FUJITSU

shaping tomorrow with ou

画像データからのテキスト抽出技術について

2011年8月3日

富士通株式会社

目次

- はじめに
- 画像からのテキスト抽出の方法と課題
- PDFからのテキスト抽出の方法と課題
- ガイドライン (案)
- まとめ

FUJITSU

shaping tomorrow with you

はじめに

課題と目的

FUJITSU

■課題

　読書障碍者にとって、音声や点字を用いて読むことのできる電子出版物、特に雑誌やコミックが少ない。

　一方、画像データ（＝出版版下データ）では、その編集の過程で見出し画像等からテキスト情報が消失しているケースがあり、読み上げや検索に支障がある。

■目的

　ニーズの高い雑誌データを主なターゲットとして、テキスト抽出の実証実験を行う。

実証実験について

FUJITSU

- 実験の詳細

  画像を対象にした場合と、PDFを対象にした場合のテキスト抽出精度を比較する。

- PDFの種別

  ・Adobe InDesignで生成されたPDF
  ・QuarkXpressで生成されたPDF

- 実証実験に用いた評価データ

  ① 「山と渓谷」、44ページ、Adobe InDesign
  (2009年11月号、株式会社 山と渓谷社)

  ② 「Hanako」、11ページ、 Adobe InDesign
  (2009年11月号、株式会社 マガジンハウス)

  ③ 書籍データ、2ページ、 QuarkXpress

FUJITSU

shaping tomorrow with you

画像からのテキスト抽出の方法と課題

画像からのテキスト抽出の方法と精度

FUJITSU

方法

- PDF文書 → 各ページを200dpiの画像に変換 → OCR(Optical Character Recognition)技術を適用

評価尺度

- 再現率：文字認識により正しく抽出できた文字の割合
- 適合率：文字認識により抽出された文字のうち、正しい文字（＝ゴミ文字以外）の割合

精度

	テキスト抽出精度
再現率	約85%
適合率	約93%

画像からのテキスト抽出の課題(1)

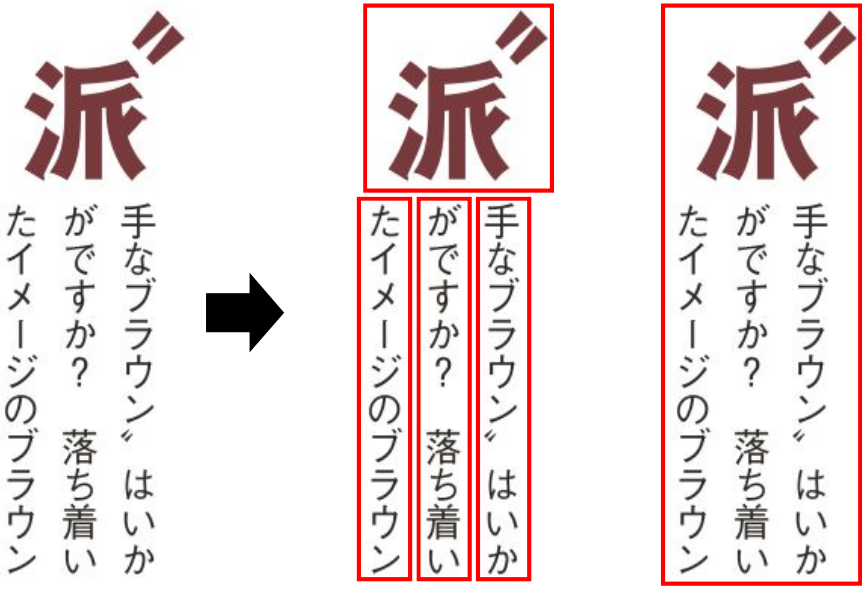
FUJITSU

■文字認識の失敗

認識結果

『山鞠襄谷』

■レイアウト解析の失敗



認識対象

期待する行単位の検出結果

誤った行単位の検出結果

認識結果例

『派蕊謡』

画像からのテキスト抽出の課題(2)

