文字がくっきりと見える高圧縮PDF変換技術

High-Compression PDF Conversion Technology

土橋　外志正　　水谷　博之
■ DOBASHI Toshimasa　■ MIZUTANI Hiroyuki

東芝ソリューション（株）は，スキャナなどから取り込まれる文書画像を対象とし，高圧縮と高画質を両立する画像圧縮技術である高圧縮PDF（Portable Document Format）変換技術の開発を進めている。

高圧縮PDF変換技術は，文書画像を文字が主体の前景画像成分とそれ以外の成分とに分離し，それぞれに適した画像処理と圧縮を施した後，PDF形式に再統合することにより文書画像の高圧縮を行う。他の画像圧縮手法と比較して，文字輪郭がくっきりと読みやすい，ファイルサイズをより小さくできる，などの優れた特長を持ち，文字認識処理と連携してテキストによる検索が可能なPDFを作成することもできる。近年，カラーデジタル複合機（MFP）などのファイリング機能として採用が進んでおり，モバイル環境や業務帳票などへの展開も可能である。

In order to meet the growing demand for efficient document image compression, Toshiba Solutions Corporation has developed a high-compression PDF conversion technology suitable for color document images obtained by scanners, multifunctional peripherals (MFDs), and so on. This high-compression PDF conversion technology realizes smaller file size and better image quality compared with JPEG technology by separating character elements and non-character elements in the document and adopting the appropriate compression method for each element.

1　まえがき

近年のカラードキュメントの増加とともに，スキャナやデジタル複合機（MFP：Multi Functional Peripherals）などで入力されるカラー文書画像ファイルをカラーのまま保存したいというニーズが高まっている。しかし，カラー文書は情報量が多く，モノクロ文書と比較してファイルサイズが大きくなりがちであり，ファイル格納時に大きなストレージ容量が必要となったり，ネットワーク送信時に通信コストが高くなるなどの問題があった。

これらの問題を解決する手段として画像圧縮を行うことが考えられる。カラー画像の圧縮技術として一般的なJPEG（Joint Photographic Experts Group）圧縮は，風景写真などの自然画には向いているが，文字画像が大半の文書画像では文字のエッジを構成する高周波成分が多く，高解像度のまま圧縮率を上げると，文字の輪郭部がざわざわするモスキートノイズや画像がモザイク状となるブロックノイズが目だつ。逆に，解像度を下げて画像の高周波成分をカットすると，文字がぼやけて読みにくくなるという問題が存在している。

これら技術上の問題に対し，文書画像の画質を保ちつつファイルサイズが削減できる技術として各社で開発や製品への採用が進められているのが，高圧縮PDF（Poratable Document Format）変換技術（以下，高圧縮PDFと略記）である。

東芝ソリューション（株）は，40年にわたり文書画像から文字を読み取るOCR（Optical Character Reader）技術の研究・開発を進めている。ここでは，当社が培ってきたOCR技術を基に開発した高圧縮PDFとその応用について述べる。

2　高圧縮PDFの概要

高圧縮PDFとは，文書画像を対象とした画像圧縮技術である。文書画像を，**図1**に示すように，文字が中心の画像成分（以下，前景画像成分と呼ぶ）から成る前景レイヤと，それ以外の写真領域や背景領域の画像成分（以下，背景画像成分と呼

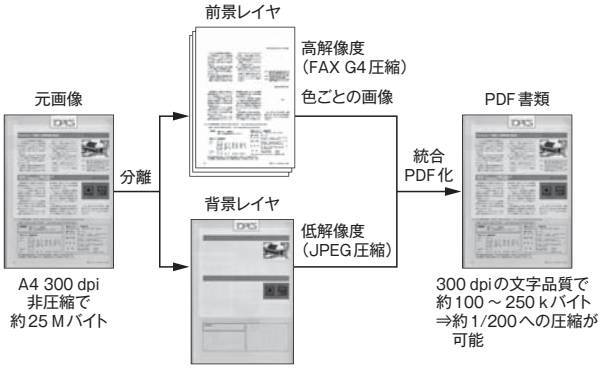


図1. 高圧縮PDFの概要 ― 高圧縮PDFは，文書画像を前景レイヤと背景レイヤに分離し，それぞれに適した手法を用いて圧縮することにより，画質の劣化を防ぎつつファイルサイズの大幅な削減を可能としている。

