Detection and Segmentation of Touching Characters in Mathematical Expressions

A. Nomura,
K. Michishita,
S. Uchida,
M. Suzuki
Kyushu University, Japan

Introduction

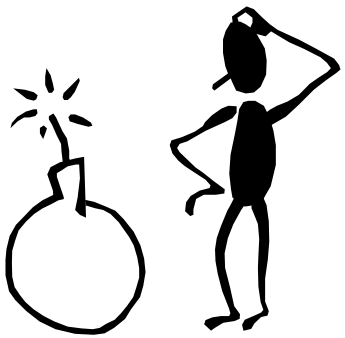
OCR for mathematical document

- Aim
  - Recognition of ordinary texts and mathematical expressions
- Merits
  - Storage size reduction
    bitmap image → ASCII codes
  - Search services
    theorem search, definition search, ...
  - Format conversion
    from scanned image
    to LaTeX, XML, Mathematica Notebook, Braille, ...

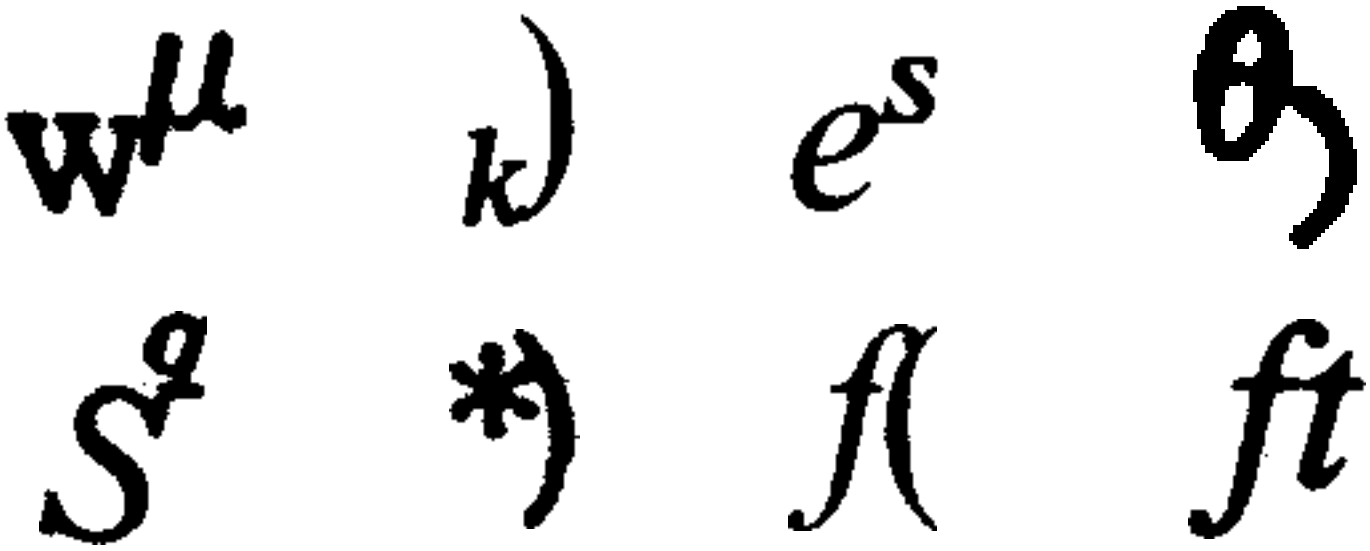
Hurdles

- Large categories (>500)
  - alphabets, numerals, Greek, operators, parentheses, big symbols (e.g., "Σ"), ...
- Various fonts
  - roman, italic, calligraphic, ...
- Various sizes and positions
  - sub-/super-scripts, fractions, ...

- Touching characters
  - *50% and more misrecognitions were due to touching characters*



