA TELECOM

Tavaroli: «Spiavo anche Moggi»

E conferma l'incontro organizzato con Moratti e Facchetti: «Mi chiesero di indagare su frodi sportive del 2002-03»

IARIA A PAGINA 32

IL ROMPI PALLONE
DI GENE GNOCCHI

*Cecchi Paone non si ferma più: «Nella nazionale degli stilisti ci sono due a cui piace la gnocca».*

BASKET LA MPS SUL 3-0

Ancora Siena! Vince a Milano Ora può chiudere

BARTEZZAGHI, CHIABOTTI PAG. 44-45

IL COMPLEANNO

Agostini, i 70 di una leggenda della moto

FALSAPERLA A PAGINA 39

- Tackling
  - use of colors
  - text blocks written on background different than the main background
  - frequent interleaving of very different text font sizes

# Document Processing and Management

- Steps
  - Document Image Understanding (layout structure, logical structure)
    - Layout Analysis
      - Segmentation
      - Component Classification
  - Document Understanding
- Layout Analysis fundamental for the quality and feasibility of Document Understanding

# Layout Analysis Procedure in DoMInUS

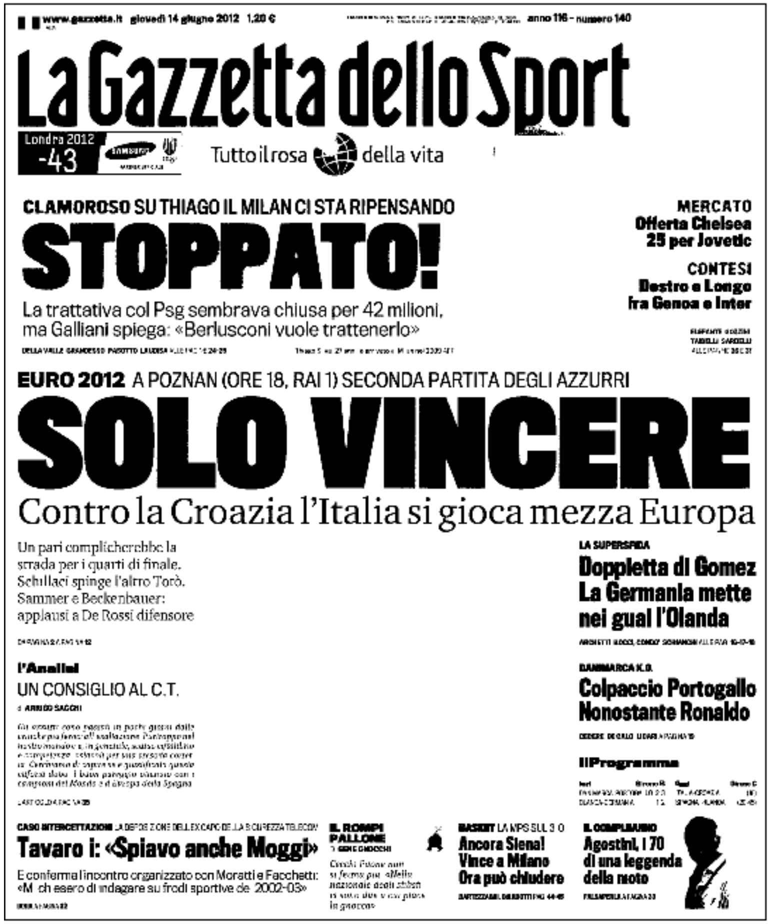
- pre-processing:
  - binarization
  - **chromatic component separation** → peculiarity #1
  - skew correction
- *classification of layout components* **in each color layer**
  - text
  - lines
  - **non-standard background** → peculiarity #2
  - images
- **text blocks identification**
  - **removal of non-textual components**
  - **extraction of text from non-standard background**
  - *text blocks aggregation* using RLSO → peculiarity #3

**Bold**: steps specifically introduced for dealing with newspapers
*Italics*: steps already present but changed for dealing with newspapers

