Recognition of Multi-Fonts Character in Early-Modern Printed Books

Chisato Ishikawa(1), Naomi Ashida(1)*, Yurie Enomoto(1), Masami Takata(1), Tsukasa Kimesawa(2) and Kazuki Joe(1)

(1) Nara Women’s University, Japan

(2) National Diet Library, Japan

* Currently work for Mitsubishi Electric co

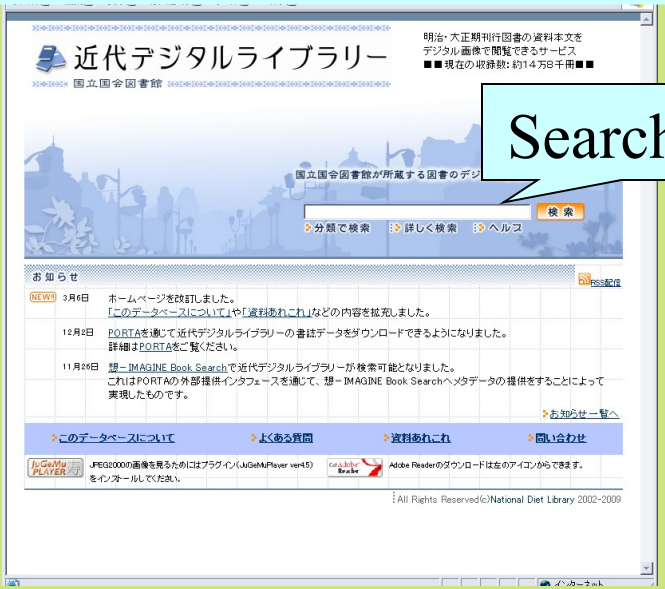
Contents

- Introduction
- Multi-fonts character recognition
  - Feature extraction from character images
  - Learning method for feature
- Experiments
  - Improvement of pre-process
- Conclusions and future work

Introduction

- The Digital Library from the Meiji Era
  (Supported by the National Diet Library in Japan)
  - Digital archive: Books published in the Meiji and Taisho eras
    1868-1926

The digital data are opened at the project Web site



Search box

Top page

Data Viewer

Introduction

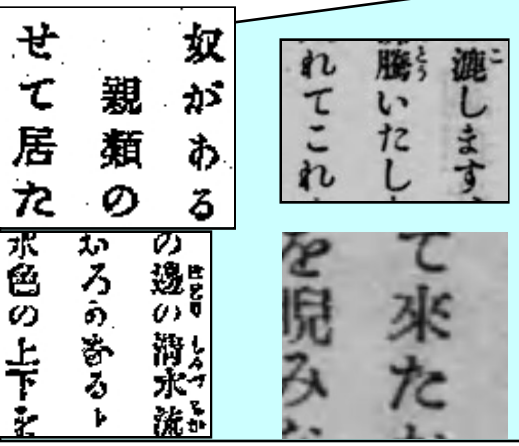
Main bodies of books

Full text search, text function:
Not supported

Image data

Conversion

Text data



–Too many kinds of fonts

–Existence of old characters

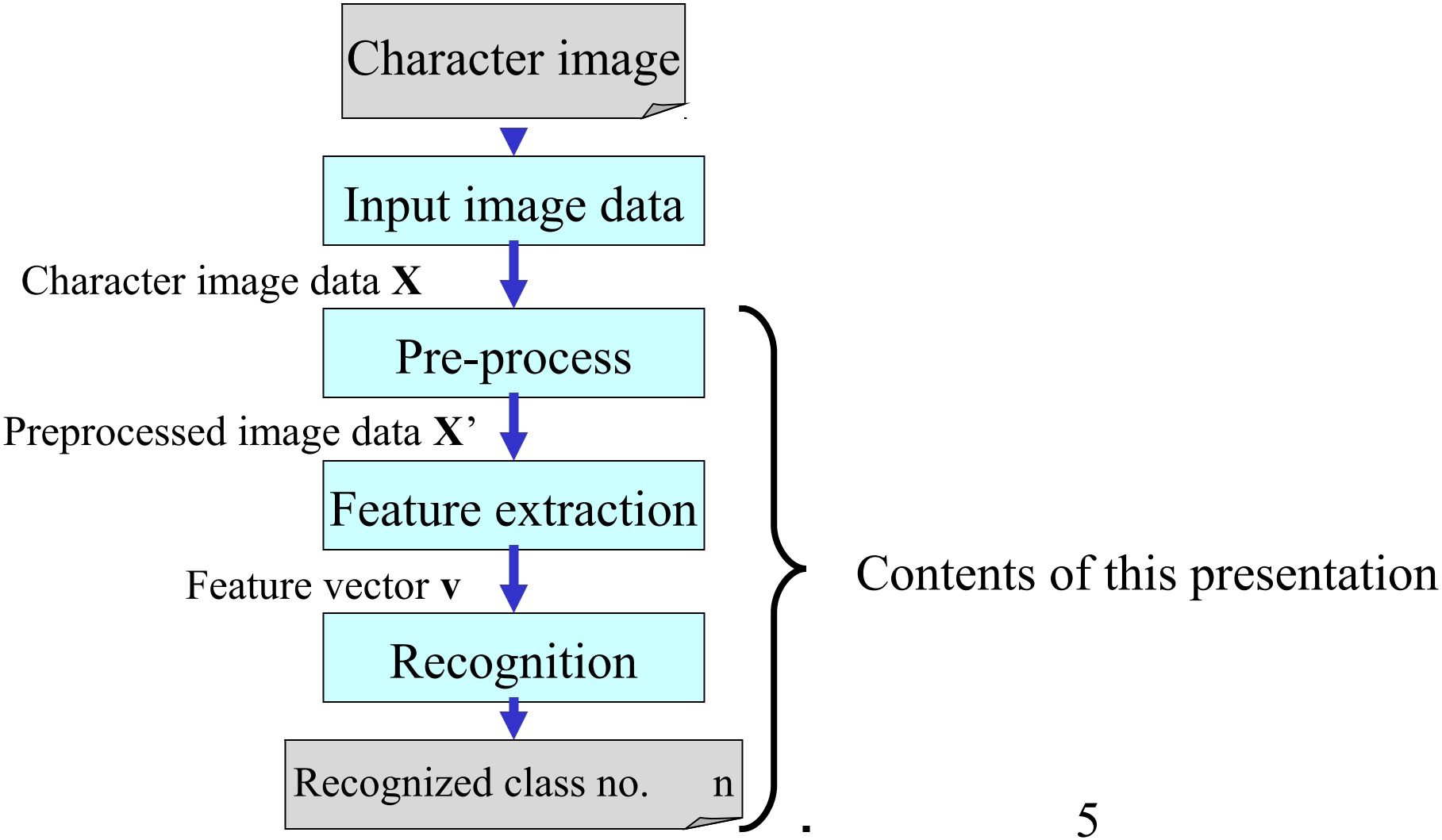
–Very noisy image

Existence OCRs are not applicable.