Outline of high-compression PDF conversion technology

ぶ）から成る背景レイヤに分離し，各成分の特徴に適した手法で圧縮することで画質や圧縮率を高めている。なお，このレイヤの分離は，単に画像の領域分割によって行うのではなく，前景画像成分を構成する画素単位での分離処理を行っている。

文書画像を前景レイヤと背景レイヤに分離するのは，それぞれのレイヤが異なる特徴を持つためである。文書画像を対象とする場合，文字を読むことに主眼が置かれるため前景画像成分には高い解像度が必要となるが，色の階調はそれほど重要ではない。また，背景画像成分では逆に，高い解像度は必要ないが色の階調は重要であることが多い。

そこで，前景レイヤの圧縮には，二値画像（白黒画像）の圧縮に適したFAX G4(注1)圧縮方式を用いる。FAX G4圧縮方式は対象が二値画像という制限はあるが，圧縮率が高いうえ，可逆の圧縮手法であるため二値化後は画質の劣化が生じない。前景画像成分が複数色から成る場合には，色ごとに前景レイヤを複数生成し，文字の読みやすさを損なわないよう高解像度のままで圧縮する。

背景レイヤは元画像から前景画像成分を除去したもので，その圧縮には階調を持つカラー画像の圧縮に適したJPEG圧縮方式を用いる。背景レイヤを構成する画素のうち，前景画像成分の分離によって画素値が欠落した画素に対しては，周辺の画素値による補間を行う。このようにして滑らかになった画像をJPEG圧縮する。

最後に，前景レイヤと背景レイヤを生成した後，背景レイヤ上に前景レイヤを重ねて配置してPDF形式にまとめたものが，高圧縮PDF画像となる。なお，前景レイヤには透過属性が付与されている（透明のシートに文字が書かれているようなもの）ので，前景レイヤどうしや前景レイヤと背景レイヤが重なっても，下レイヤに配置された画像成分が覆い隠されて見えなくなることはない。

PDF形式を画像フォーマットに採用する理由を以下に示す。

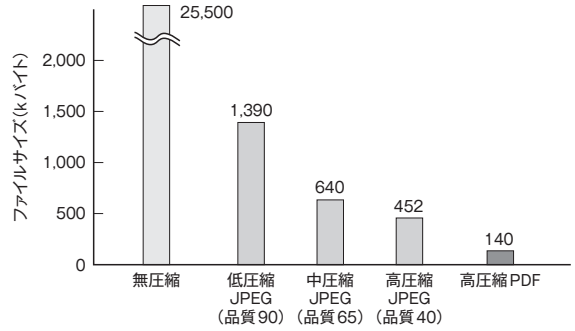
(1) PDF形式はコンテナとして異なる形式の画像やデータを混在させることができる
(2) 画像に透過属性を与えることができる
(3) 複数ページ文書を扱うことができる
(4) 広く普及したフォーマットである

3 高圧縮PDFの特長

高圧縮PDFの特長を，カラー画像圧縮で一般的に用いられているJPEG圧縮方式と比較して説明する。

3.1 コンパクトなファイルサイズ

高圧縮PDFは文書画像のファイルサイズを大幅に削減可能であり，A4サイズで300 dpi（dots per inch）の画像（約25 Mバイト）の場合，最大で約1/200（約120 kバイト）まで高圧縮が可能である。一般的なカラーのオフィス文書やパンフレットなどを高圧縮PDFで圧縮した場合と，JPEG圧縮した場合とのファイルサイズの比較を**図2**に示す。高圧縮PDFでは前景画像成分とそれ以外の成分を分離し，それぞれの成分の圧縮率が高くなるように画像変換と圧縮が施されるため，圧縮率を高圧縮に設定したJPEG圧縮よりも更に小さなファイルサイズとなっていることがわかる。