FUJITSU

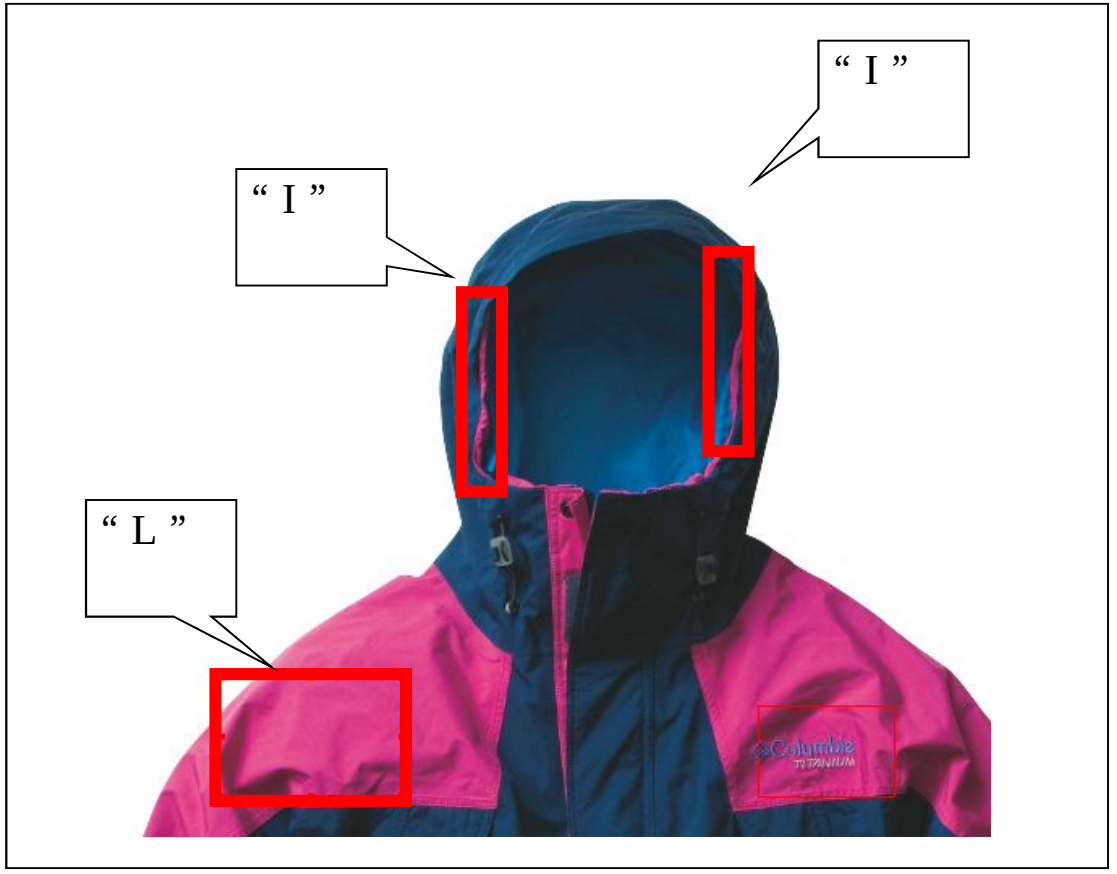
■ レイアウト解析の失敗：斜めに描画された文字

■ 領域抽出の失敗

画像からのテキスト抽出の課題(3)