Touching characters in math. expressions



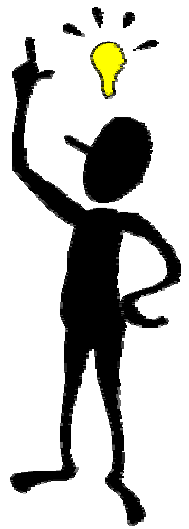
touching not only horizontally

but also diagonally

conventional segmentation techniques will fail

Our purpose

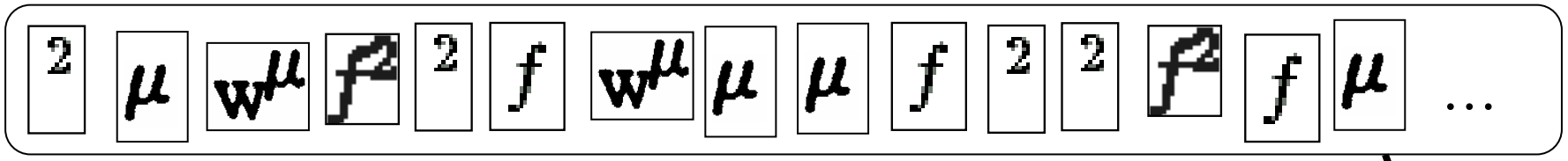
- Development of a novel segmentation technique for touching characters in mathematical expressions
- And higher recognition accuracy



Outline of the proposed technique

Outline of the proposed technique (1)

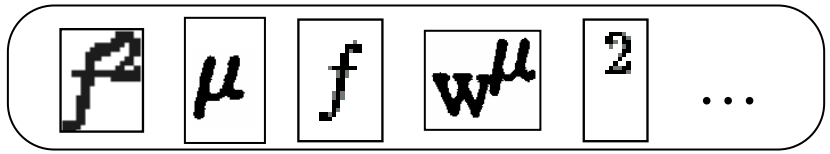
all connected components in document
(image data with initial recognition result)



clustering based on shape similarity

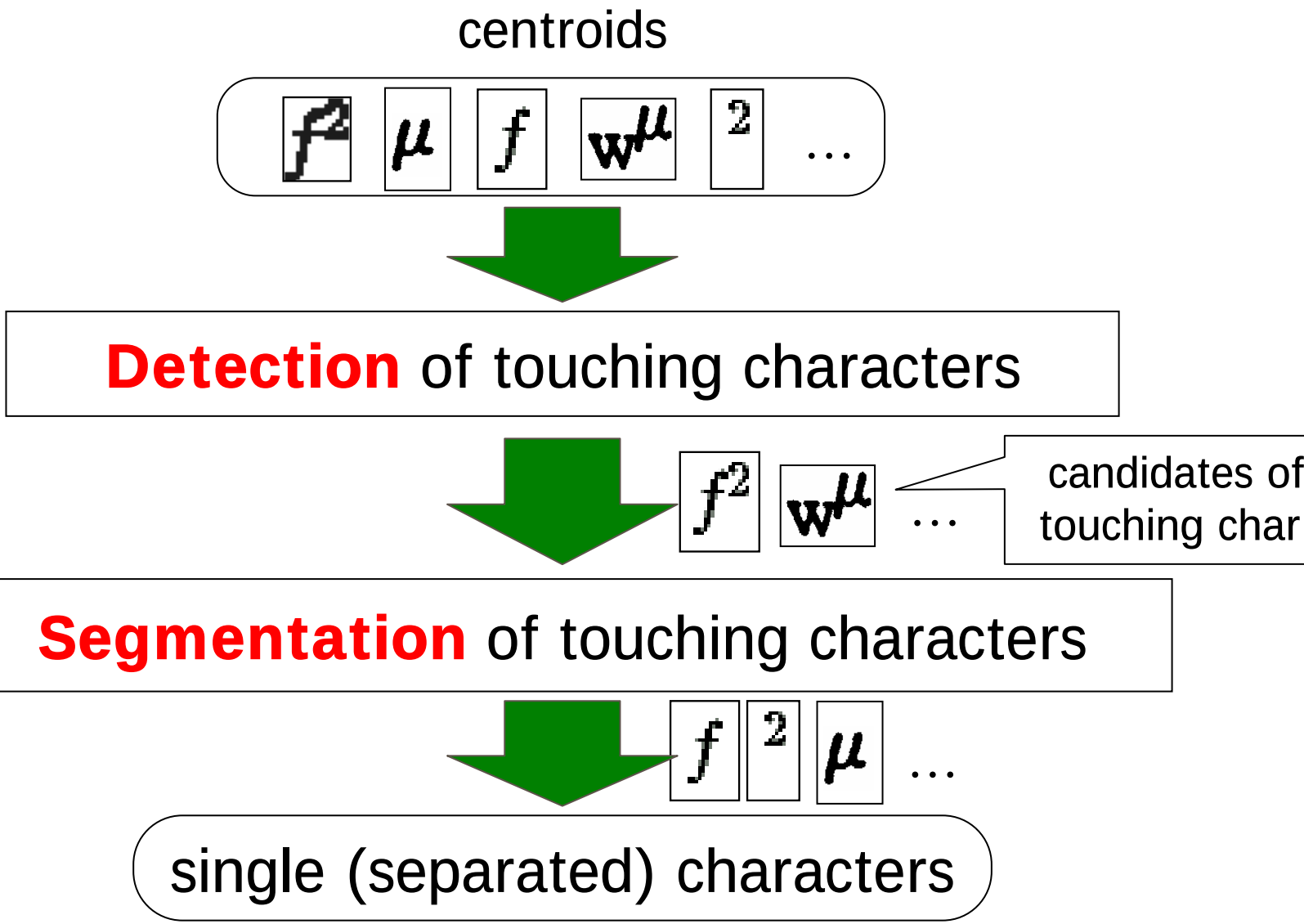
- computational efficiency
- neglect of trivial shape difference