Our goal

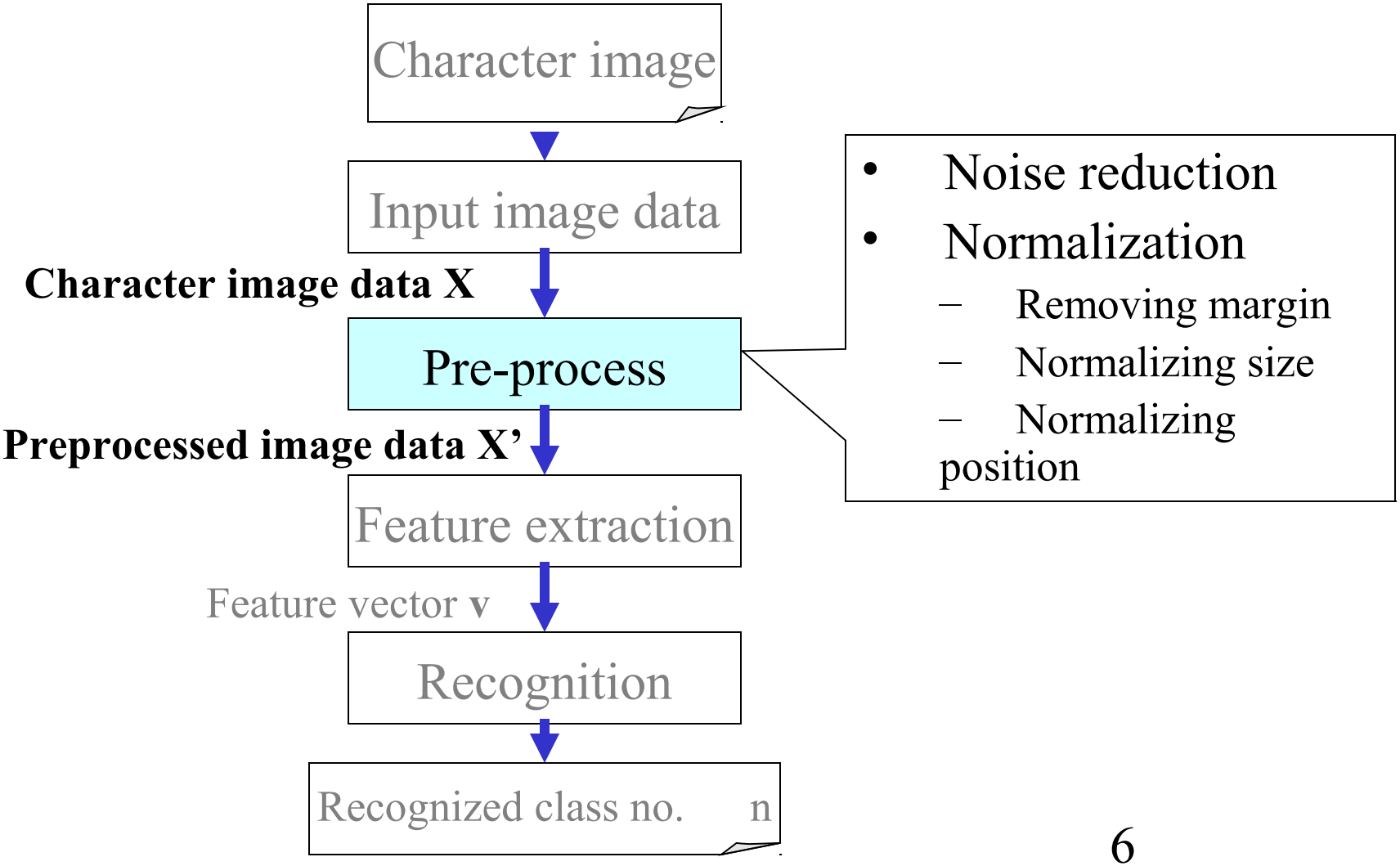
Development of an OCR for multi-fonts character in early-modern printed books

Flow of OCR


Character image
Input image data
Character image data **X**
Pre-process
Preprocessed image data **X**'
Feature extraction
Feature vector **v**
Recognition
Recognized class no. n
Contents of this presentation

