画像データからのテキスト抽出に関するガイドライン

1. ガイドラインの目的・位置づけ、対象読者

本ガイドラインは、新たに作成する画像データや過去の膨大な画像データの利活用に向けて、それらに対して音声読上げや検索といった電子出版のアクセシビリティを確保するうえで、テキスト抽出に適した画像データに必要とされる特性を提示することを目的とする。合わせて、実証実験で判明した課題の解決に必要とされる機能を提案する。
本ガイドラインは、雑誌、書籍の制作に関わる出版業界、印刷業界を対象とする。

2. 用語定義

ガイドラインの用語については次のように定義する。

表 1 用語定義

用語	語義
アウトラインフォント	文字データを輪郭線で構成したフォント。
外字	コンピュータの文字システムにあらかじめ登録されていない文字のこと。後述の文字コードが割り当てられていないため、外字登録をしたコンピュータ以外では扱えない。
再現率	文書に含まれる正解のテキストのうち、テキスト抽出技術により正しく抽出できた文字の割合。（例. 正解テキスト 100 文字の中で 95 文字を抽出できれば再現率は 95/100*100=95%）
代替テキスト	図や写真など読上げが困難なものに対して、その内容を説明するテキスト。但し、属性として付与されているのみで表示はされていない。
タグ	文書構造に関する属性情報。タイトル、サブタイトル、段組み、画像、画像の代替テキスト等。
適合率	テキスト抽出技術で抽出されたテキストのうち、正しく抽出できている文字の割合。（例. テキスト抽出技術で抽出された文字が 100 文字で、そのうち正解テキストが 80 文字の場合、20 文字がノイズということになる。このときの適合率は 80/100*100 = 80%）

文字コード	文字を一意に識別するために各文字に割り当てられるバイト表現。Unicode、SJIS などがある。
OCR	光学的文字認識（Optical Character Recognition）。 一般的に、スキャンされた文書などの画像を入力とし、当該画像中に含まれる文字情報を抽出・認識し、テキスト化する技術。
PDF	Portable Document Format。Adobe Systems 社により開発された電子文書のためのフォーマットで、PDF1.7 は ISO 標準として認定されている。 なお、スキャンされた文書は画像のみから成る PDF であり、MS Word 等のアプリケーションから作成された PDF は文字コード、図形、画像が混在する。さらに、タグが付与された PDF と付与されていない PDF に大別される。

3. テキスト抽出における要件定義

読上げや検索に向けたアクセシビリティ機能を実現するために、PDF 内に含まれるどのテキストを読上げの対象にするのかを事前に定義する必要がある。ここでは、画像のみから成る PDF と、文字コード／図形／画像が混在する PDF のそれぞれについて、定義しておくべき項目について記述する。

表 1 画像のみから成る PDF に対する要件定義

・図に含まれる文字を読上げ対象にするかどうか
・写真に含まれる文字を読上げ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の設定 （再現率重視の設定では、なるべく多くの文字が抽出される。反面、文字以外のノイズも抽出されやすくなる。適合率重視の設定では、ノイズは抽出されにくくなるものの、その分、正しいテキストの抽出されにくくなる。）
・文字のサイズ、色、アンダーラインなどの属性に対する読上げルールの設定

表 2 文字コード／図形／画像が混在する PDF に対する要件定義

・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）を読上げ対象にするかどうか
・ルビがある場合に、漢字と仮名の読上げ方をどうするか
・図に含まれる文字を読上げ対象にするかどうか、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上げルール
・タイトル、見出し、小見出しに対する読上げルール
・(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
・外字に対する読上げルールの設定 （ex. OCR を用いて、最も字形の近い文字として扱う）

4. テキスト抽出に適した画像データ

テキスト抽出に適した画像データは以下のような特性を持つことが望ましい。

【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】

レイアウトの変更は、雑誌のもつアクティブな体裁を損ねてしまうため、あくまで、テキスト抽出がしやすいデータの参考になるものである。

① 1 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。

5. 画像データからのテキスト抽出に必要な機能

過去に作成された画像データ等、1.4 の特性をもっていないものについては、以下のような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。

上記を実現するには、従来の OCR 技術と、画像データからのテキスト抽出技術との連携が 100%の抽出ではないが、対応可能と考える。
以下に、上記①③に対する具体例を挙げて説明する。