centroids

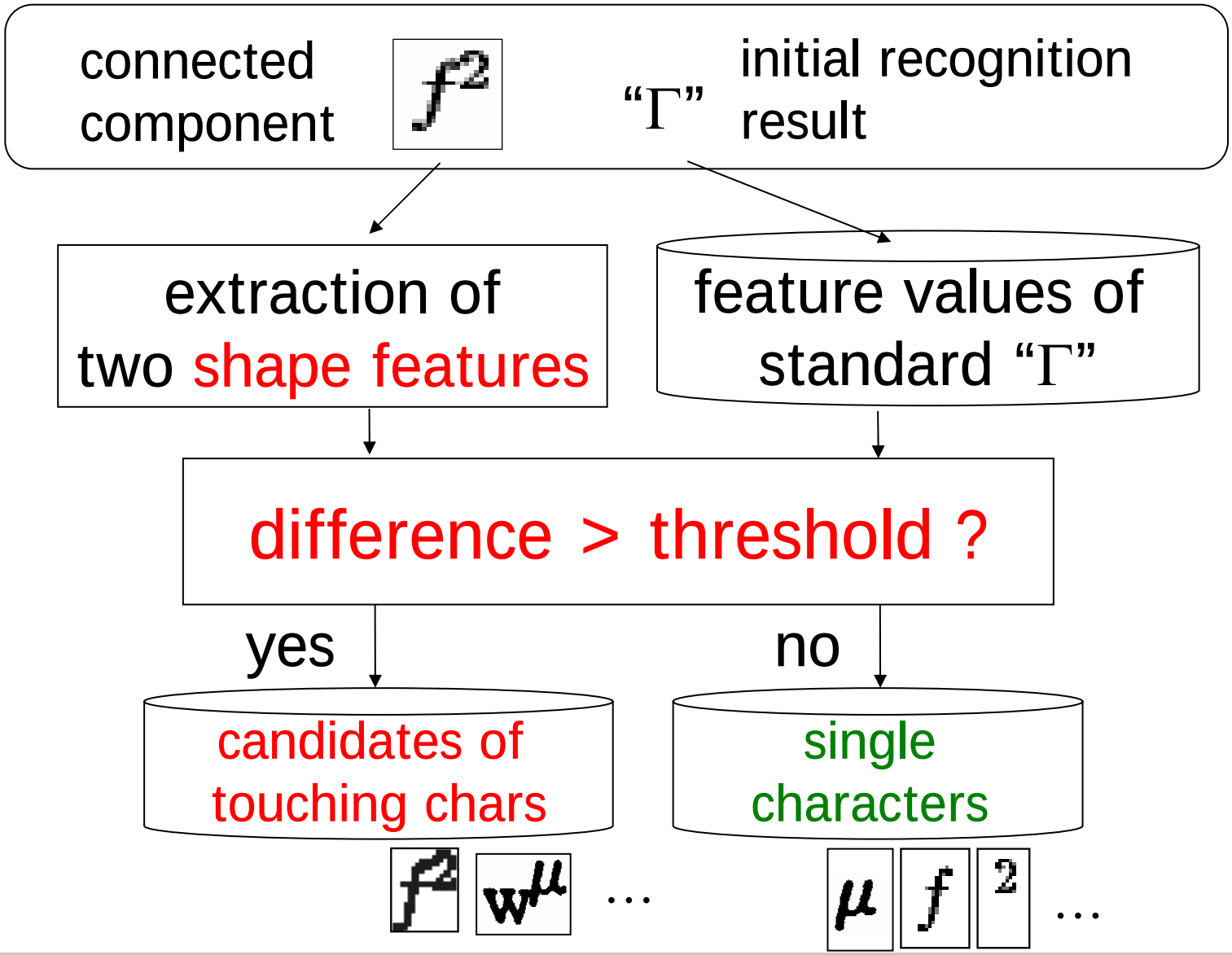


about 1/10
compressio

