

# 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

- 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

読書障害者にとって、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

「山と溪谷」、44ページ、 Adobe InDesign  
(2009年11月号、株式会社 山と溪谷社)

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

書籍データ、2ページ、 QuarkXpress



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

## ■ 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)

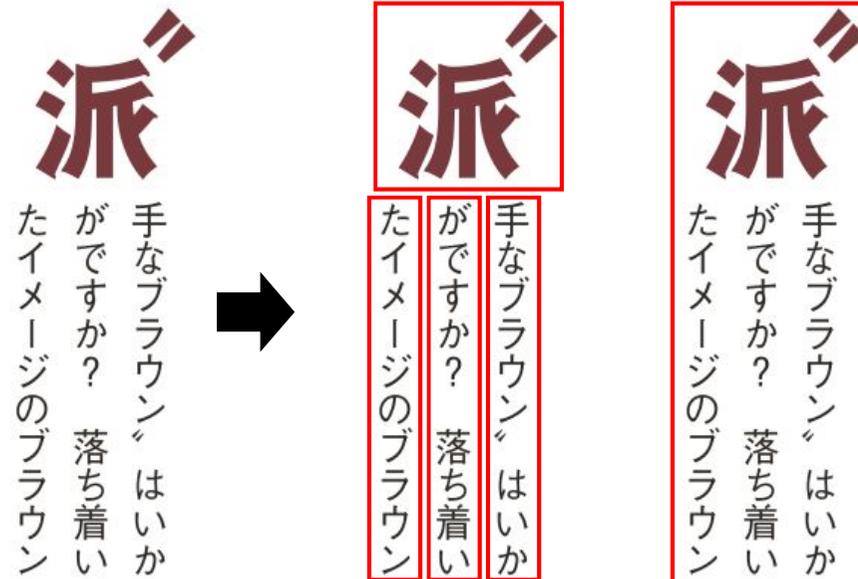
## ■ 文字認識の失敗



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

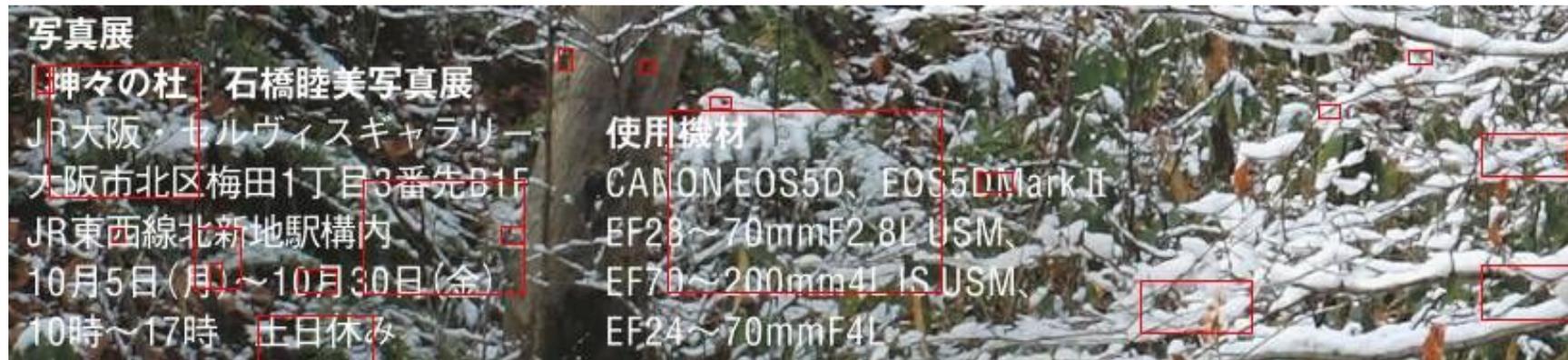
『派蕊謡』

# 画像からのテキスト抽出の課題(2)

## ■ レイアウト解析の失敗: 斜めに描画された文字

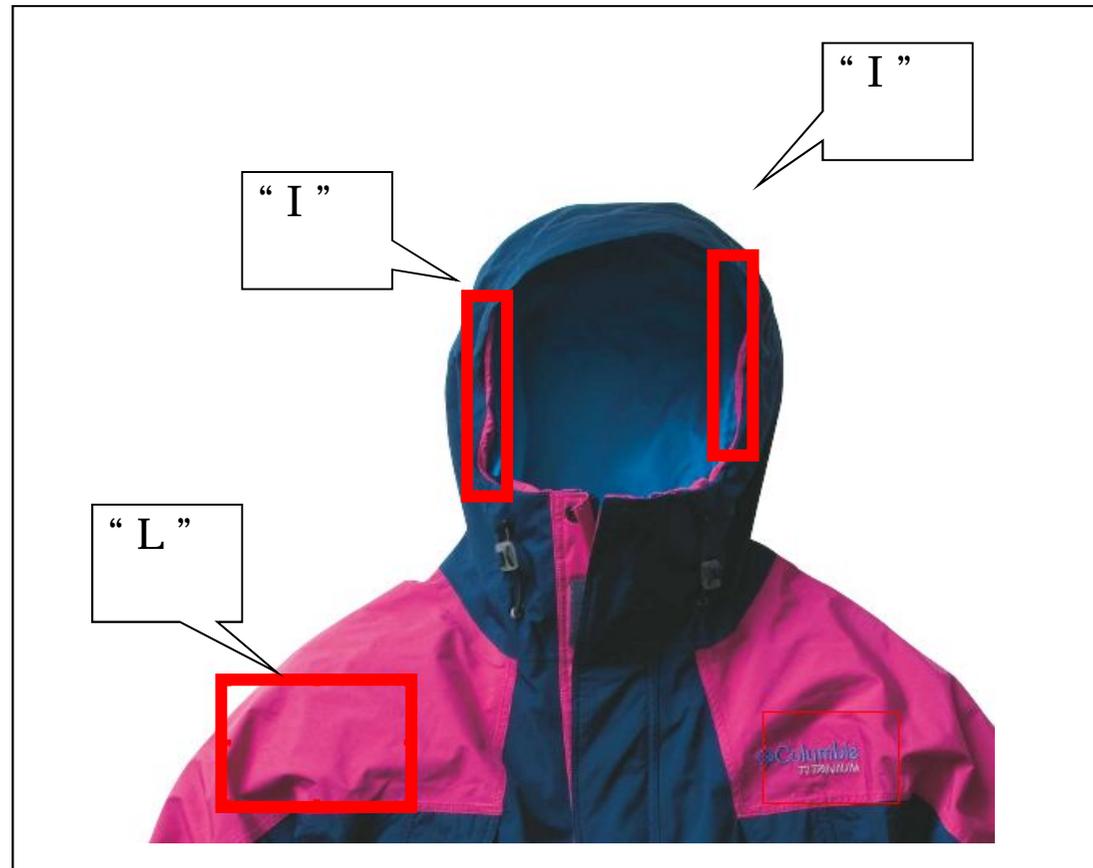


## ■ 領域抽出の失敗



# 画像からのテキスト抽出の課題(3)

## ■ 文字のない領域からの誤検出



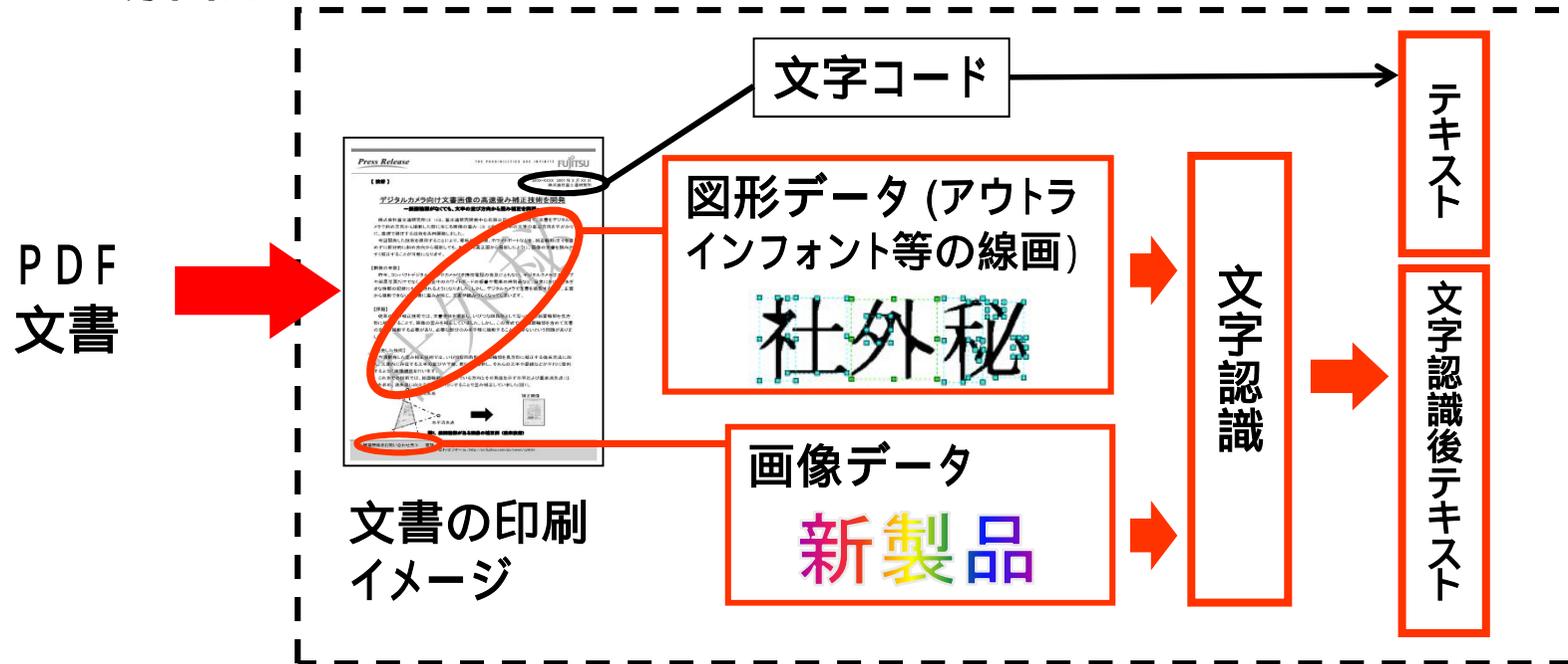