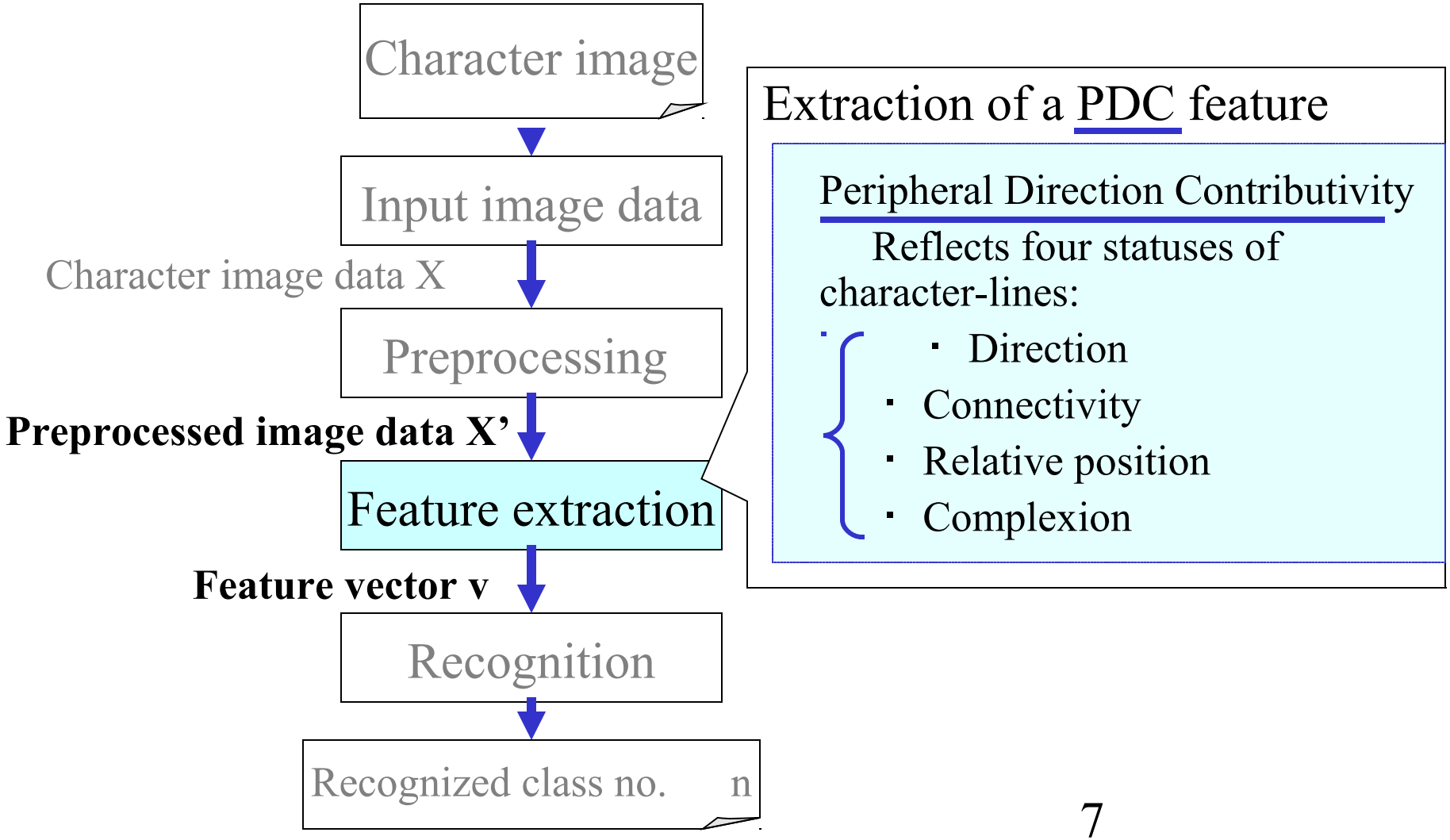
Flow of our OCR

Pre-process



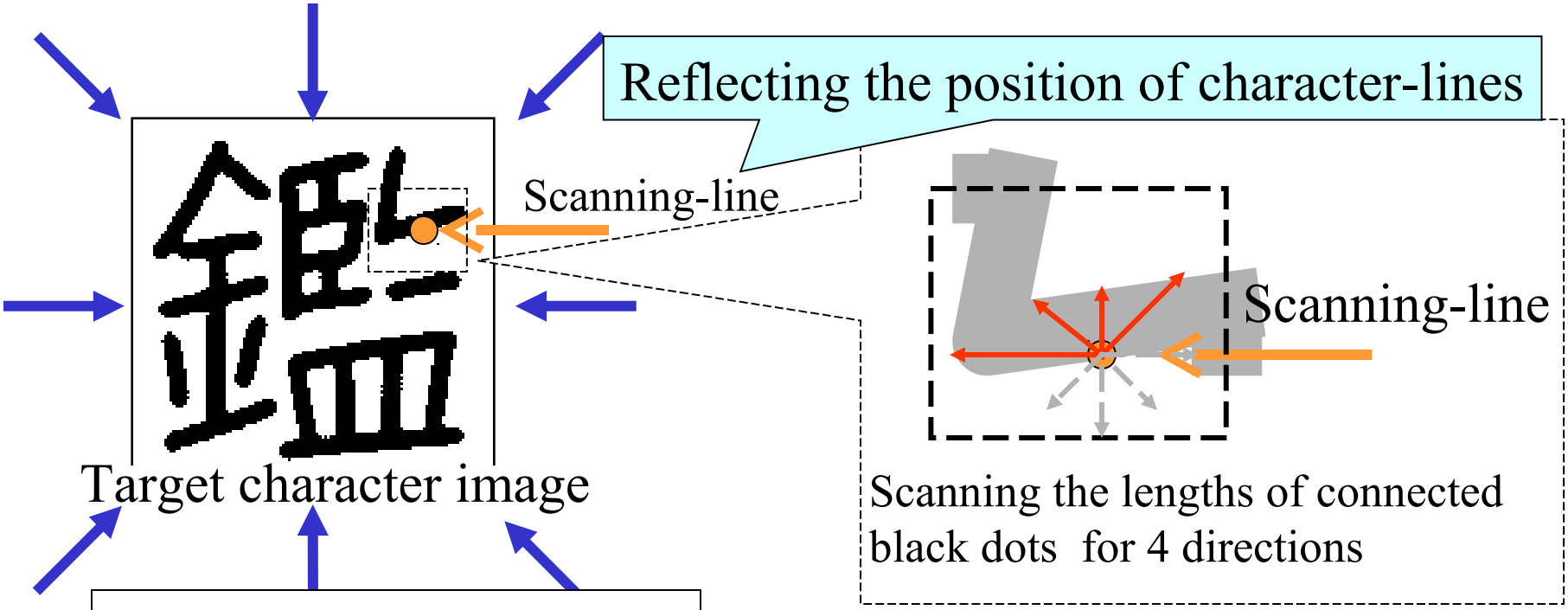
Flow of our OCR

Feature Extraction



PDC Feature

Scanning from 8 directions

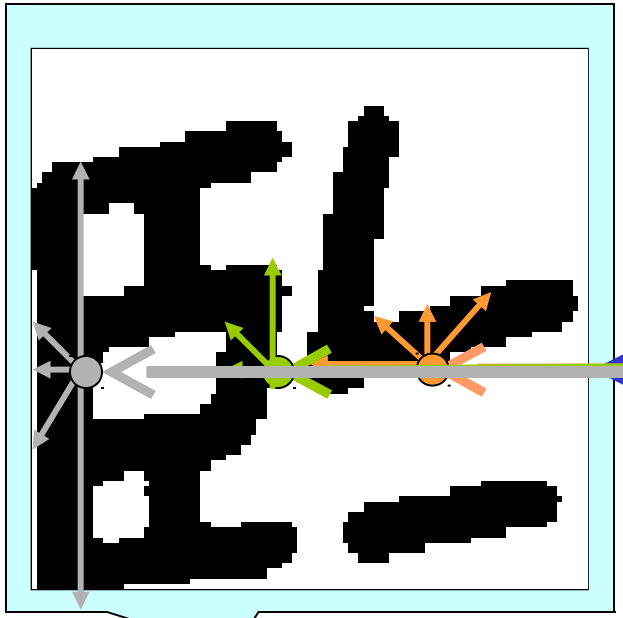


Direction contributivity is calculated from the scanned lengths

Reflecting the direction and the connectivity of character-lines

PDC Feature

Reflecting the complexity of character-lines



Scanning-line

1st depth

2nd depth

3rd depth

Direction contributivity

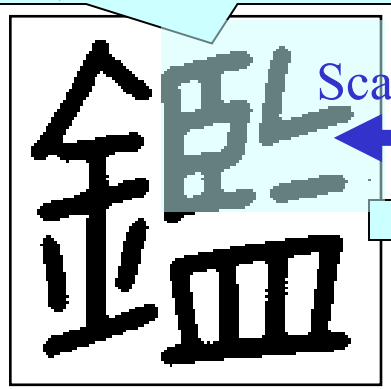
Direction contributivity

Direction contributivity

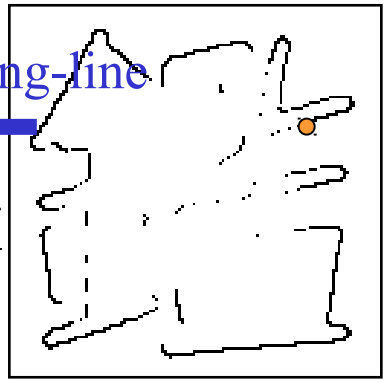
Deeper level's

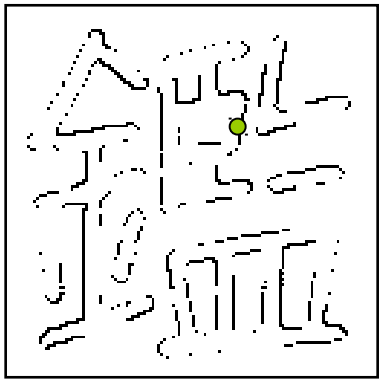
