Character Recognition Internals

Dr. István Marosi
Nuance-Recognita, Inc., Hungary

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
  - User assisted correction
  - Result exportation

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
    - Get image
      - B/W Scanning
      - Gray Scanning
      - Color Scanning
      - Load from image file
    - Preprocess image
  - Layout recognition
  - Text recognition
  - User assisted correction
  - Result exportation

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
    - Get image
    - Preprocess image
      - Color separation
      - Thresholding
      - Despeckling
      - Rotation
      - Deskewing
  - Layout recognition
  - Text recognition
  - User assisted correction
  - Result exportation

Color Separation

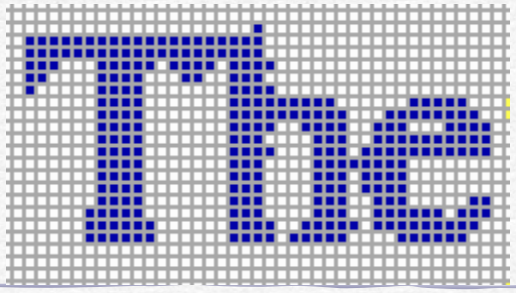
De-speckle, de-skew

The Preprocessed Image

Joined chars

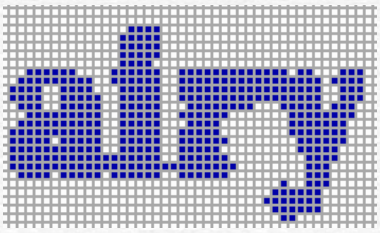
The Preprocessed Image

Joined chars



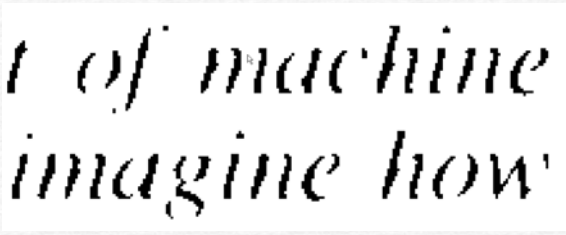
The Preprocessed Image

Joined chars



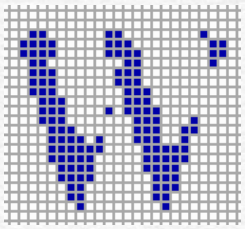
The Preprocessed Image

Broken chars



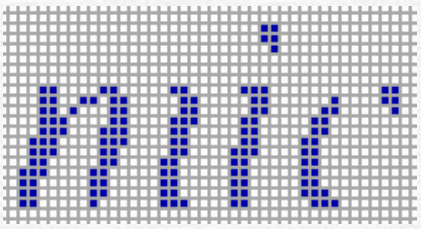
The Preprocessed Image

Broken chars



The Preprocessed Image

Broken chars

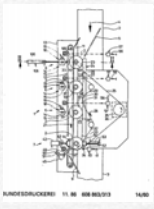


OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
    - Text zones
      - Columns of flowed text
      - Tables
      - Inverse text
    - Graphic zones
  - Text recognition
  - User assisted correction
  - Result exportation

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
    - Text zones
    - Graphic zones
      - Line Art
      - Photo
  - Text recognition
  - User assisted correction
  - Result exportation



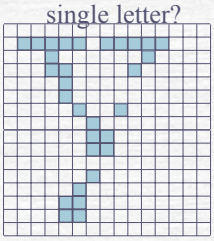
OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
    - *Segmentation*
    - *Calculation of Feature Vector Elements*
    - *Classification*
    - *Language Analysis*
    - *Voting*
  - User assisted correction
  - Result exportation

Istvan Marosi

Segmentation

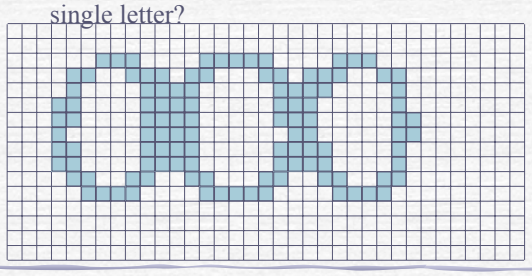
- What are those pixel groups belonging to a single letter?