shaping tomorrow with you

# 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

- 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)

## ■ 読み順と異なる描画順

<p>6 日本の伝統息づく金糸を、現代に置き換えたジュエリー。</p>	<p>1 物ができる母親にも教わりました。作品ではなく、製品であること。その意識を持って作り続ける。</p>	<p>4 かぎ針を用いたニッティング。によって生まれる繊細なモチーフと大胆なカラリングが目を惹く。 [shingo matsushita]</p>	<p>7 コットンから始まり、シルク系、ウール、レザー、アンゴラなど、松下山はニッティングする素材自体も徹底追究。そして2007年、松下山さんが新たに使い始めた素材がこの、金糸。だった。 「よりよい素材を、と思つて辿り着いたのが金糸です」 主に高級絹織物で知られる西陣織をはじめ高価な着物の帯に使われている金糸。これによって、松下山の表現の幅はさらに広がる。 「パールを足したりチェーンを転がしたり。やはり金糸は和に転がりすぎるころがあるので、全体のバランスを大切にしています」 金糸を筆頭に、近頃はパールをメインとしたネックレスやピアスも充実している。 「パールは規則的なものよりも、不規則な配列をしているものが好き。組み合わせもビーズやスパンコール、シルク素材のアンティークスカーフを細かく裂いて編み込んだりしています。だから杉山さんのいうように、僕の作るものは、可愛らしくならないのかも。