are not 0 → Complex character-lines

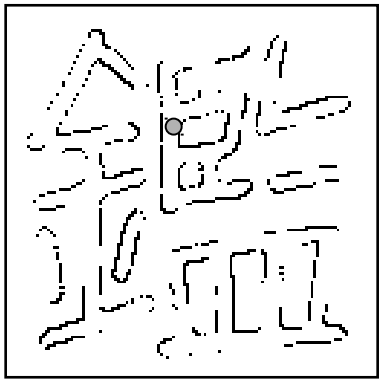
are 0 → Simple character-lines



Scanning-line







Base image

1st depth

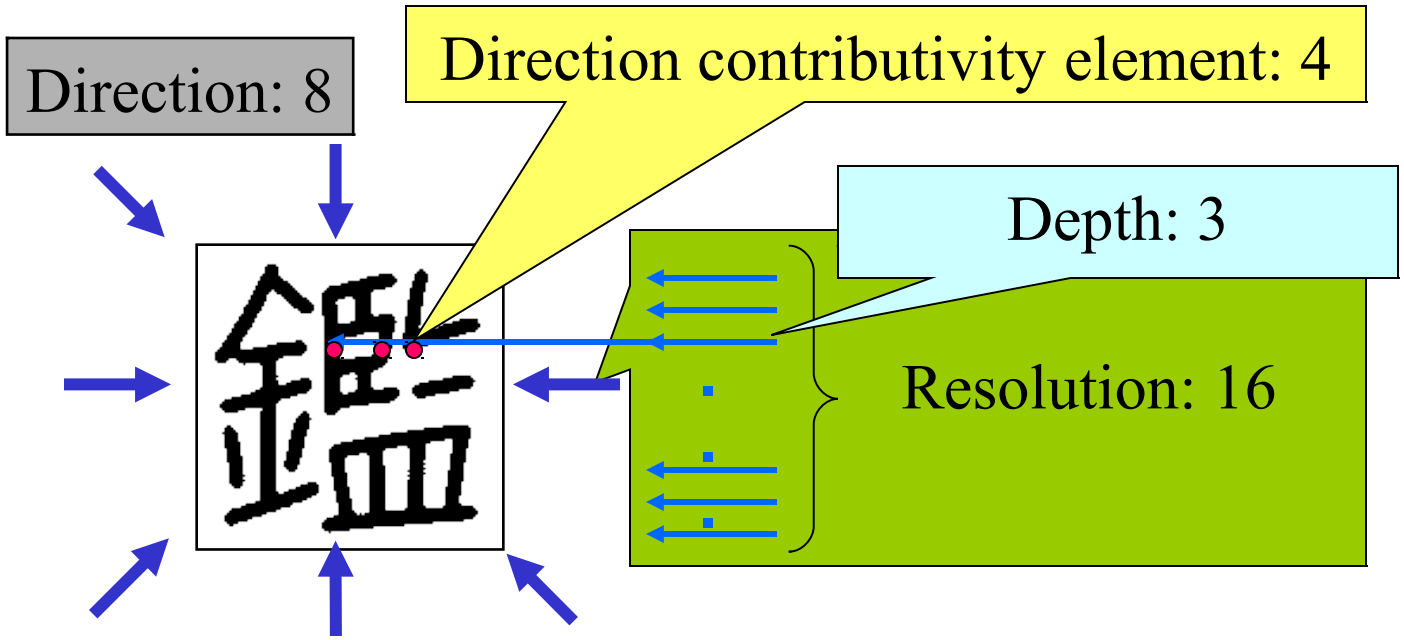
2nd depth

3rd depth

Black dot: Direction contributivity is not 0

PDC Feature

- PDC feature vector: Direction contributivities set

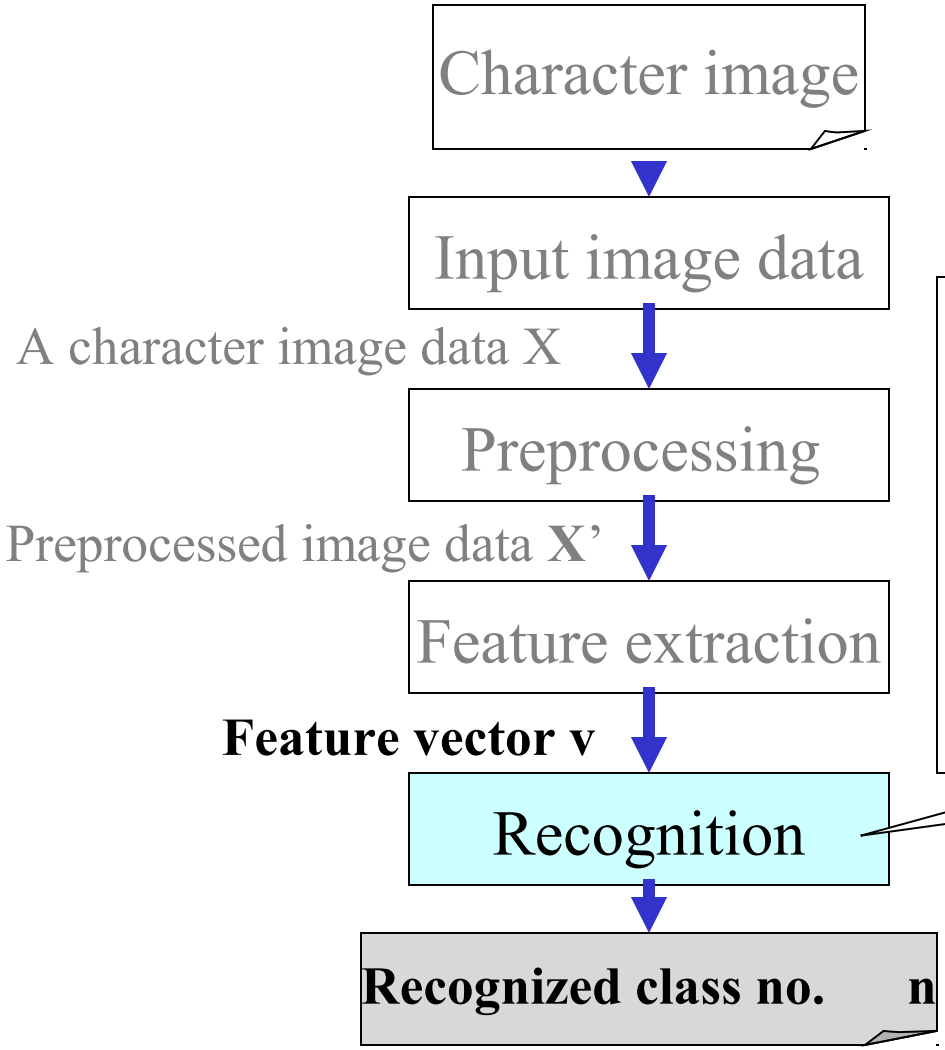


Dimension number=

Direction(8)*Resolution(16)*Depth(3)*Element(4)=1536

Flow of our OCR

Recognition



Recognition by an SVM

Support Vector Machine

–High generalization capability

–Independence of the number of target vector dimension

–Low calculation cost

Experiments

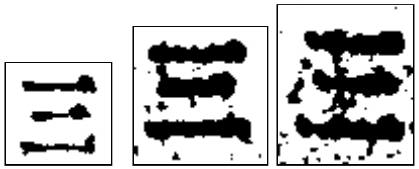