Istvan Marosi

Segmentation

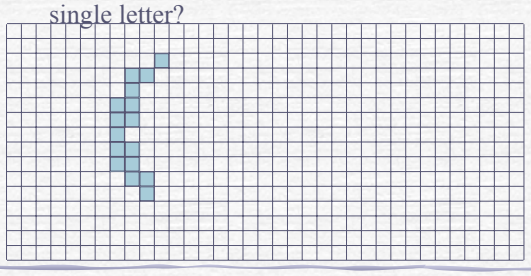
- What are those pixel groups belonging to a single letter?



Istvan Marosi

Segmentation

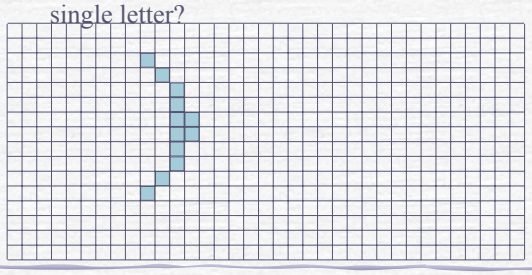
- What are those pixel groups belonging to a single letter?



Istvan Marosi

Segmentation

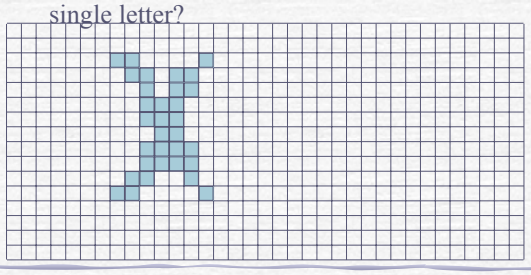
- What are those pixel groups belonging to a single letter?



Istvan Marosi

Segmentation

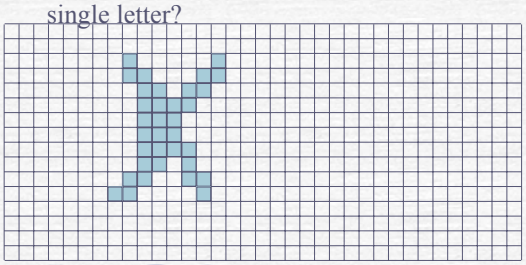
- What are those pixel groups belonging to a single letter?



Istvan Marosi

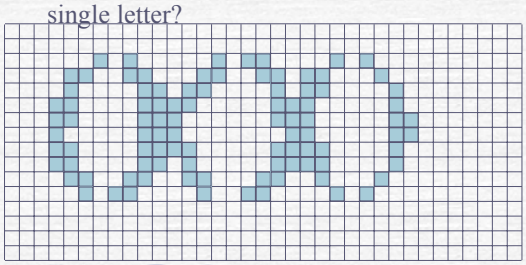
Segmentation

- What are those pixel groups belonging to a single letter?



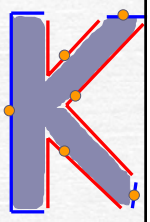
Segmentation

- What are those pixel groups belonging to a single letter?



OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
    - *Segmentation*
    - *Calculation of Feature Vector Elements*
    - *Classification*
    - *Language Analysis*
    - *Voting*
  - User assisted correction
  - Result exportation



Calculation of FV Elements:
Contour Tracing

- Find a (new) white-black transition
- Follow the "edge" of the pixels using the MIN or MAX rule
- Administrate the already traced white-black transitions
- Collect information while going around
- And repeat the process on new shapes ...

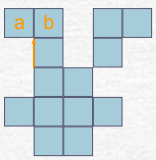
Contour Tracing

- Find a (new) white-black transition

- Follow the "edge" of the pixels using the MIN or MAX rule
- Administrate the already traced white-black transitions
- Collect information while going around
- And repeat the process on new shapes ...