FUJITSU

■文字のない領域からの誤検出



FUJITSU

shaping tomorrow with you

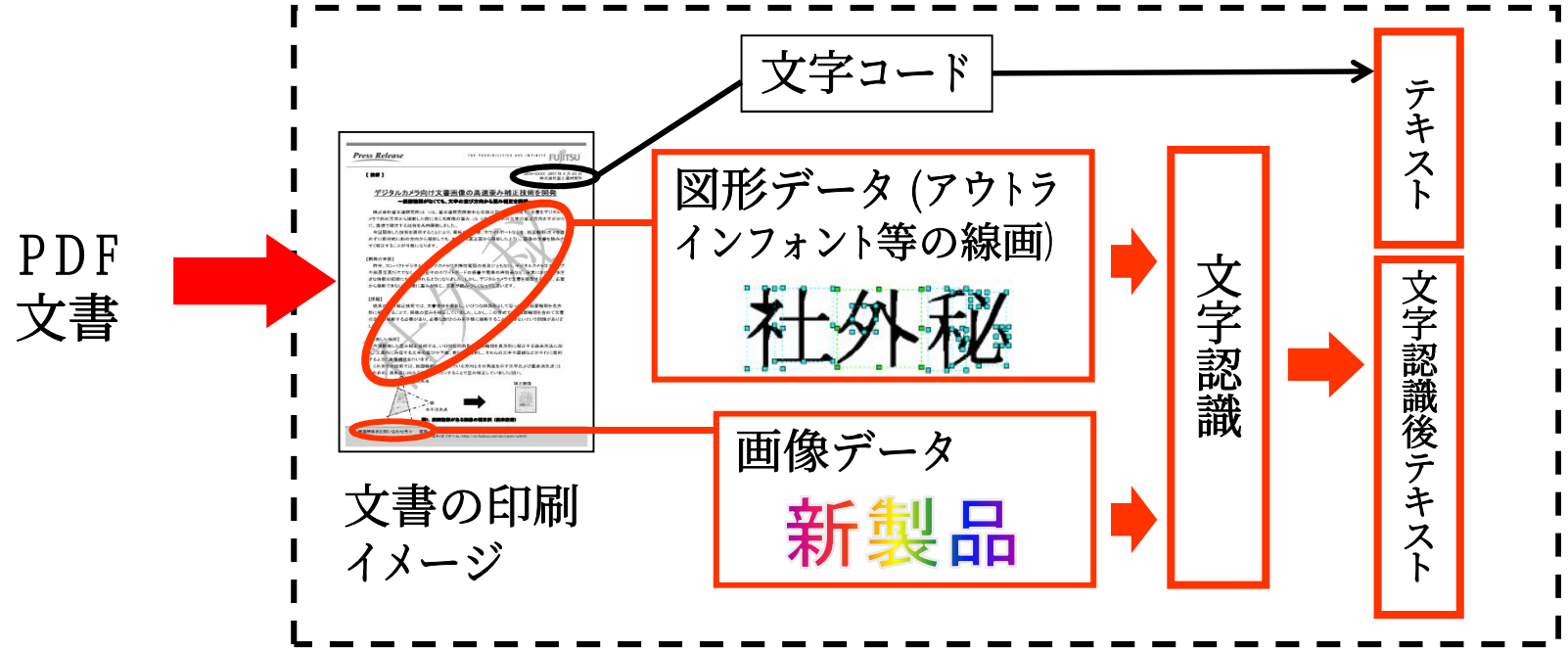
PDFからのテキスト抽出の方法と課題

PDFからのテキスト抽出の方法と精度