# Layout Analysis Procedure in DoMInUS

- 1.b: artificially colored parts of the page (ignore saturation)
  - Sequence of filtered versions of the page:
    background (white); graylevel; other colors
  - Reversed background layer = color-independent binarization of the page
- 2.c: reverse all ‘Image’ connected components in each layer;
  run again the classifier: is the inverted block classified as Text?
- 3.a: remove all non-text components in the various color layers
- 3.b turn the original non-standard background into standard background;
  represent the text as standard foreground
  - Binarized image: only textual components on standard background
- 3.c: obtain aggregate text blocks using RLSO (non-Manhattan layout), *but*:
  applied as a last step; applied on a filtered image containing only text;
  applied iteratively

# Partial Processing Steps



www.gazzetta.it giovedì 14 giugno 2012 1,20 € anno 116 – numero 140

La Gazzetta dello Sport

Londra 2012 -43

Tutto il rosa della vita

CLAMOROSO SU THIAGO IL MILAN CI STA RIPENSANDO

**STOPPATO!**

La trattativa col Psg sembrava chiusa per 42 milioni, ma Galliani spiega: «Berlusconi vuole trattenerlo»

MERCATO

**Offerta Chelsea 25 per Jovetic**

CONTESI

**Destro e Longo fra Genoa e Inter**

EURO 2012 A POZNAN (ORE 18, RAI 1) SECONDA PARTITA DEGLI AZZURRI

**SOLO VINCERE**

Contro la Croazia l'Italia si gioca mezza Europa

Un pari complicherebbe la strada per i quarti di finale. Schillaci spinge l'altro Torò. Sammer e Beckenbauer: applausi a De Rossi difensore

l'Analisi

UN CONSIGLIO AL C.T.

di ARRIGO SACCHI

LA SUPERSFIDA

**Doppietta di Gomez La Germania mette nei guai l'Olanda**

DANIMARCA K.O.

**Colpaccio Portogallo Nonostante Ronaldo**

Il Programma

CASO INTERCETTAZIONI LA DEPOSIZIONE DELL'EX CAPO DELLA SICUREZZA TELECOM

**Tavaroli: «Spiavo anche Moggi»**

E conferma l'incontro organizzato con Moratti e Facchetti: «Mi chiesero di indagare su frodi sportive del 2002-03»