- Experimental sample data
  - Character images obtained from “The Digital Library from the Meiji era”
  - Target characters ：

Class no.	No.1	No.2	No.3	No.4	No.5
Character	行	三	人	生	十
Number of samples	102	103	134	100	100

Class no.	No.6	No.7	No.8	No.9	No.10
Character	來	小	中	年	彼
Number of samples	135	100	209	153	100

Examples of Sample Images

No.1 (行)



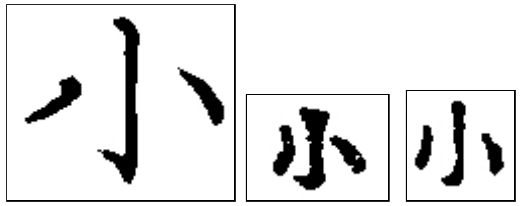
No.2 (三)

No.3 (人)

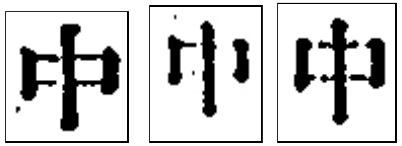
No.4 (生)

No.5 (十)

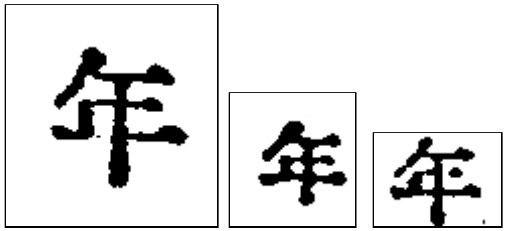
No.6 (來)



No.7 (小)



No.8 (中)



No.9 (年)

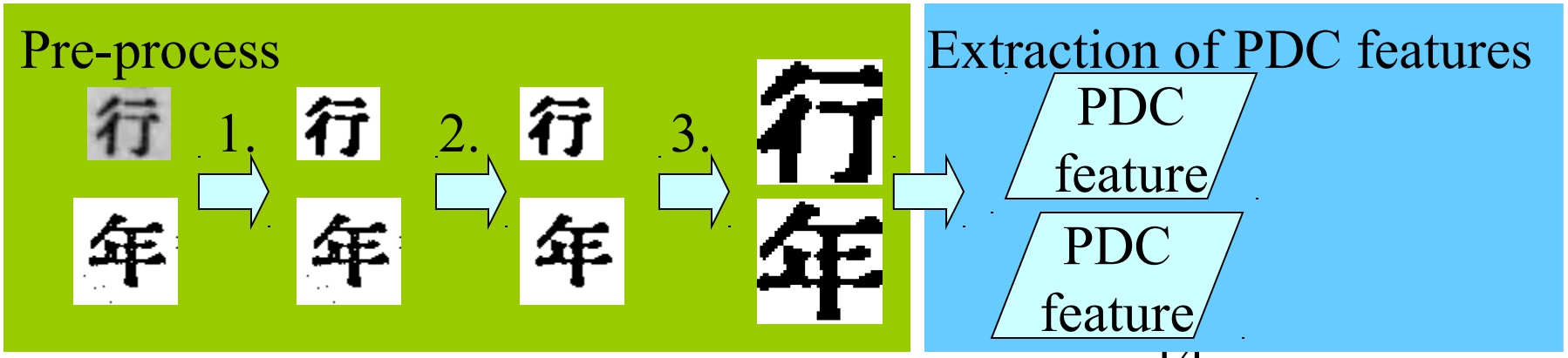
No.10 (彼)

Monochrome or 256-grayscale

Experiments Description(1/2)