Contour Tracing

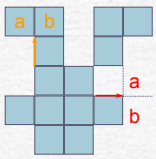
- Find a (new) white-black transition
- Follow the "edge" of the pixels using the MIN or MAX rule



if black(a) then turn(ccw)
else if black(b) then forward
else turn(cw)

Contour Tracing

- Find a (new) white-black transition
- Follow the "edge" of the pixels using the MIN or MAX rule



if black(a) then turn(ccw)
else if black(b) then forward
else turn(cw)

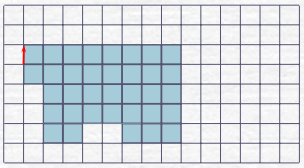
if white(b) then turn(cw)
else if white(a) then forward
else turn(ccw)

Contour Tracing

- Find a (new) white-black transition
- Follow the "edge" of the pixels using the MIN or MAX rule
- Administrate the already traced white-black transitions 
- Collect information while going around
- And repeat the process on new shapes ...

Some Easily Calculatable Data

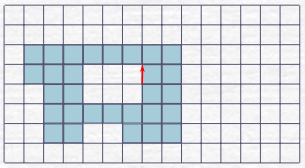
- Problem #1



Turning CW: $I_n=I_{n-1}+1$
Turning CCW: $I_n=I_{n-1}-1$
Going Forward: $I_n=I_{n-1}$

Some Easily Calculatable Data

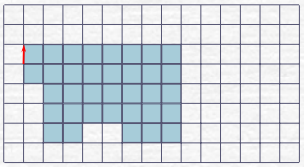
- Problem #2



Turning CW: $I_n=I_{n-1}+1$
Turning CCW: $I_n=I_{n-1}-1$
Going Forward: $I_n=I_{n-1}$

Some Easily Calculatable Data

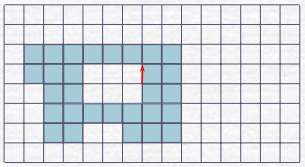
- Problem #3



Going Up: $I_n=I_{n-1}-X_n$
Going Down: $I_n=I_{n-1}+X_n$
Going Right: $I_n=I_{n-1}$
Going Left: $I_n=I_{n-1}$

Some Easily Calculatable Data

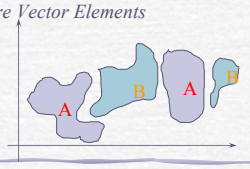
- Problem #4



Going Up: $I_n=I_{n-1}-X_n$
Going Down: $I_n=I_{n-1}+X_n$
Going Right: $I_n=I_{n-1}$
Going Left: $I_n=I_{n-1}$

OCR Internals

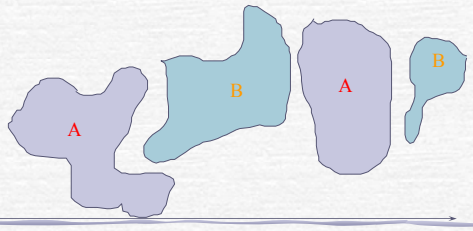
- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
    - *Segmentation*
    - *Calculation of Feature Vector Elements*
    - *Classification*
    - *Language Analysis*
    - *Voting*
  - User assisted correction
  - Result exportation



Istvan Marosi

Classification; Training models

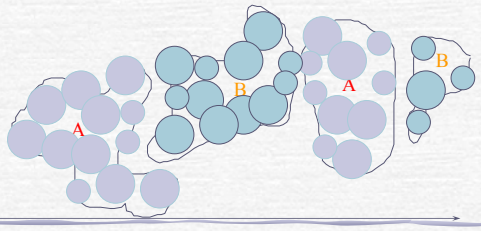
- Restricted Coulomb Energy (RCE) Network (Dr. Leon Cooper, Dr. Charles Elbaum)