(笑)」 そう笑顔を見せる松下山さんは、どんな手法を使ってもこだわっていることがある。 「制作を始めた当時は、ヘッドドレスとか、一点ものを作っていました。今もそういう作品を作ることが好きです。でも [shingo matsushita] で作るものは、作品ではなく、製品。例えばなじみもなく高級品とされる金糸を使っていても、デザイナーとして、日常のなかで身につけてもらえようようなジュエリーを作り続けたいです」</p>	<p>3 「2年前ジュエリーを拝見して一発で気に入ったんです。時代を感じさせない普遍的な魅力があるし女性らしさがあまり前面に打ち出されていないところも好きです。それに色使いが着物の色合いに近しいというか、配色も気になって。そうしたら、これはデザイナーの(松下)真悟さんに後から聞いて初めて知ったんですが、着物に用いられている、金糸。がニッティングされていたんです。和の素材そのもので和を表現せずに、相手にもそれを気づかせることもないでも結果的にその、匂い、を感じ。素敵ですね」と話す杉山優子さん。以来、定期的に作品に触れているという杉山さんが、今回初めて松下山さんの作業場を訪れた。</p>	<p>2 「パールは規則的なものよりも、不規則な配列をしているものが好き。組み合わせもビーズやスパンコール、シルク素材のアンティークスカーフを細かく裂いて編み込んだりしています。だから杉山さんのいうように、僕の作るものは、可愛らしくならないのかも。(笑)」 そう笑顔を見せる松下山さんは、どんな手法を使ってもこだわっていることがある。 「制作を始めた当時は、ヘッドドレスとか、一点ものを作っていました。今もそういう作品を作ることが好きです。でも [shingo matsushita] で作るものは、作品ではなく、製品。