Conversion of character images to feature vectors

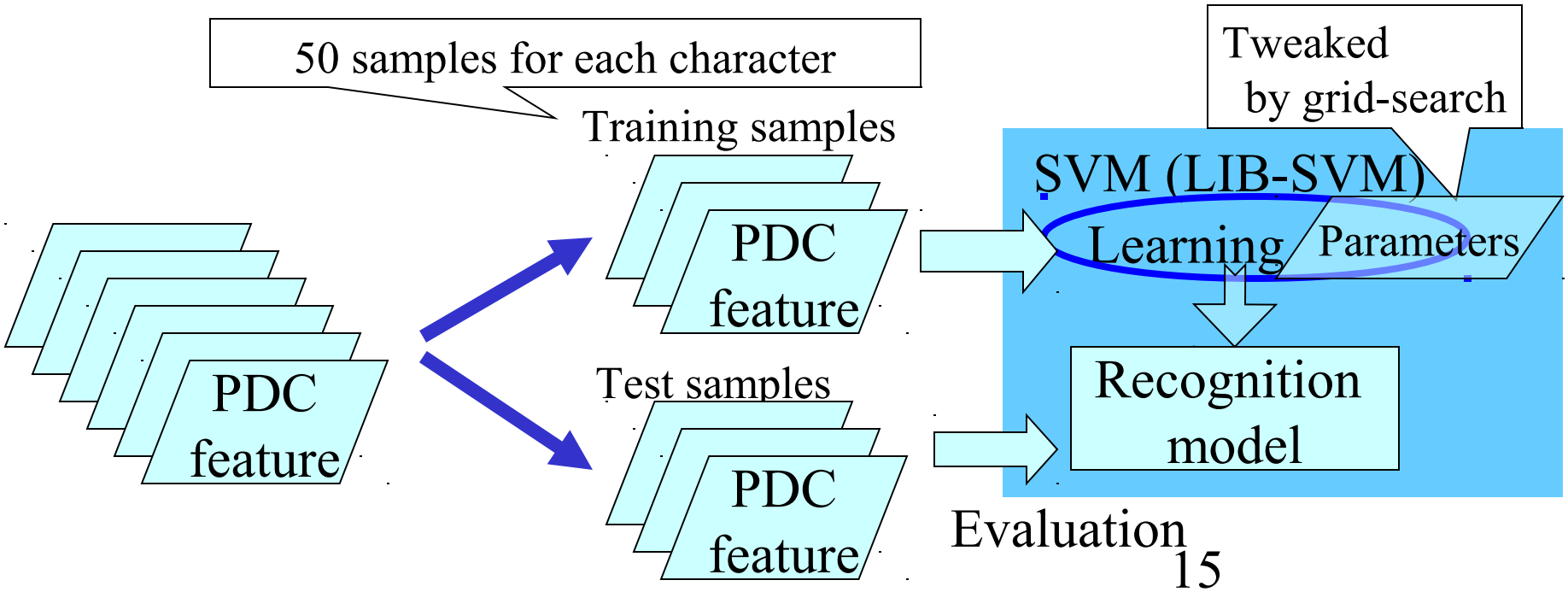
- Pre-process
  1. Binarization Threshold: 128
  2. Noise Reduction Median filter (Filter size ： 3×3)
  3. Normalization Removing margin and scaling to 128×128
- Extraction of PDC features
  - Vector dimension: 1536



Experiments Description(2/2)

Learning and evaluation of a recognition model

- Learning recognition model with training samples to SVM
  - Used SVM: LIB-SVM
  - Parameters of SVM: Tweaked by grid search
- Evaluation of the recognition model by using test samples



Result of Recognition Model Evaluation

※We have shown this result at

73th Mathematical Modeling and Problem Solving (MPS) in March, 2009.

- Recognition rate: 97.8%

Class	Character	The number of test samples	Error	Recognition rate[%]
1	行	52	0	100.0
2	三	53	1	98.1
3	人	84	1	98.8
4	生	50	0	100.0
5	十	50	1	98.0
6	来	85	1	98.8
7	小	50	0	100.0
8	中	159	12	92.5
9	年	103	0	100.0
10	彼	50	0	100.0

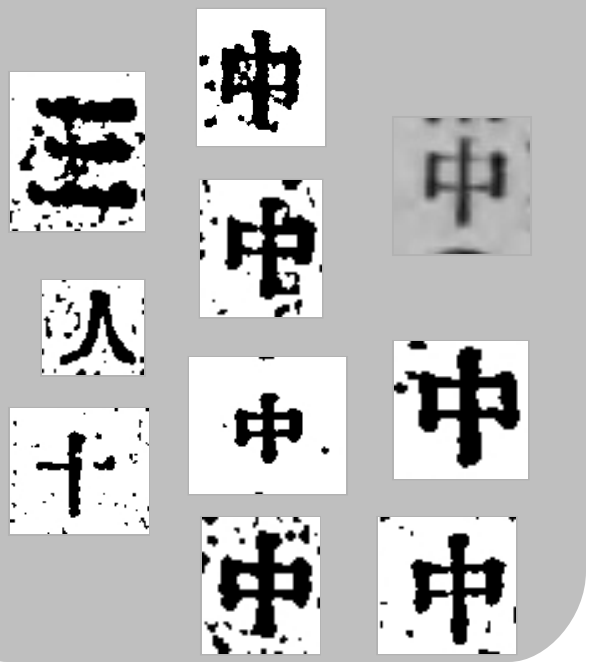
cf. Recognition rate by neural network(NN) ・・ 77.6%

Computation time ・・ SVM: NN= 1 : 7.16

Recognition Error in Result

- Some images are not recognized because of …

noise



or similarity of character forms

Diminishable by an improvement of pre-process

Improvement of Pre-process

- Pre-process
  1. Binarization
     - ~~Threshold:t=128~~ → Discriminant Analysis
  2. First noise reduction
     - Median filter， Filter size ： 3×3
  3. Normalization

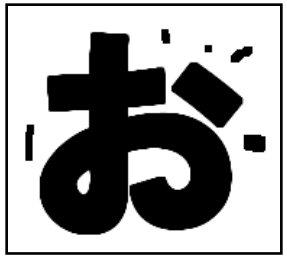
4.Second noise reduction

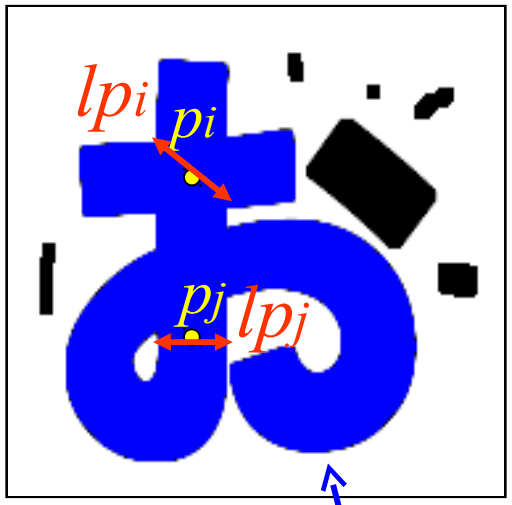
- Based on estimated width of character-line