Istvan Marosi

Classification; Training models

- Restricted Coulomb Energy (RCE) Network (Dr. Leon Cooper, Dr. Charles Elbaum)



Istvan Marosi

Classification; Training models

- Nestor Learning System (NLS)

Istvan Marosi

Classification; Training models

- Nestor Learning System (NLS)

Default radius R_{max}

Istvan Marosi

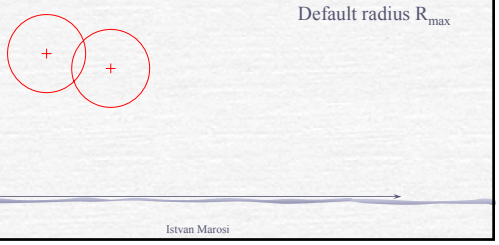
Classification; Training models

- Nestor Learning System (NLS)

Istvan Marosi

Classification; Training models

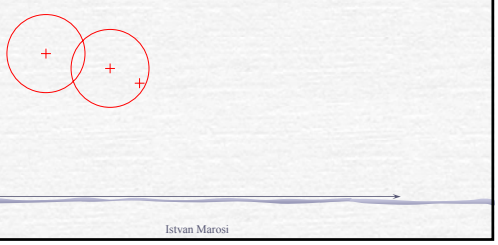
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

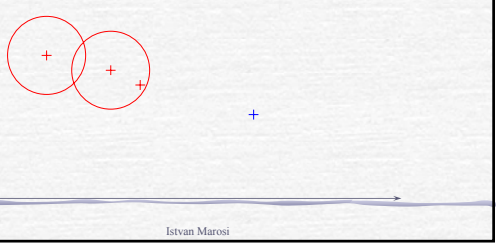
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

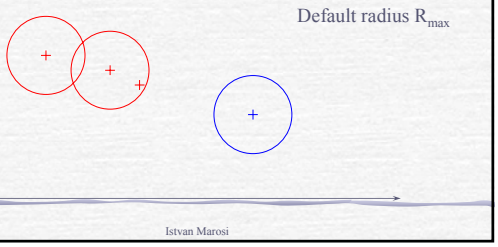
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

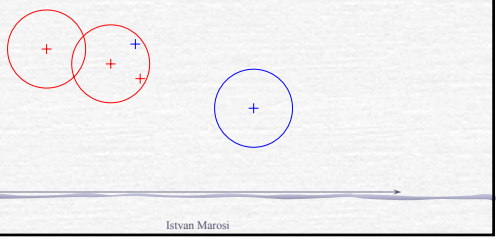
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

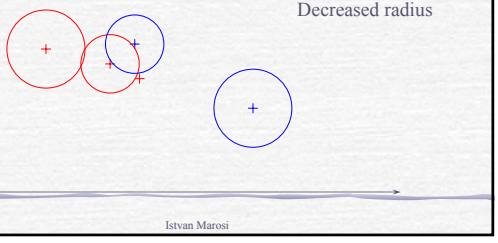
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

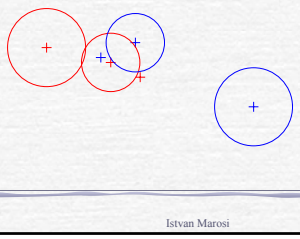
- Nestor Learning System (NLS)



Istvan Marosi

Classification; Training models

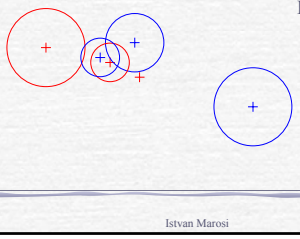
- Nestor Learning System (NLS)



Classification; Training models

- Nestor Learning System (NLS)

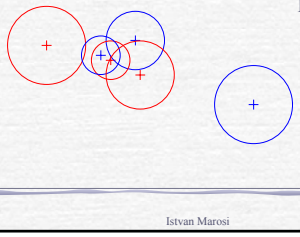
Decreased radius R_{min}



Classification; Training models

- Nestor Learning System (NLS)