例えばなじみもなく高級品とされる金糸を使っていても、デザイナーとして、日常のなかで身につけてもらえようようなジュエリーを作り続けたいです」</p>	<p>5 必要に迫られて始めた、ニッティングという手法。</p>	<p>2 「制作を始めた当時は、ヘッドドレスとか、一点ものを作っていました。今もそういう作品を作ることが好きです。でも [shingo matsushita] で作るものは、作品ではなく、製品。例えばなじみもなく高級品とされる金糸を使っていても、デザイナーとして、日常のなかで身につけてもらえようようなジュエリーを作り続けたいです」</p>	<p>4 [shingo matsushita] は、松下山さんが美容師、スタイリストを経て2003年に活動をスタートさせる。紙で制作した花のピアスや人毛で編んだネックレスなど、様々なスタイルでジュエリーを制作するなかで、現在の金属パーツにさまざまな素材をニッティングするという主なスタイルを確立したのは、2006年のこと。 「アイデアが浮かんだときに、それを作るにはどうしてもニッティングの技術が必要だったんです。でも編み物はやったこともなかったんで、本屋に駆け込んで編み物の本を読んで試したり。(笑) 編み</p>	<p>2 「制作を始めた当時は、ヘッドドレスとか、一点ものを作っていました。今もそういう作品を作ることが好きです。でも [shingo matsushita] で作るものは、作品ではなく、製品。例えばなじみもなく高級品とされる金糸を使っていても、デザイナーとして、日常のなかで身につけてもらえようようなジュエリーを作り続けたいです」</p>
---	--	---	--	---	---	--------------------------------------	--	---	--