＊カラーのオフィス文書，パンフレットなどA4，16ページの平均値
＊元画像，JPEG，及び高圧縮PDF前景レイヤ解像度は300 dpi，高圧縮PDFの背景レイヤ解像度は150 dpi
＊JPEGの品質は，100を最大値（高画質），1を最小値（低画質）とするJPEG圧縮パラメータ

図2．高圧縮PDFとJPEGのファイルサイズ比較 — 高圧縮PDF画像のほうがJPEG画像よりファイルサイズが小さくなることがわかる。

File size comparison (high-compression PDF vs. JPEG)

3.2 文字がくっきり，高画質

高圧縮PDFは文字の読みやすさを重要視した圧縮手法であり，文字がくっきりと見えて読みやすい点で優れている。高圧縮PDF画像の文字がくっきりとしている理由の一つは，FAX G4圧縮の際に前景画像成分の二値化処理を施すことにより，スキャンによって生じた文字輪郭のぼけが解消されるからである（**図3**）。

高圧縮PDF画像と，その画像と同じファイルサイズとなるように作成したJPEG画像とで画質を比較すると，高圧縮PDF画像では元の画像の文字のシャープさを保ちながら画質も安

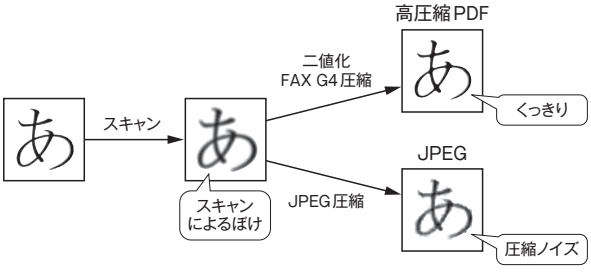


図3．文字の二値化による効果 — 文字画像の二値化によりスキャンによって生じたぼけが解消され，くっきりとした文字になる。

Effect of character image binarization

(注1) デジタル回線用のファクシミリの国際規格で，高解像度画像を小さいファイルサイズに圧縮できるのが特長。

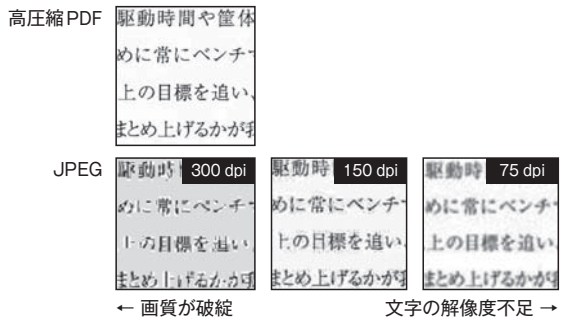


図4．高圧縮PDFとJPEGとの画質比較 — 同じファイルサイズの画像の画質を比較すると，高圧縮PDFのほうがより高画質であることがわかる。

Image quality comparison (high-compression PDF vs. JPEG)

定しているのに対して，JPEG画像はノイズが目だつ（**図4**）。一般に，JPEG圧縮は文字などのエッジがくっきりとした対象の圧縮には適していないため，文書画像を対象とした圧縮では，画質面でも高圧縮PDFが優位であることがわかる。

3.3　テキストによる検索（OCR処理との連携）

JPEG形式やTIFF（Tag Image File Format）形式のような画像フォーマットとは異なり，PDF形式にはユーザーには見えないデータを持たせることができる。そこで，文書画像に対してOCR処理を行い，認識結果として得られたテキストデータ（OCRテキスト）を透明テキストとしてPDFに埋め込み，高圧縮PDF画像のテキスト検索を可能にした。

OCRテキストが埋め込まれた高圧縮PDF画像は，PDFビューアの検索機能や，Windows Vista®(注2)，Google(注3)デスクトップ検索などの各種デスクトップ検索でのテキスト検索が可能となる。検索性の向上で蓄積された大量の文書画像の活用が容易となる。