FUJITSU

■方法：PDF文書に対してPDF解析を行う

■PDF解析



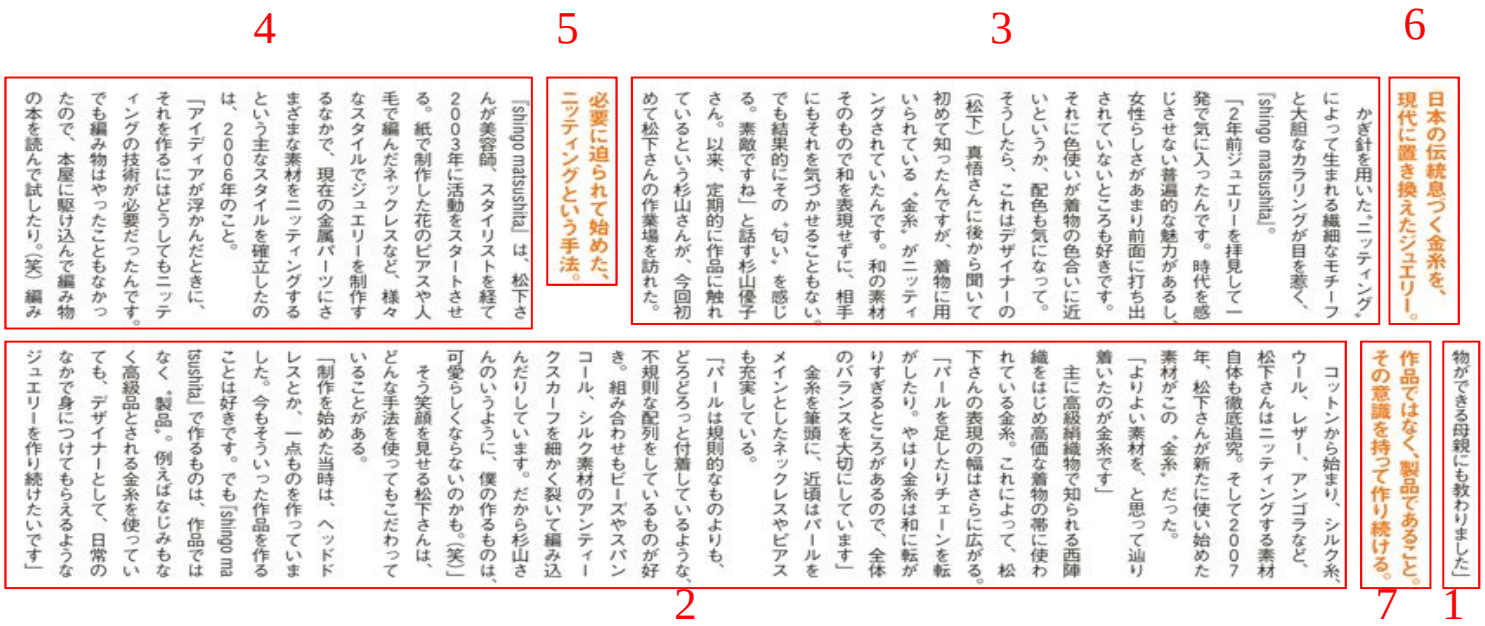
■精度

	テキスト抽出精度
再現率	約97%
適合率	約97%

PDFからのテキスト抽出の課題(1)