図 1 に PDF からのテキスト抽出結果の例を示す。1 行の中で文字サイズの異なる文字が混在していると、PDF の内部では異なる行として扱われることが多い。その場合、見た目の文字並びと一致しないことが読上げの際に課題となる。

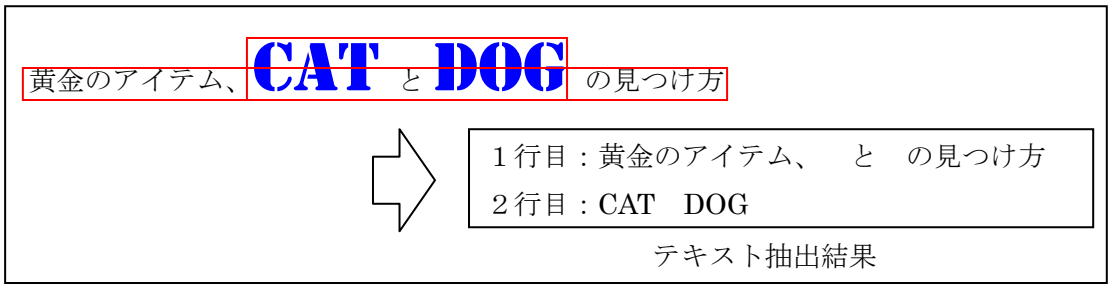

図 1 PDF からのテキスト抽出結果

図 2 には同じ文字列を画像として扱った場合の OCR によるテキスト認識結果の例を示す。一般的に OCR ではレイアウトの解析処理が入っているため、同じ行の文字は同一文字列として扱われやすい。一方で、文字認識精度が 100%ではないため、誤認識の可能性を含む。この例では、「ＤＯＧ」の”O”と”G”をそれぞれ”０”と”６”に誤読している。

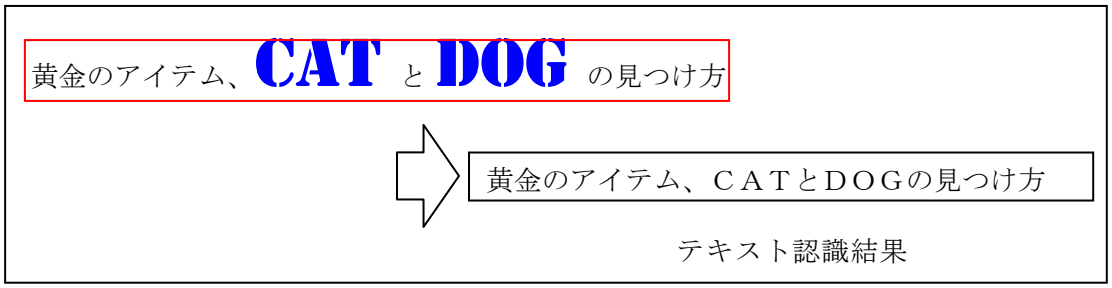

図 2 OCR によるテキスト認識結果例

PDF からのテキスト抽出では、テキスト情報を正確に取り出せる反面、見た目の文字並びと異なるという課題があり、OCR によるテキスト認識では、レイアウト情報を見た目の文字並びどおりに取り出しやすい反面、文字認識誤りを完全には回避できない。そこで、図 3 のように、両者を理想的にマージすることができれば、それぞれの弱点を補った高機能なテキスト抽出技術を構築できる。

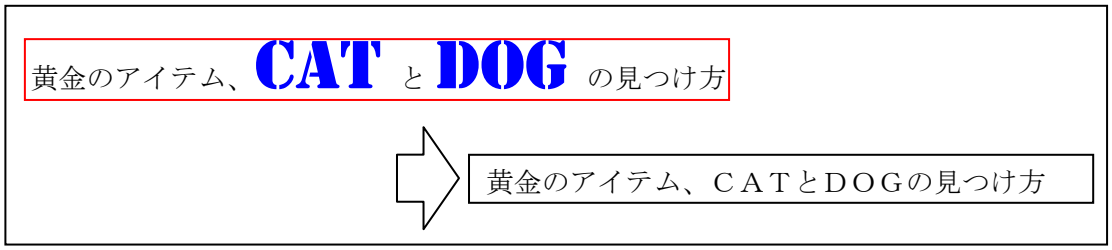

図 3 PDF からのテキスト抽出と OCR によるテキスト認識のマージ例