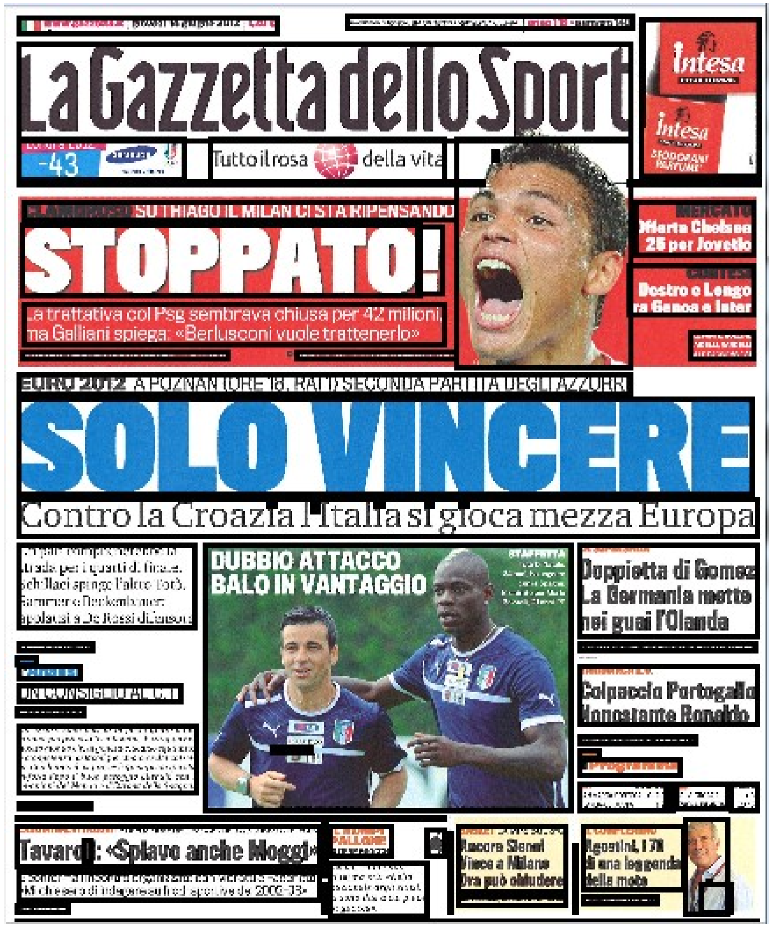
IL ROMPIPALLONE di GENE GNOCCHI

BASKET LA MPS SUL 3-0

**Ancora Siena! Vince a Milano Ora può chiudere**

IL COMPLEANNO

**Agostini, i 70 di una leggenda della moto**



# Component Type Classification Features

- block height (*h*)
- block width (*w*)
- block area ($a = w \times h$)
- block eccentricity ($w/h$)
- number of black pixels in the block (*b*)
- number of black-white transitions in the block rows (*t*)
- percentage of black pixels in the block ($b/a$)
- average number of black pixels per black-white transition ($b/t$)
- short run emphasis (*F1*): blocks containing many short runs
  - small-sized characters (e.g., newspaper articles)
- long run emphasis (*F2*): blocks containing many runs having medium length
  - quite large characters (e.g., newspaper subtitles)
- extra long run emphasis (*F3*): blocks containing few runs, all of which very long
  - text of very large size (e.g., main titles of newspaper pages).
    - Requires two parameters, *T1* and *T2*

# Component Type Classification Classes

- **Text**: a group of alphanumeric characters or symbols
  - even just one character or symbol
- **Horizontal Line**
- **Vertical Line**
- **Graphic**: an artificial image
  - (e.g., produced using vector graphics tools)
- **Image**: a (possibly halftone) raster image
- **Mixed**: a combination of text and image(s), but clearly disjoint (text within images would fall in the Image class)
- **Undefined**: none of the above
  - A portion of an image, a particularly eroded line, ...

# Component Type Classification Additional Features

- Spread: $s = n/b \times \min(w,h)^2$
  - spatial distribution of black pixels in a pattern
    - *b* = # black pixels (raising the density reduces the distance among pixels),
    - *n* = # black runs (the more the runs, the more fragmented the black zones),
    - Area of square sections:
      - $a \times sq = w \times h \times \min(w,h) / \max(w,h) = \min(w,h)^2$
- # components
  - blocks having large area and many components ~ text
  - blocks having small area and 1 component ~ character
- # black-white transitions in the block columns
  - complementary perspective with respect to feature #6
- F3 ($T_1 = 30$, $T_2 = 5$)
- F3 ($T_1 = 5$, $T_2 = 5$)