FUJITSU

■読み順と異なる描画順



1 茨城県 生瀬富士

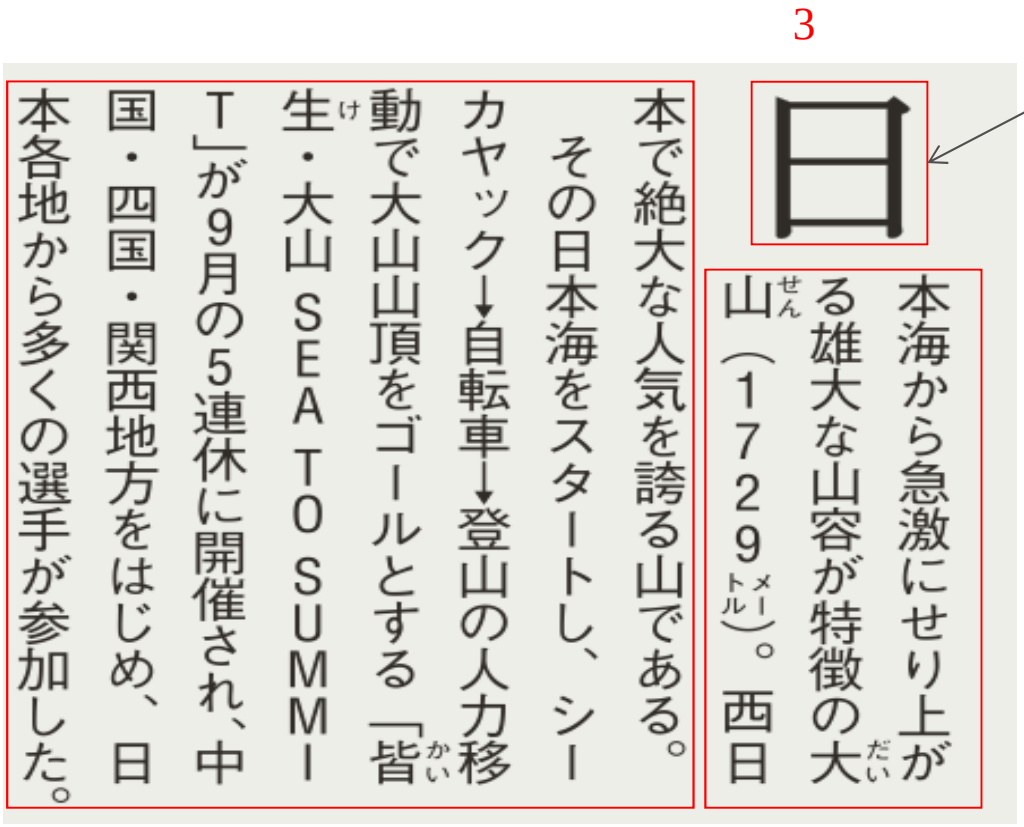
2 宮城県と山形県 笹谷峠

3 神奈川県 姫次と袖平山

PDFからのテキスト抽出の課題(2)

FUJITSU

■ 読み順と異なる描画順：文字コード／図形の混在



PDFからのテキスト抽出の課題(3)

FUJITSU

■ 文字コード／図形／画像の文字の混在

図形

図形

TREND STUDY

Hanako's View | TOKYOの流行りモノ研究

図形

文字コードによる描画

PDFからのテキスト抽出の課題(4)

FUJITSU

■ ルビ

PDF内の文字コード並び：

『「栂つが海み新道に』

■ 2重テキスト

PDF内の文字コード並び：

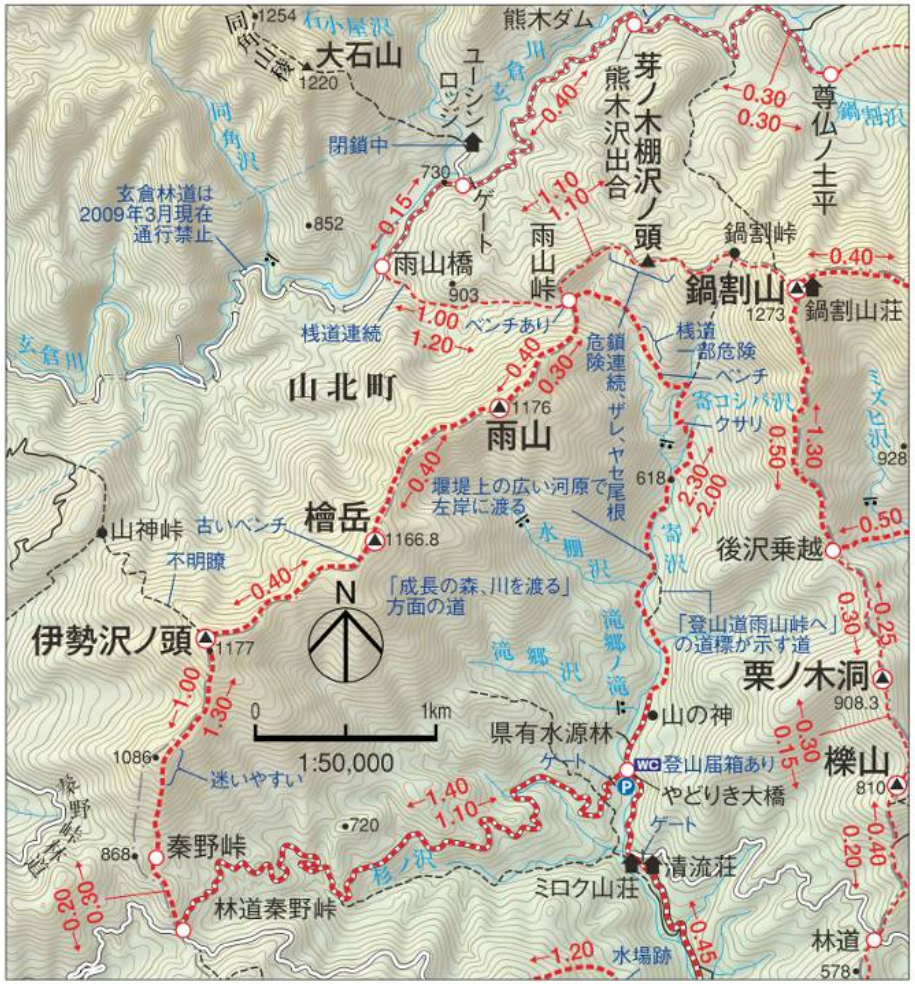
『追憶の一葉、追憶の一葉
述懐の山述懐の山』

PDFからのテキスト抽出の課題(5)