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)

## ■ 読み順と異なる描画順：文字コード / 図形の混在

3

図形

1

2

本で絶大な人気を誇る山である。  
その日本海をスタートし、シー  
カヤック↓自転車↓登山の人力移  
動で大山山頂をゴールとする「皆  
生・大山 SEATOSUMMIT」が9月の5連休に開催され、中  
国・四国・関西地方をはじめ、日  
本各地から多くの選手が参加した。

本海から急激にせり上が  
る雄大な山容が特徴の大  
山（1729メートル）。西日

# PDFからのテキスト抽出の課題(3)

## ■ 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)

## ■ ルビ

「つがみ  
梅海  
新道に

PDF内の文字コード並び：

『「梅つが海み新道に」』

## ■ 2重テキスト

追憶の山、  
述懐の山

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)

## ■ 地図上に書かれた文字

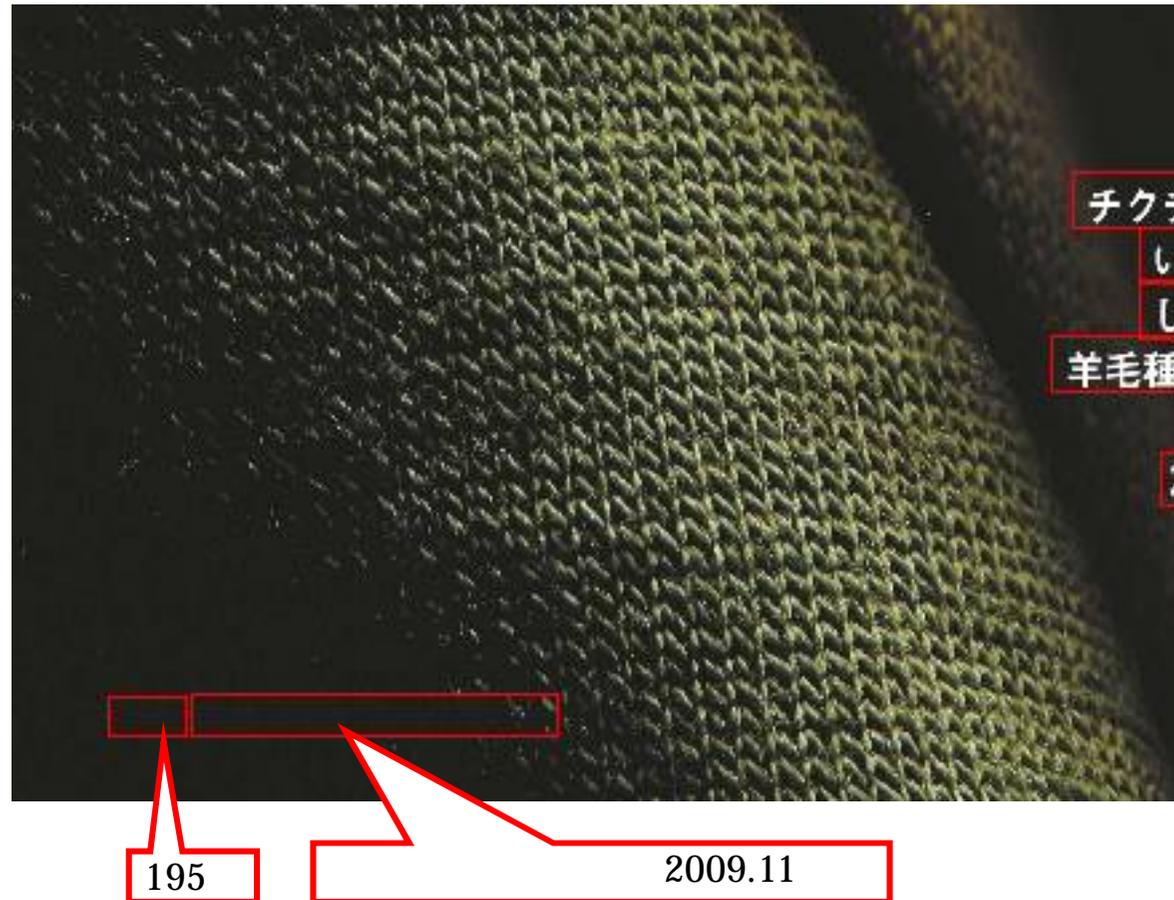
どこまでの文字を抽出し、文字並びをどう定義するか？



# PDFからのテキスト抽出の課題(6)

## ■ 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。





shaping tomorrow with you

# ガイドライン(案)

## ■ 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

## ■ 雑誌のレイアウトに対し、影響がない特性

可能な限り、テキストデータを残す。

雑誌上見えない文字については、削除する。

「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。

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

1行の中での文字サイズは同じ。

文字の並び(行 or 列)の方向は水平もしくは垂直。

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。

文字の並びの中に図形が混在しない。

特殊な文字(矢印記号、「」など)や外字が使用されていない。

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。

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)。

文字同士が隣接している(= 文字列として抽出しやすい)。

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



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。

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。

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。

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。

見た目のレイアウトに基づき、2重テキストを回避する。

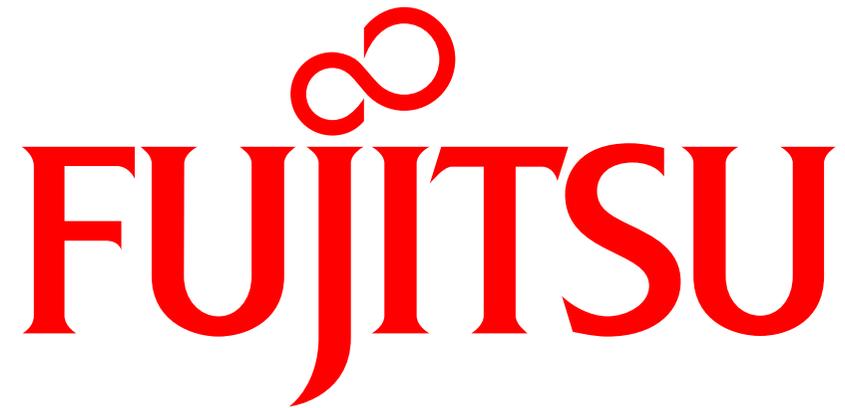
「タイトル」「見出し」「ページ」「図」「表」「テキストブロック」「箇条書き」などの属性を表すタグを付与する。

**FUJITSU**

shaping tomorrow with you

まとめ

- 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