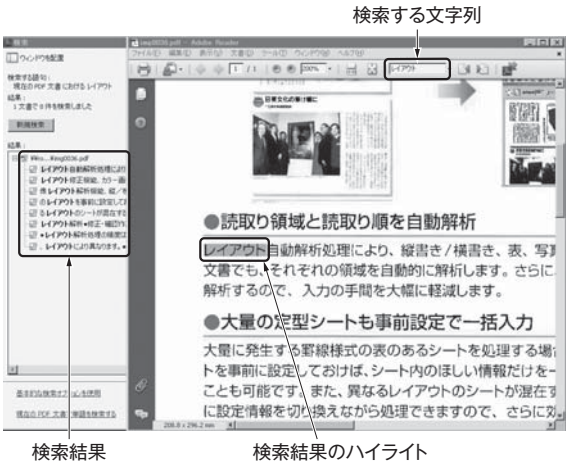


図5．テキスト検索の例 — 高圧縮PDF画像にOCRテキストを埋め込むことにより，高圧縮PDF画像に対するテキスト検索が可能となる。

Example of high-compression PDF text retrieval

OCRテキストが埋め込まれた高圧縮PDF画像に対して，Adobe® Reader®(注4)でテキスト検索を行った例を**図5**に示す。埋め込まれたOCRテキストは文字位置や文字サイズ情報を持っているので，検索文字列が含まれている文書を検索できるだけではなく，文書中の位置まで特定できていることがわかる。

4　開発した高圧縮PDFの処理の流れと特長

開発した高圧縮PDFの処理の流れを**図6**に示す。高圧縮PDF画像への変換過程は，多くの画像処理や認識処理で構成されており，これらの性能が高圧縮PDF画像の画質やファイルサイズに大きく影響する。

高圧縮PDFは前景画像成分と背景画像成分を分離して圧縮を行うため，前景画像成分に文字の抽出漏れがあるとその文字部分の可読性が悪化したり，写真領域の一部を誤抽出するとその領域の色の階調性が失われたりするといった画質への悪影響が生じる。そのため，色文字や反転文字及び複雑な背景（グラデーションや網点下地など）にも対応する前景成分抽出技術や，カラーレイアウト解析技術などを新たに開発し，文書画像中の文字の抽出漏れや誤抽出をできるだけ生じさせないよう対策した。

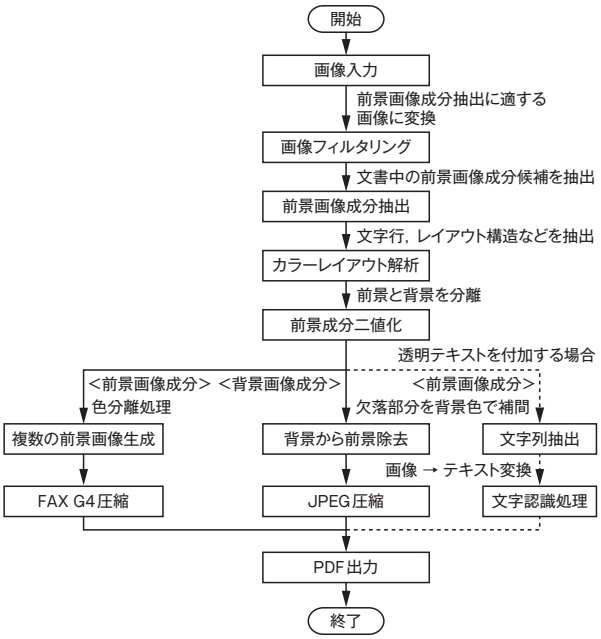


図6．開発した高圧縮PDFの処理の流れ — 高圧縮PDFは，様々な画像処理や認識処理から構成されている。

Flow of high-compression PDF conversion

（注2）　Windows Vistaは，米国Microsoft Corporationの米国及びその他の国における登録商標又は商標。
（注3）　Google は，Google Inc.の登録商標。
（注4）　Adobe，Adobe Readerは，Adobe Systems, Incの米国及びその他の国における登録商標又は商標。

また，テキスト検索可能な高圧縮PDF画像を出力する場合，文字認識処理が高圧縮PDF変換処理と密接に統合されている点も，当社で開発した高圧縮PDFの特長である。文書画像に対する文字認識処理と，高圧縮PDFに必要な処理とは重複する処理が多い。そこで，高圧縮PDFの処理過程で得られるレイアウト解析結果や，前景画像成分の二値化処理結果をそのまま文字認識処理への入力とすることで，処理の重複を避け高速化を図っている。また，圧縮によって劣化する前の文書画像に対し文字認識処理を施すことができるため，両処理が統合されていない場合に比べ文字認識の精度が高くなるというメリットもある。