FUJITSU

■地図上に書かれた文字

どこまでの文字を抽出し、文字並びをどう定義するか？

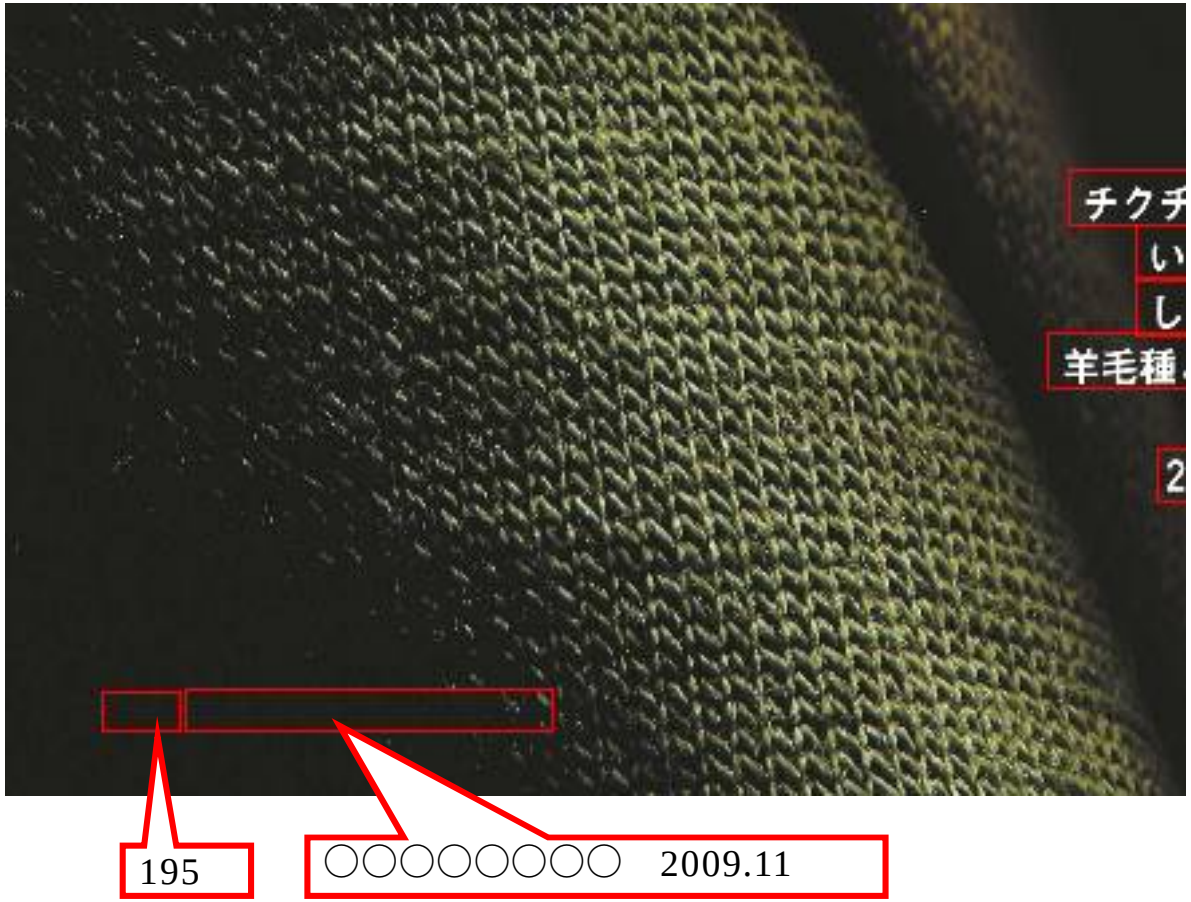


PDFからのテキスト抽出の課題(6)

FUJITSU

■非表示テキストの存在

見かけ上は見えないが、PDF内には文字コードが記録されている。



FUJITSU

shaping tomorrow with you

ガイドライン（案）

テキスト抽出における要件定義

FUJITSU

■画像のみから成るPDFに対して

・図に含まれる文字を読み上げ・検索対象にするかどうか
・写真に含まれる文字を読み上げ・検索対象にするかどうか
・OCRの適用に際して、再現率を重視するか適合率を重視するか
・文字のサイズ、色、アンダーラインなどの属性に対する
読み上げルールの設定