# Component Type Classification Additional Classes

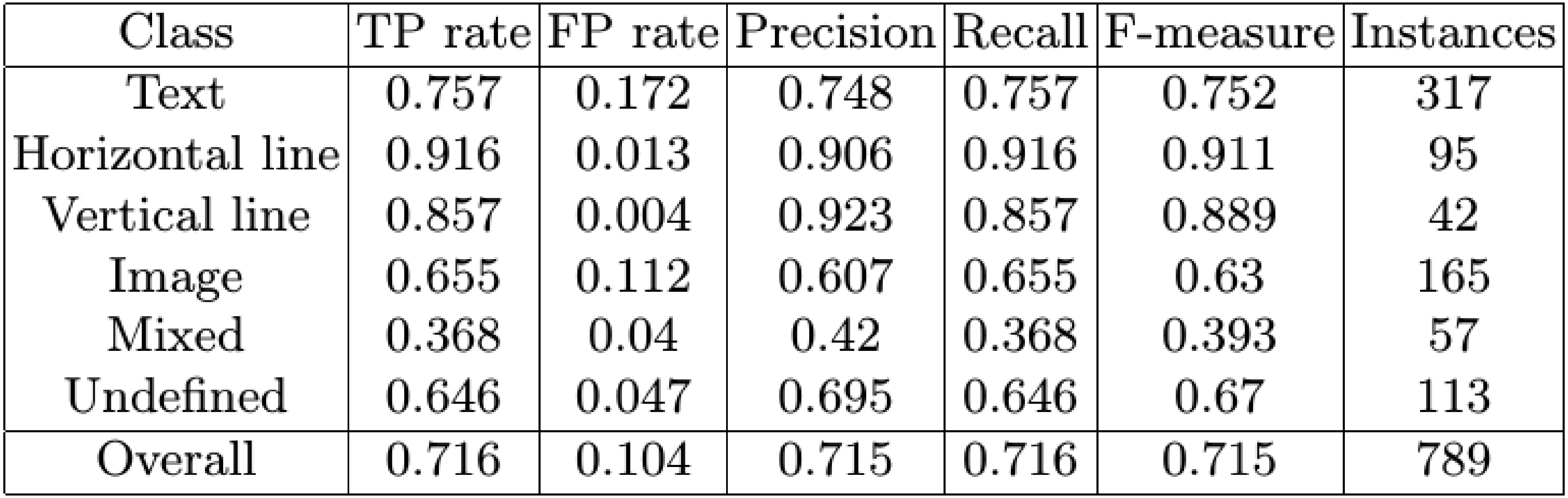
- splitting the class Text
  - Text
  - Character
  - Reverse Text
  - Reverse Character

# Experiments
# Baseline

- Dataset
  - 30 images of newspapers' first pages
    - some in color, some in black and white
  - 789 connected components
    - No graphic or diagonal line
      - However, these classes are meaningful

- Learning setting
  - 10-fold cross-validation
  - Decision tree learner J48 (WEKA)
  - Worst accuracy: Mixed
    - Very subtle (and mostly semantic) differences compared to Image, especially when they include text
    - Some newspapers superimpose text to images

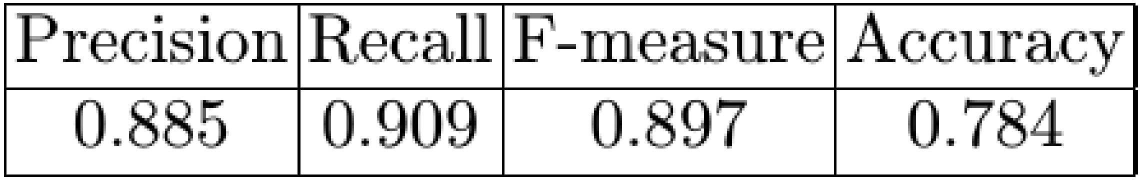
# Baseline experimental results for component type classification



| Class | TP rate | FP rate | Precision | Recall | F-measure | Instances |
|---|---|---|---|---|---|---|
| Text | 0.757 | 0.172 | 0.748 | 0.757 | 0.752 | 317 |
| Horizontal line | 0.916 | 0.013 | 0.906 | 0.916 | 0.911 | 95 |
| Vertical line | 0.857 | 0.004 | 0.923 | 0.857 | 0.889 | 42 |
| Image | 0.655 | 0.112 | 0.607 | 0.655 | 0.63 | 165 |
| Mixed | 0.368 | 0.04 | 0.42 | 0.368 | 0.393 | 57 |
| Undefined | 0.646 | 0.047 | 0.695 | 0.646 | 0.67 | 113 |
| Overall | 0.716 | 0.104 | 0.715 | 0.716 | 0.715 | 789 |

  - Last row = weighted average for performance columns, total for the number of components

- Layout Analysis performance on 45 additional newspapers:

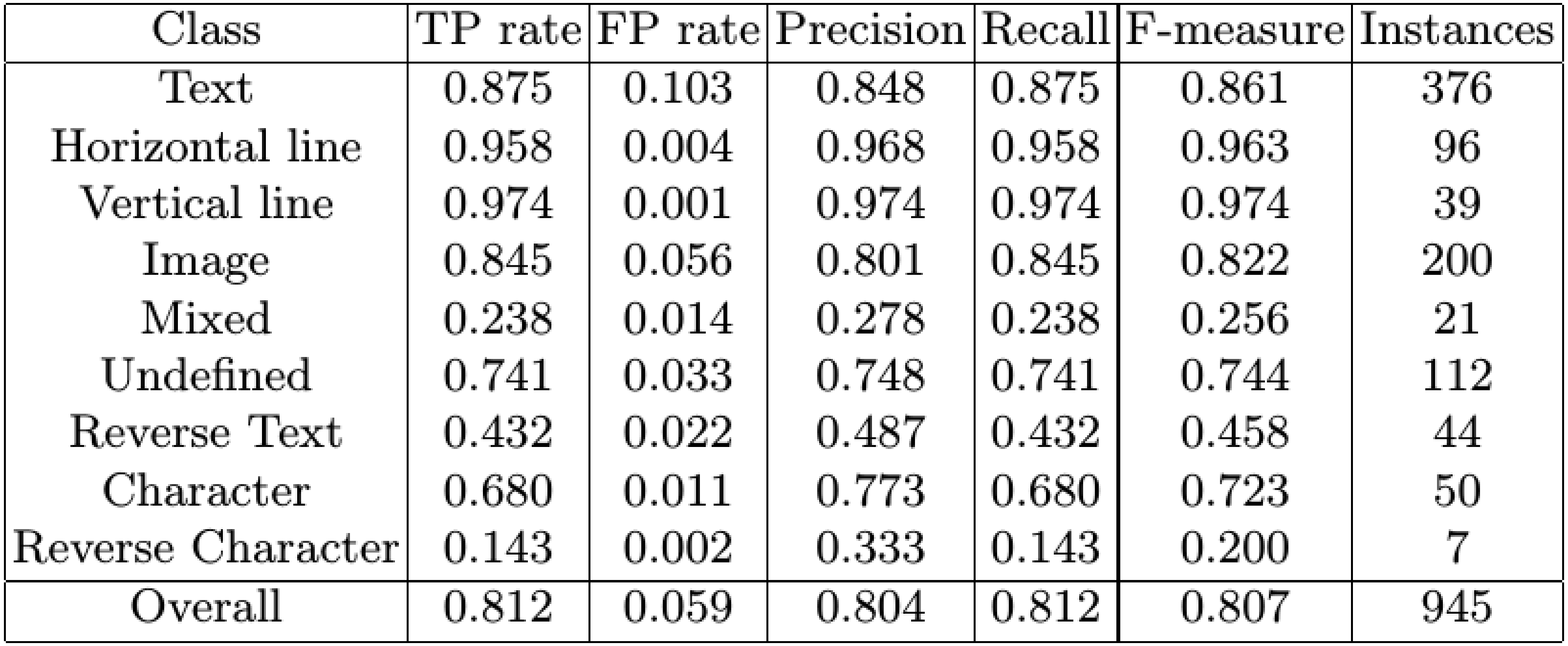


| Precision | Recall | F-measure | Accuracy |
|---|---|---|---|
| 0.885 | 0.909 | 0.897 | 0.784 |