Pass 2
Decreased radius



OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
    - *Segmentation*
    - *Calculation of Feature Vector Elements*
    - *Classification*
    - *Language Analysis*
    - *Voting*
  - User assisted correction
  - Result exportation

children

Voting

- Text recognition in OmniPage Pro
  - OCR Engines available:
    - Caere's engine *(codename: Salt & Pepper)*
    - Recognita's engine *(codename: Paprika)*
    - ScanSoft's engine *(codename: Fireworx)*
    - Redesigned engine *(codename: Mango)*

Voting

- Text recognition in OmniPage Pro
  - OCR Engines available:
    - Caere's engine *(Salt & Pepper)*
      - Uses a Matrix Matching based algorithm
        - feature set: 40 cells of an 8x5 grid
        - good overall description of a shape
        - weaker at detailed structure
    - Recognita's engine *(Paprika)*
      - Uses a Contour Tracing based algorithm
        - feture set: convex and concave arcs on the contour
        - good detailed description of a shape
        - weaker at overall structure

Voting

- Text recognition in OmniPage Pro
  - OCR Engines available:
    - Caere's engine *(Salt & Pepper)*
    - Recognita's engine *(Paprika)*
    - ScanSoft's engine *(Fireworx)*
    - Redesigned engine *(Mango)*
  - Segmentation algorithms:

Voting

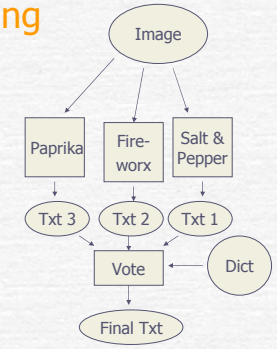
- Text recognition in OmniPage Pro
  - OCR Engines available:
    - Caere's engine *(Salt & Pepper)*
    - Recognita's engine *(Paprika)*
    - ScanSoft's engine *(Fireworx)*
    - Redesigned engine *(Mango)*
  - Segmentation algorithms:
    - Developed by independent groups
    - Have different strengths and weaknesses

Voting

- Text recognition in OmniPage Pro
  - OCR Engines available
  - Segmentation algorithms
- Conclusion:
  - They are complementary
  - Let's create a voting system

Voting

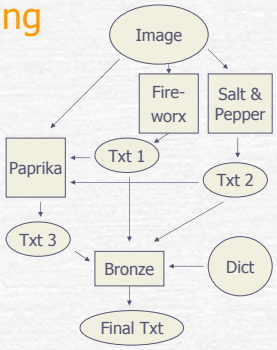
- Voting strategies
  - External „Black box" voting