■文字コード／図形／画像が混在するPDFに対して

・目視できない文字（ex. 図の下に隠れた文字、PDFに属性としてのみ付与された文字）を読み上げ・検索対象にするかどうか
・ルビがある場合に、漢字と仮名の読み上げ方をどうするか
・図に含まれる文字を読み上げ・検索対象にするか、
あるいは代替テキストを読むかどうか
・文字のサイズ、色、アンダーラインなどの属性に対する読み上げルール
・タイトル、見出し、小見出しに対する読み上げルール
・（2重テキストとなっている場合に）見た目の文字数を優先するか、テキストを2回読むかどうか
・外字に対する読み上げルールの設定

テキスト抽出に適した画像データ

FUJITSU

■雑誌のレイアウトに対し、影響がない特性

①可能な限り、テキストデータを残す。
②雑誌上見えない文字については、削除する。
③「タイトル」「見出し」「ページ」「図」「表」「テキストブロック」「箇条書き」などの属性を表すタグが付与されている。
④タグ全体について、その読み順（Reading Order）が付与されている。

■雑誌のレイアウトに対し、影響がある特性

①1行の中での文字サイズは同じ。
②文字の並び（行 or 列）の方向は水平もしくは垂直。
③文字の背景は単色で、かつ文字色と背景色の濃度差が大きい。
④文字の並びの中に図形が混在しない。
⑤特殊な文字（矢印記号、「①」など）や外字が使用されていない。
⑥文字フォントはゴシック、明朝などの一般的に使用頻度の高いフォント。
⑦文書のレイアウトが単純（＝図、写真、テキストブロックが入り組んでいない）。
⑧文字同士が隣接している（＝文字列として抽出しやすい）。
⑨行頭に倍角文字が使われていない。

画像データからのテキスト抽出に必要な機能

FUJITSU

所望の特性を持っていない画像データ（ex. 過去のデータ等）については、以下の機能を持つツールが必要とされる。

①見た目のレイアウトに基づき、1行の文字並びを正しく判定する。

②見た目のレイアウトに基づき、段落の並び順を正しく判定する。

③図で表記された文字、文字コードが混在する場合に、見た目の読み順どおりにそれらを正しく判定する。

④見た目のレイアウトに基づき、2重テキストを回避する。

⑤「タイトル」「見出し」「ページ」「図」「表」「テキストブロック」「箇条書き」などの属性を表すタグを付与する。

FUJITSU

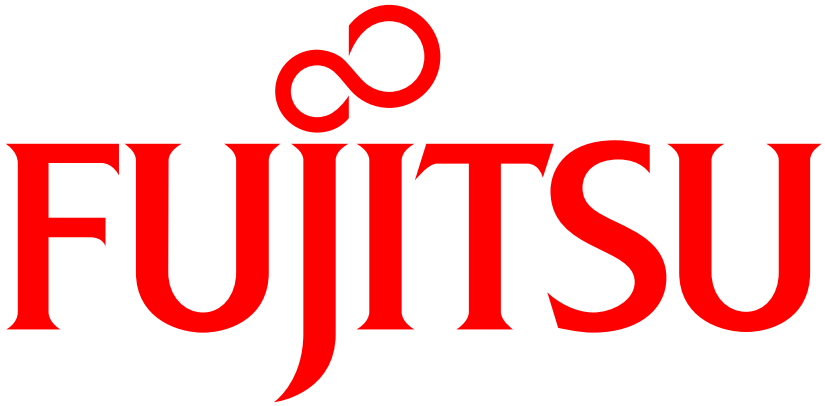
shaping tomorrow with you

まとめ

FUJITSU

- 画像データ（特に雑誌データ）に対してテキスト抽出の実証実験を実施。
  - 画像、PDFデータのそれぞれについてテキスト抽出における課題を抽出

- 実証実験を踏まえたガイドライン（案）として、要件定義とテキスト抽出に必要な機能の検討を実施。



shaping tomorrow with you