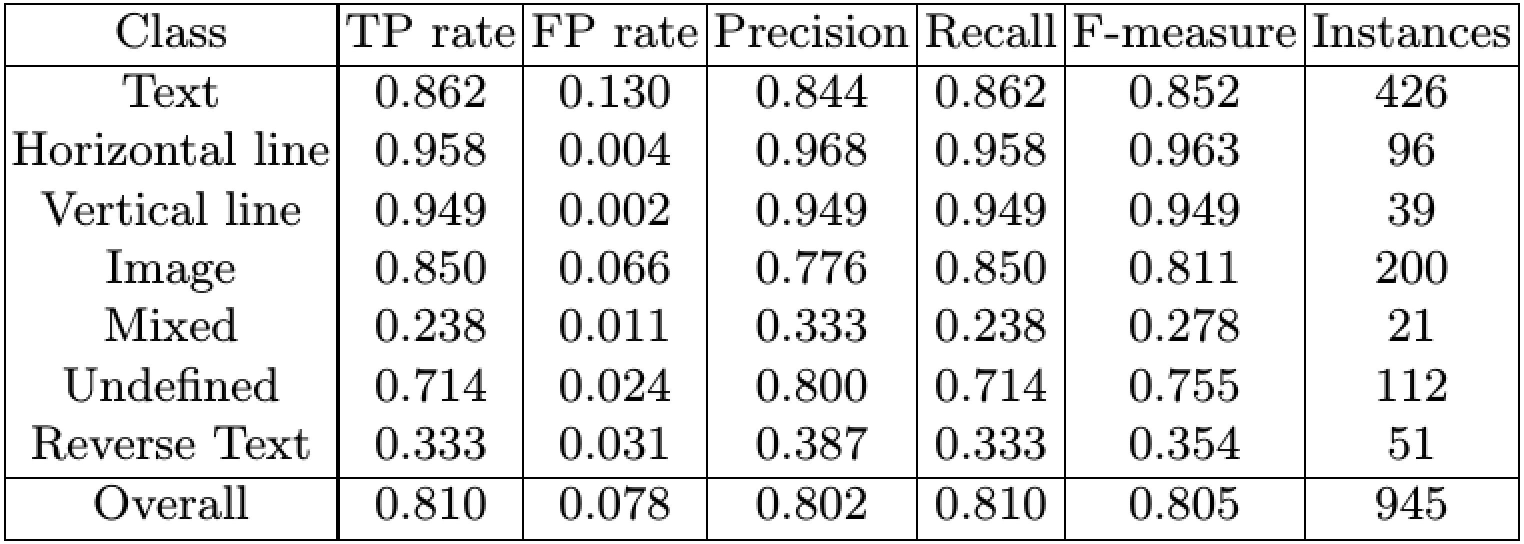
# Experiments

- New dataset made up of 10 newspapers
    - Previous dataset unavailable
- Always used the extended set of features
    - *F3* ($T_1$ = 30, $T_2$ = 5) never considered
- Different set of classes
    - same classes as the baseline
    - separate class for reversed text only
    - specific classes for text/characters, normal/reversed
- All settings much better than the baseline
    - Some better on some classes, some better on others

# Experimental results with additional features and classes



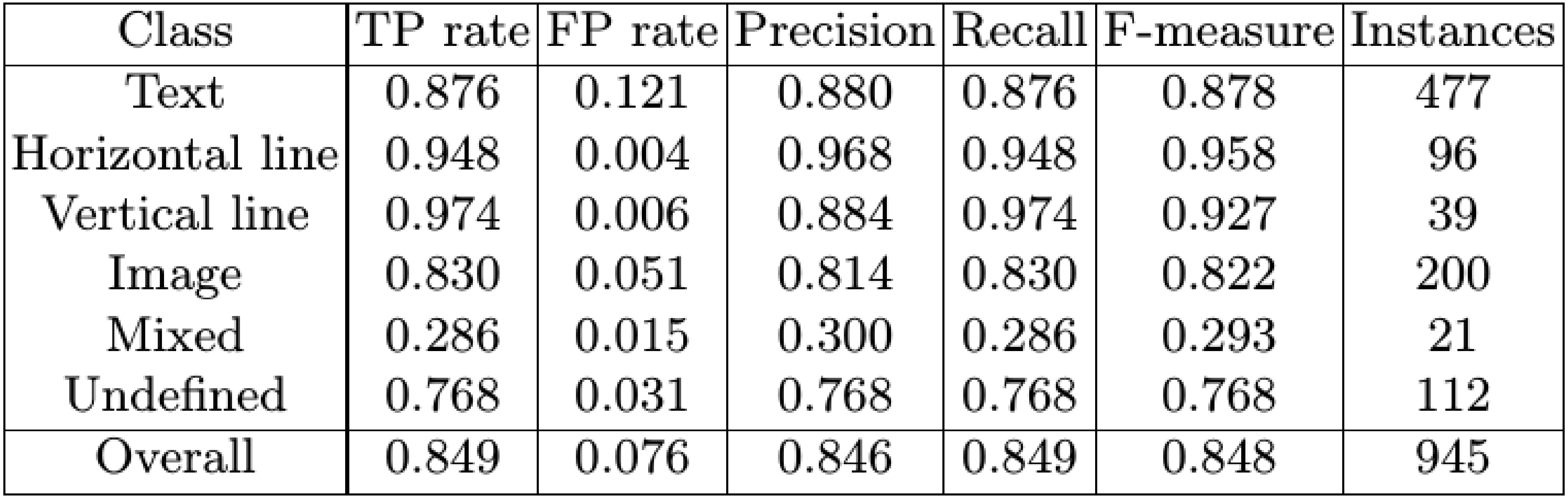
| Class | TP rate | FP rate | Precision | Recall | F-measure | Instances |
|---|---|---|---|---|---|---|
| Text | 0.875 | 0.103 | 0.848 | 0.875 | 0.861 | 376 |
| Horizontal line | 0.958 | 0.004 | 0.968 | 0.958 | 0.963 | 96 |
| Vertical line | 0.974 | 0.001 | 0.974 | 0.974 | 0.974 | 39 |
| Image | 0.845 | 0.056 | 0.801 | 0.845 | 0.822 | 200 |
| Mixed | 0.238 | 0.014 | 0.278 | 0.238 | 0.256 | 21 |
| Undefined | 0.741 | 0.033 | 0.748 | 0.741 | 0.744 | 112 |
| Reverse Text | 0.432 | 0.022 | 0.487 | 0.432 | 0.458 | 44 |
| Character | 0.680 | 0.011 | 0.773 | 0.680 | 0.723 | 50 |
| Reverse Character | 0.143 | 0.002 | 0.333 | 0.143 | 0.200 | 7 |
| Overall | 0.812 | 0.059 | 0.804 | 0.812 | 0.807 | 945 |