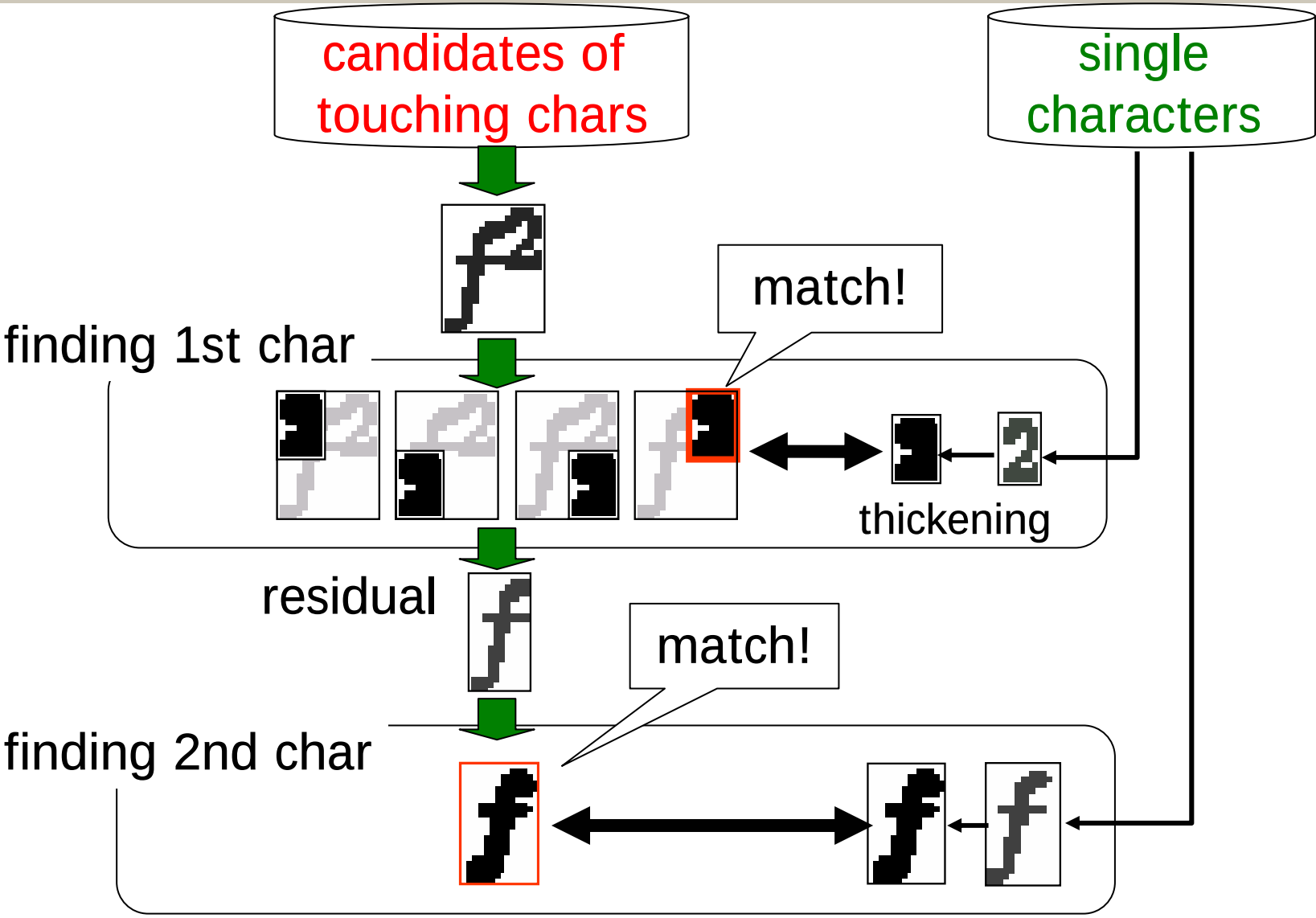
Outline of the proposed technique (2)