4.Normalization

Noise Reduction based on Estimation of Character-line Width

Target image





The largest component X

- Estimation of line width by using the largest connected component X

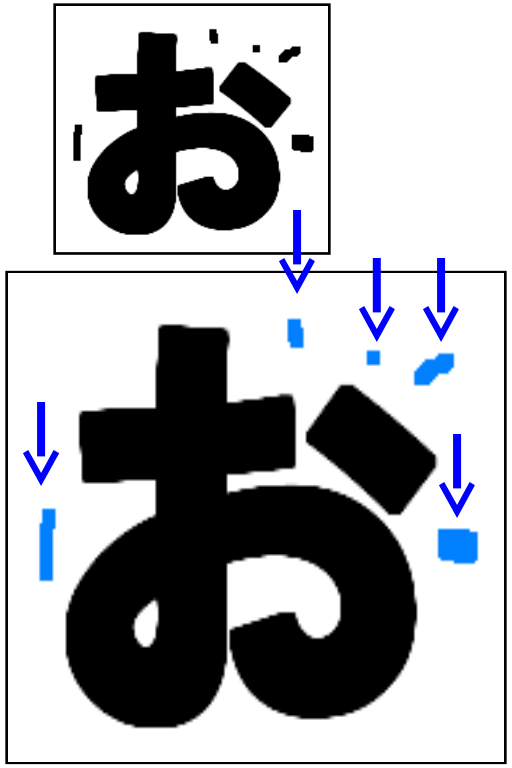
  lp_n : Length of the shortest connected line pass through pixel p_n $(p_n \subset X)$

Estimated width of character-line:
b=median value of lp_n

- Elimination of connected component whose area is smaller than $\frac{b^2}{2}$

Noise Reduction based on Estimation of Character-line Width

Target image



- Estimation of line width by using the largest connected component X

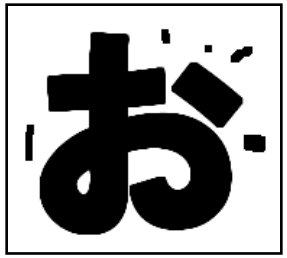
  lp_n : Length of the shortest connected line pass through pixel p_n $(p_n \subset X)$

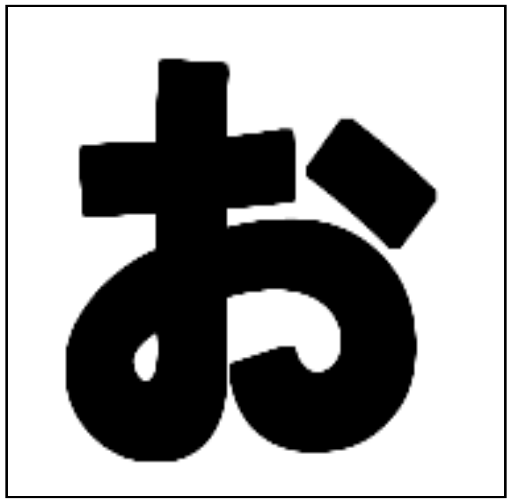
Estimated width of character-line:
b=median value of lp_n

- Elimination of connected components whose area are smaller than $\frac{b^2}{2}$

Noise Reduction based on Estimation of Character-line Width

Target image





- Estimation of line width by using the largest connected component X

  lp_n : Length of the shortest connected line pass through pixel p_n $(p_n \subset X)$

Estimated width of character-line:
b=median value of lp_n

- Elimination of connected components whose area are smaller than $\frac{b^2}{2}$

Result of Improved Pre-process Adoption

- Recognition rate 97.8%→99.0%

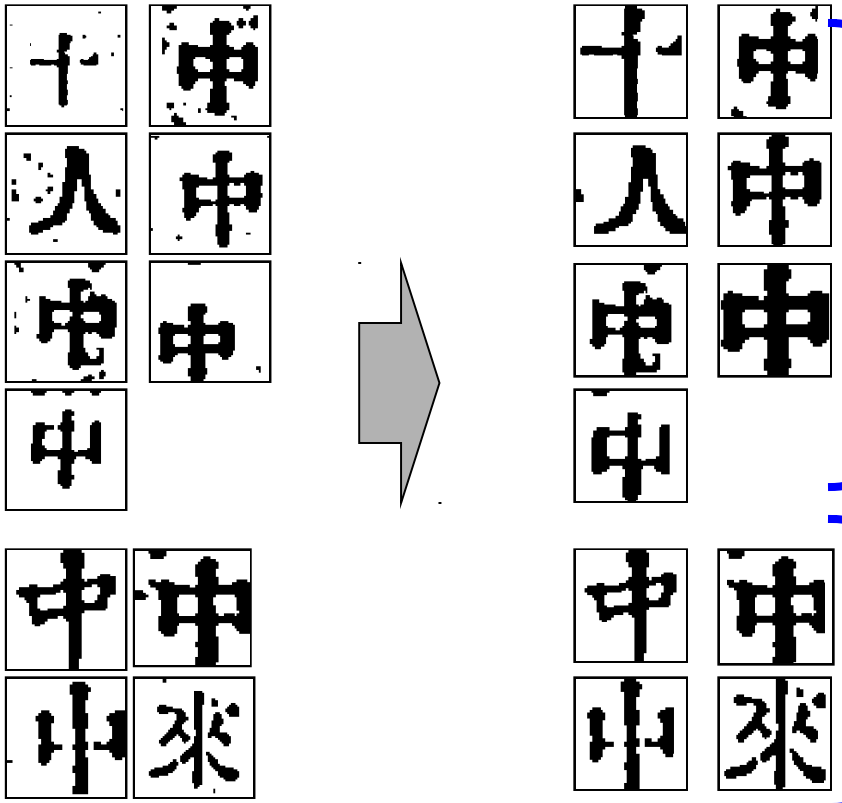
Class	Character	The number of unknown input data	Previous result	New noise reduction	
			Recognition rate[%]		Error
1	行	52	100.0%	100.0%	0
2	三	53	98.1%	98.1%	1
3	人	84	98.8% ↑	100.0%	0
4	生	50	100.0%	100.0%	0
5	十	50	98.0% ↑	100.0%	0
6	来	85	98.8% ↑	100.0%	0
7	小	50	100.0%	100.0%	0
8	中	159	92.5% ↑	96.9%	5
9	年	103	100.0% ↓	99.0%	1
10	彼	50	100.0%	100.0%	0