~20% gain

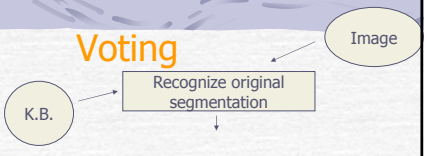
Voting

- Voting strategies
  - External „Black box" voting
  - Internal „Shape" voting



Voting

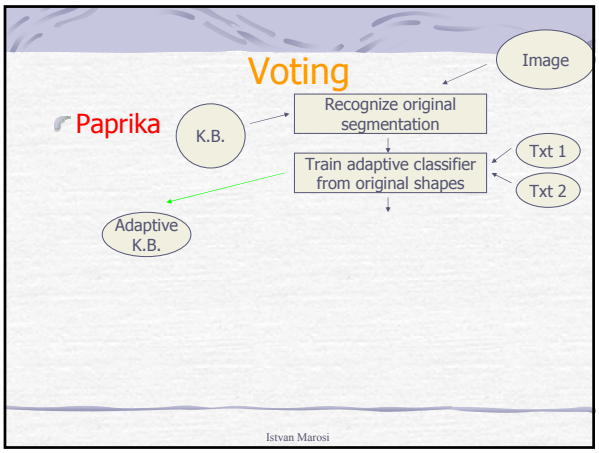
- Paprika

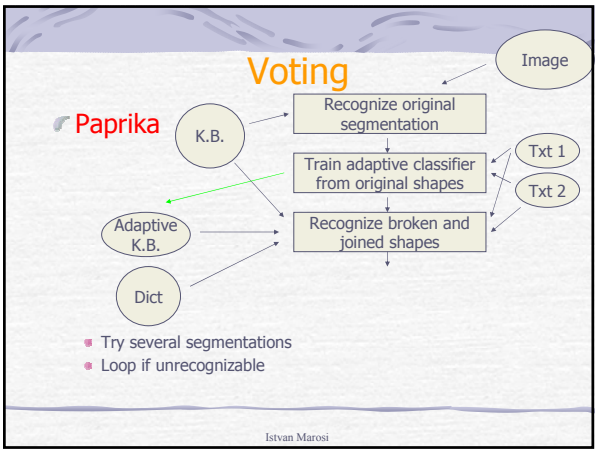


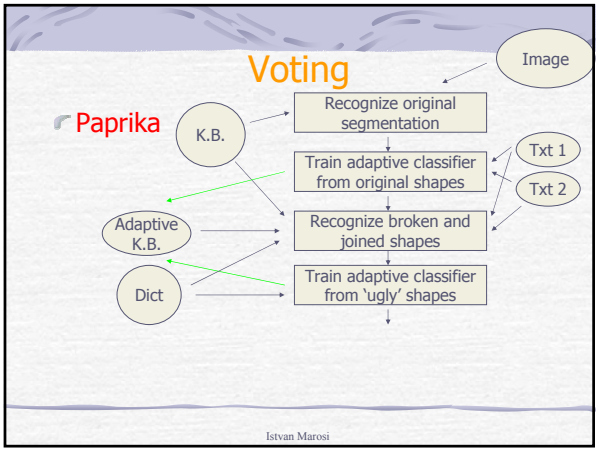
Original segmentation:

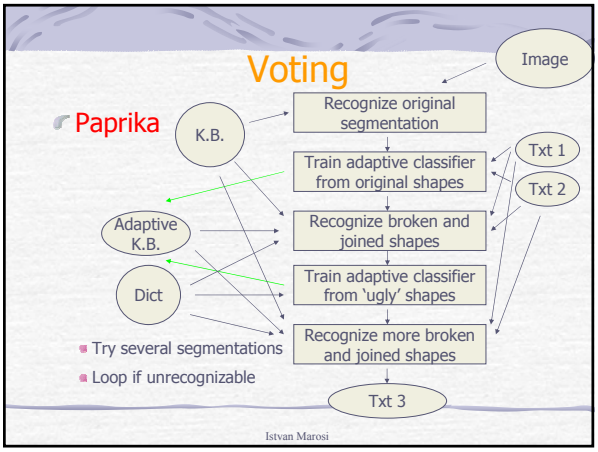
Every independent connected component is a character

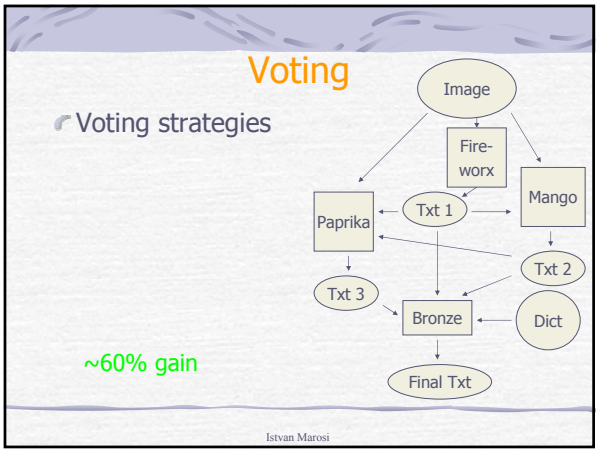
Good segmentation: recognize
Bad segmentation: reject