Detail of detection procedure



Detail of segmentation procedure



Notes

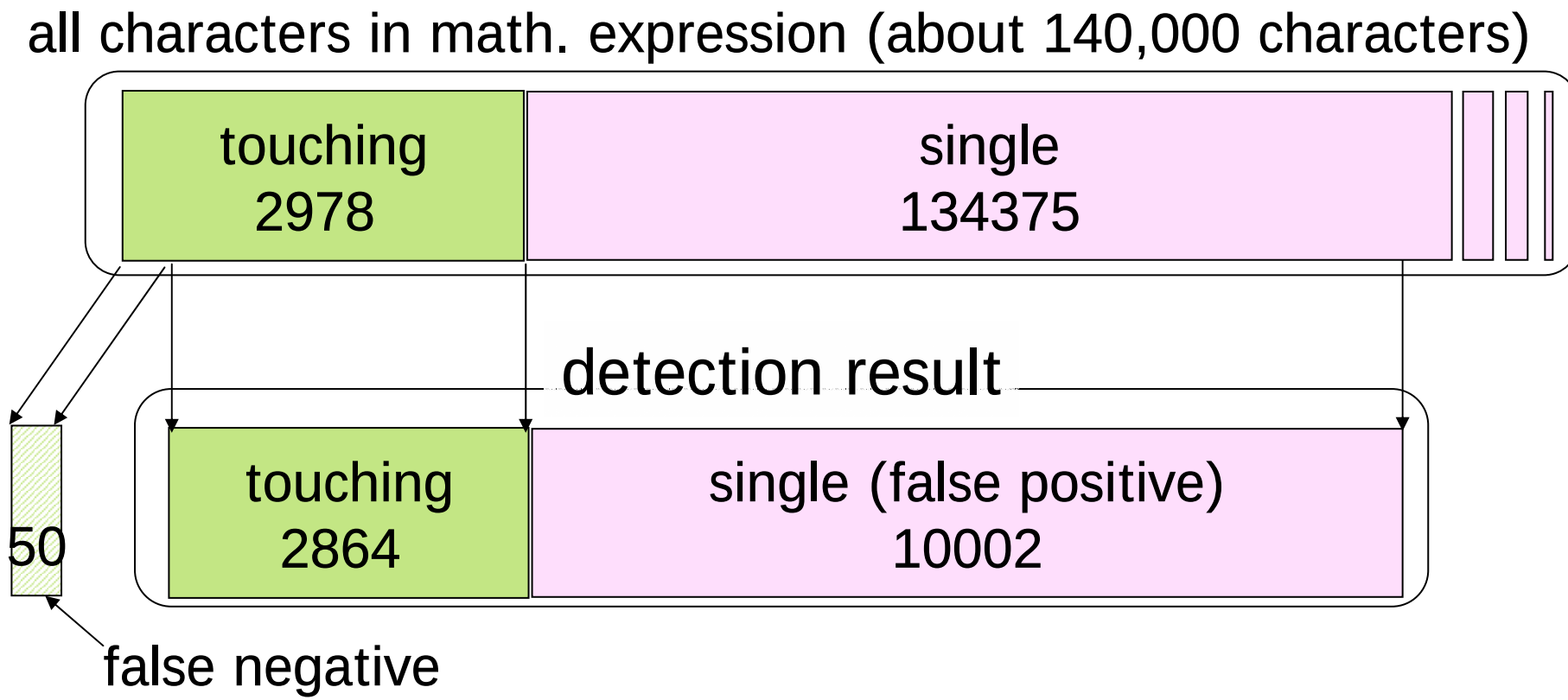
1. Single characters from the same document are utilized
   - Components of touching characters are usually found as single characters in the document
   - Font-style/size adaptive separation is realized
2. Recognition result is provided
3. Tolerant to false positives
   - If no match is found, the candidate is rejected as a single character