Discussion

Case: better recognition(Error→Correct)

Previous pre-process　Improved pre-process

Error　　*Correct*



Quality of
test samples are improved

Quality of
training samples are improved

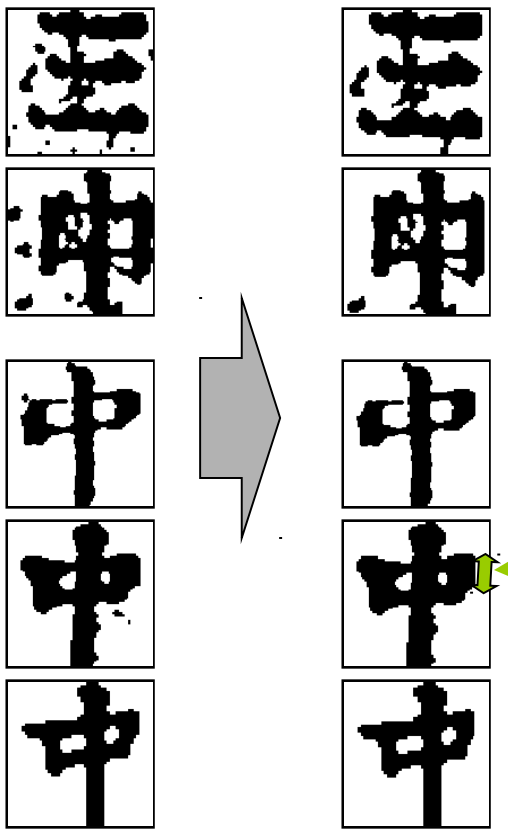
More efficient
recognition model

Discussion

Case: unchanged(Error→Error)

Previous *Error*

Improved *Error*

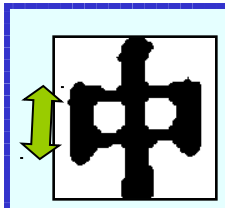


Connected to character-line

Residual noise
⇨ Error

Similar form of character no.5(＋)
⇨ Error

Shorter than major form
→Similar with one horizontal line



Major form of no.8

Discussion

Case: worse recognition (Correct→Error)

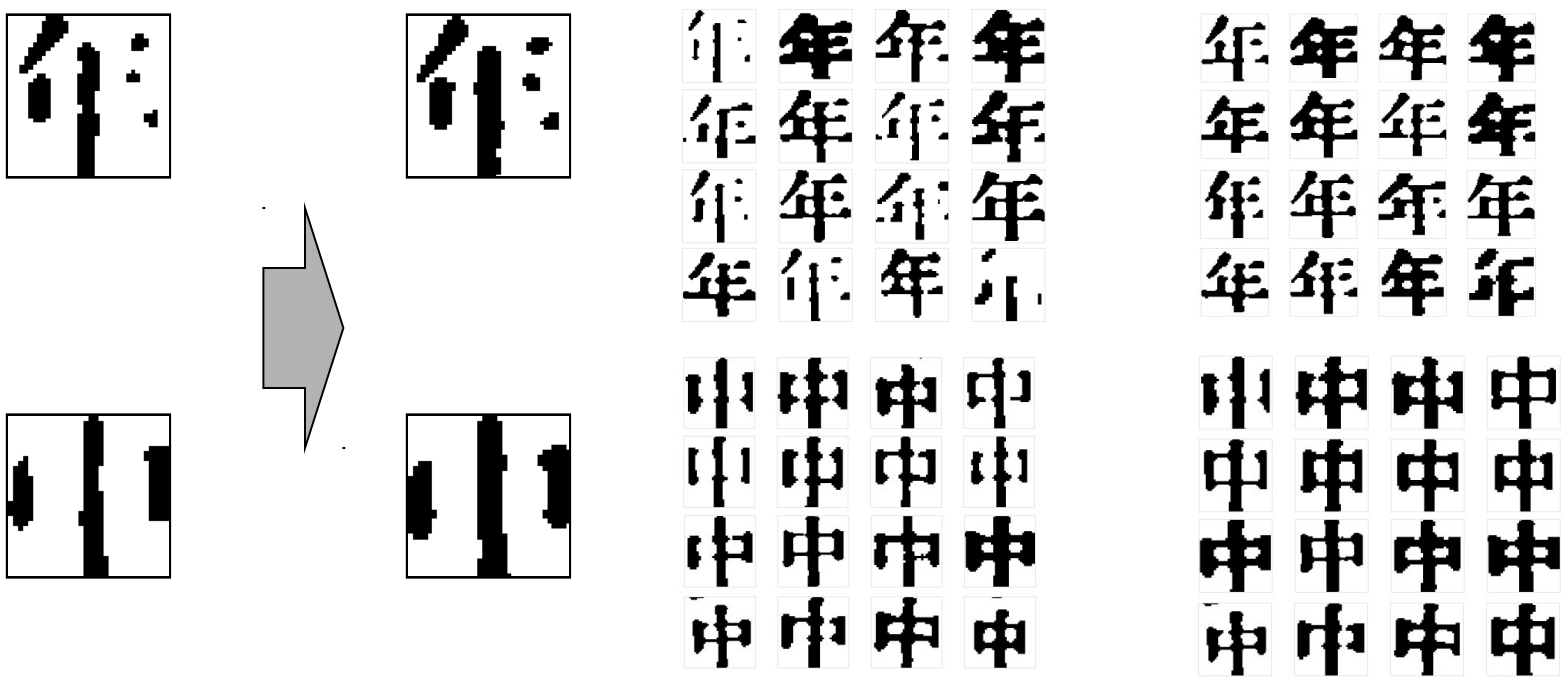
Previous *Correct*

Improved *Error*

Pre-processed images

Previous

Improved



Training samples with lack of line are reduced

Recognition rate of data with lack of line becomes low

Conclusions and Future work

- Recognition of multi-fonts character in Early-Modern Printed Books
  - Proposal of our method which uses PDC feature and SVM
  - Experimentations of applying our method
    - The results show high recognition rate
    - Improvement of noise reduction leads higher recognition rate
      - Recognized 10 kinds of character at 99 % accuracy
- Future works
  - Dealing lots of character kinds ⇒ Hierarchical recognition method
    - Recognition of similar form characters
  - Automation of extracting character area

Thank you for your attention!