5　高圧縮PDFの応用分野

5.1　スキャナやMFPによるドキュメントファイリング

高圧縮PDFの有用性がもっとも期待されるのは，スキャナやMFP利用時のドキュメントファイリング分野である。

高圧縮PDFは，近年，カラーMFPなどでファイリング機能として採用が進んでおり，今後も更なる普及が見込まれる。その実装形態は様々であり，MFP本体に組み込まれるほか，パソコン（PC）上で動作する画像ファイリングアプリケーションとして，あるいはスキャナなど画像入力機器のドライバの一部として実装されている。

高圧縮PDFのメリットの一つであるファイルサイズの大幅な縮小によって，限られたストレージに大量の文書画像が保存でき，容量制限のあることが多いEメールでの画像添付もより手軽なものとなるなど，文書画像の利用がよりいっそう促進されると考えられる。

5.2　モバイル環境[1]

今日ほとんどの携帯電話がカメラを備えており，文書画像の入力手段として用いられる機会も増えている。名刺やメモをはじめ，黒板やホワイトボードなどの文書を記録用に撮影することも少なくない。このような画像は，高い解像度で撮影されるためファイルサイズが大きくなりがちで，携帯電話の限られたメモリ容量を圧迫し，携帯電話とサーバ間で画像の送受信を行う際に時間がかかるなどの問題があった。

このような問題を解決するために，携帯電話などリソースが乏しい環境でも動作可能な高圧縮PDF変換機能を試作した。実装にあたっては，処理の簡易化及び高速化を行い，携帯電話の画像撮影機能やカメラ入力系に起因するシェーディングやぼけによる悪影響を回避する機能などを新たに開発した。高圧縮PDFは，携帯電話上で動作する文書管理や名刺管理アプリケーションなどへの応用が考えられる。名刺画像の表示例を**図7**に示す。

図7. 携帯電話での高圧縮PDFの例 — 携帯電話のような限られたリソース上でも動作可能な高圧縮PDFを開発した。

High-compression PDF conversion technology on mobile device

5.3　業務帳票への応用

各種の帳票をスキャンしてOCR処理を行ったり，画像をPCやサーバに蓄積したりといった業務用途においては，帳票内の印影や署名などは照合のために高画質で保存したいが，それ以外の記入部分はそれほどの画質を必要としないというような場合も多い。このような対象に対しては，印影部などの重要な部分だけを高解像度で低圧縮に，それ以外の部分は低解像度で高圧縮にすることで画像サイズを大幅に削減できる。

高圧縮PDFを，PDF形式の文書中に複数の画像形式や解像度を混在させられることを利用した画像圧縮手法と広くとらえると，その適用範囲は更に広がると予想される。

6　あとがき

前景画像成分と背景画像成分を分離してそれぞれに適した圧縮を施すことで，ファイルサイズの大幅な削減を実現する高圧縮PDFについて述べた。高圧縮PDF画像には文字がくっきり見えるという特長があり，文書画像に適した圧縮手法といえる。高圧縮PDFは多くの画像処理や認識処理で構成されており，当社が長年にわたり研究・開発してきた様々な文書画像処理技術とOCR技術が生かされている。

今後は，製品化に向けた開発を進め，更なる応用分野を開拓していく。

文　献

(1)　土橋外志正，ほか．“携帯端末による高圧縮PDF変換技術”．Media Computing Conference 2006．千葉，2006-06，画像電子学会．2006，p.97.

土橋　外志正　DOBASHI Toshimasa
東芝ソリューション（株）プラットフォームソリューション事業部要素技術開発部主任。文書画像処理の研究・開発に従事。
Toshiba Solutions Corp.

水谷　博之　MIZUTANI Hiroyuki
東芝ソリューション（株）プラットフォームソリューション事業部要素技術開発部参事。文字・パターン認識の研究・開発に従事。電子情報通信学会会員。
Toshiba Solutions Corp.