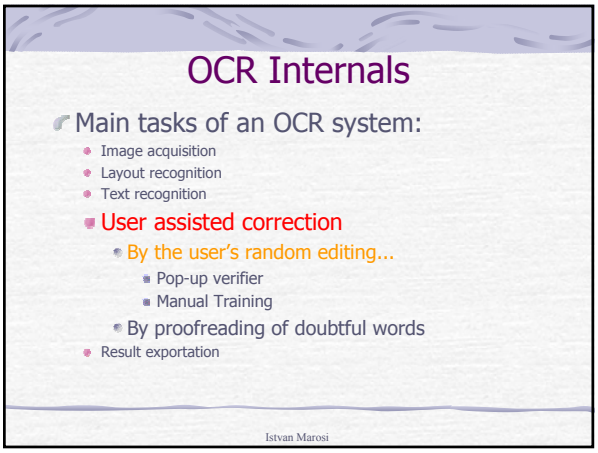

Voting
Image
Paprika
K.B.
Recognize original segmentation
Train adaptive classifier from original shapes
Txt 1
Txt 2
Adaptive K.B.
Istvan Marosi


Voting
Image
Paprika
K.B.
Recognize original segmentation
Train adaptive classifier from original shapes
Txt 1
Txt 2
Adaptive K.B.
Recognize broken and joined shapes
Dict
Try several segmentations
Loop if unrecognizable
Istvan Marosi


Voting
Image
Paprika
K.B.
Recognize original segmentation
Train adaptive classifier from original shapes
Txt 1
Txt 2
Adaptive K.B.
Recognize broken and joined shapes
Dict
Train adaptive classifier from 'ugly' shapes
Istvan Marosi


Voting
Image
Paprika
K.B.
Recognize original segmentation
Train adaptive classifier from original shapes
Txt 1
Txt 2
Adaptive K.B.
Recognize broken and joined shapes
Dict
Train adaptive classifier from 'ugly' shapes
Recognize more broken and joined shapes
Try several segmentations
Loop if unrecognizable
Txt 3
Istvan Marosi


Voting
Voting strategies
Image
Fire-worx
Mango
Paprika
Txt 1
Txt 2
Txt 3
Bronze
Dict
Final Txt
~60% gain
Istvan Marosi


OCR Internals
Main tasks of an OCR system:
Image acquisition
Layout recognition
Text recognition
User assisted correction
By the user's random editing...
Pop-up verifier
Manual Training
By proofreading of doubtful words
Result exportation
Istvan Marosi

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
  - User assisted correction
    - By the user's random editing...
    - By proofreading of doubtful words
      - Correct: User dictionary
      - Changed: IntelliTrain
        - Remember trained characters
        - Apply them on following pages
  - Result exportation

Istvan Marosi

IntelliTrain

Recognized word: *sorneUüng*

Istvan Marosi

IntelliTrain

Recognized word: *sorneUüng*

Fixed word: *something*

Istvan Marosi

IntelliTrain

Recognized word: *sorneUüng*

Fixed word: *something*

Istvan Marosi

IntelliTrain

Recognized word: *sorneUüng*

Fixed word: *something*

Substitutions found: *m* → *rn*

thi → *Uü*

Istvan Marosi

IntelliTrain

Recognized word: *sorneUüng*

Fixed word: *something*

Substitutions found: *m* → *rn*

thi → *Uü*

Perform automatically:

- Learn image pattern and substitution info
- Find similar substituted ('*blue*') text on actual page
- Match against pattern of substitution and correct
- Find such errors on following pages, too

Istvan Marosi

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
  - User assisted correction
  - Result exportation
    - Combine pages into a Document
      - Header / Footer recognition
      - Page numbers
      - Hyperlinks (e.g. „See Table 20")
    - Save results

OCR Internals

- Main tasks of an OCR system:
  - Image acquisition
  - Layout recognition
  - Text recognition
  - User assisted correction
  - Result exportation
    - Combine pages into a Document
    - Save results
      - doc file
      - e-mail
      - Speech synthesizer