| Class | TP rate | FP rate | Precision | Recall | F-measure | Instances |
|---|---|---|---|---|---|---|
| Text | 0.862 | 0.130 | 0.844 | 0.862 | 0.852 | 426 |
| Horizontal line | 0.958 | 0.004 | 0.968 | 0.958 | 0.963 | 96 |
| Vertical line | 0.949 | 0.002 | 0.949 | 0.949 | 0.949 | 39 |
| Image | 0.850 | 0.066 | 0.776 | 0.850 | 0.811 | 200 |
| Mixed | 0.238 | 0.011 | 0.333 | 0.238 | 0.278 | 21 |
| Undefined | 0.714 | 0.024 | 0.800 | 0.714 | 0.755 | 112 |
| Reverse Text | 0.333 | 0.031 | 0.387 | 0.333 | 0.354 | 51 |
| Overall | 0.810 | 0.078 | 0.802 | 0.810 | 0.805 | 945 |

# Experimental results with additional features only

- Overall weighted averaged F-measure significantly better than the other settings
  - Real improvement due to the extension to the set of features



| Class | TP rate | FP rate | Precision | Recall | F-measure | Instances |
|---|---|---|---|---|---|---|
| Text | 0.876 | 0.121 | 0.880 | 0.876 | 0.878 | 477 |
| Horizontal line | 0.948 | 0.004 | 0.968 | 0.948 | 0.958 | 96 |
| Vertical line | 0.974 | 0.006 | 0.884 | 0.974 | 0.927 | 39 |
| Image | 0.830 | 0.051 | 0.814 | 0.830 | 0.822 | 200 |
| Mixed | 0.286 | 0.015 | 0.300 | 0.286 | 0.293 | 21 |
| Undefined | 0.768 | 0.031 | 0.768 | 0.768 | 0.768 | 112 |
| Overall | 0.849 | 0.076 | 0.846 | 0.849 | 0.848 | 945 |

# Conclusions

- Adaptation of existing approach to block type classification of digitized newspapers
    - colors, text on non-standard background, frequent interleaving of very different font sizes
    - Implemented and embedded in DoMInUS
    - Experimental results showed that using additional features may be beneficial
- Future work
    - Larger dataset
    - Effect on the final layout analysis performance