Experimental results

Database

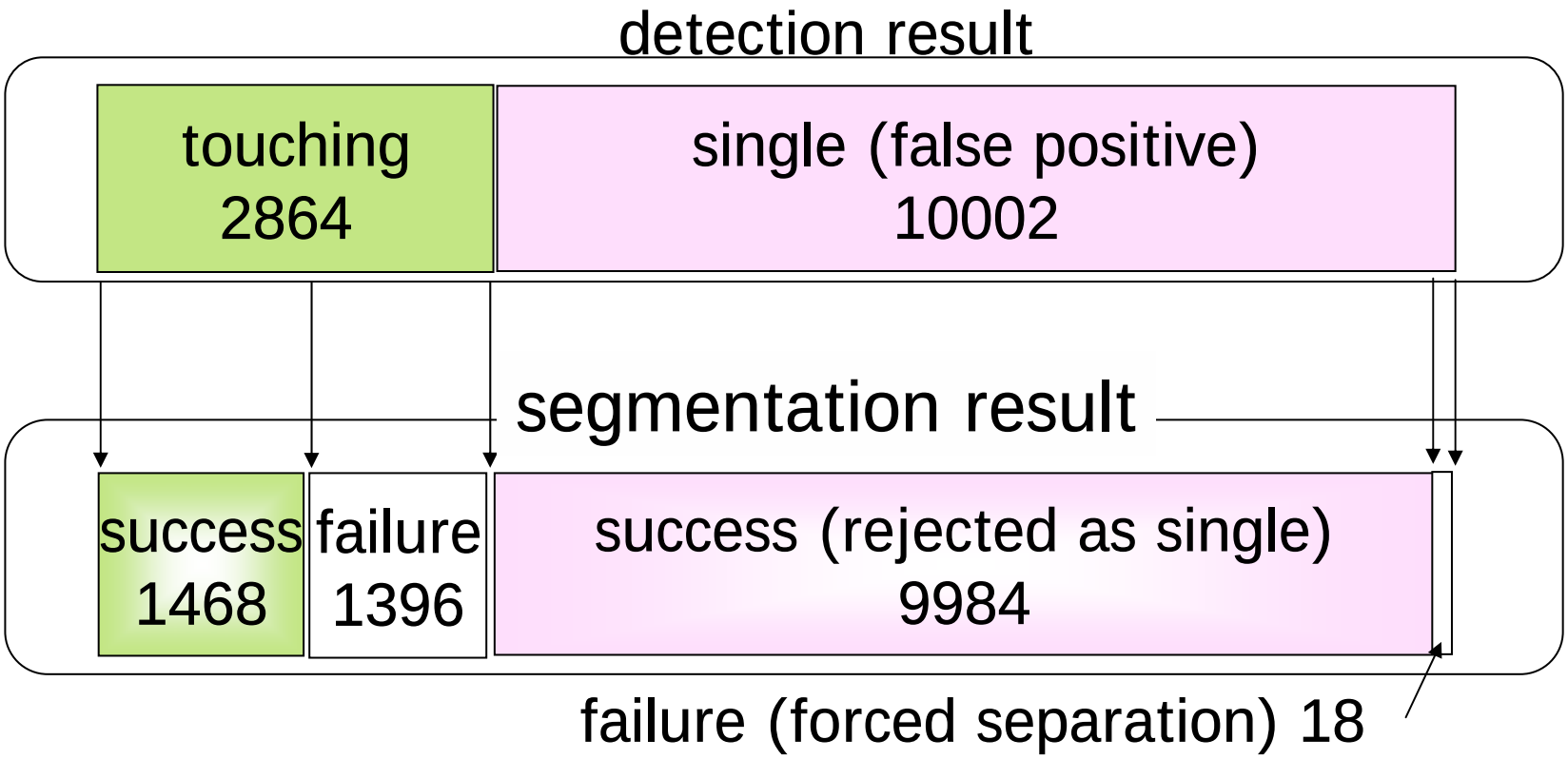
- 391 pages from 21 math. documents
- 140,000 characters in math. expressions
  - groundtruth was manually attached to each character
  - characters in ordinary text parts were excluded in our experiment
- 2,978 touching char images (~ 6,000 chars)
  - 4.2% of all 140,000 characters

Detection result



- False negatives were small (1.6%)
- False positives were large (but will be rejected in the segmentation procedure)

Segmentation result



- 50% of touching chars were successfully separated
- Forced separations were very small

Segmentation result

$$w^{\mu} \rightarrow w + \mu$$

$$e^{s} \rightarrow e + s$$

$$k) \rightarrow k +)$$

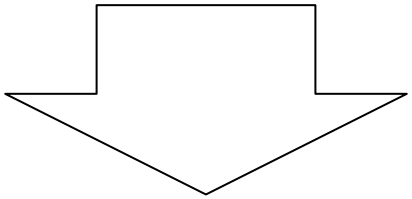
Effect on total recognition rate

- Recognition rate of all 140,000 characters

  92.9 % → 95.1 % (2.2% up)

- The number of misrecognitions

  9710 → 6792 (30% reduction)



the proposed technique is very meaningful !

Conclusion

- Detection and segmentation procedures for touching characters in math. expressions were investigated
- About 50% of touching characters were successfully detected and separated
- Total character recognition rate was improved from 92.9 % to 95.1 %.