高圧縮PDF技術

Compact PDF Techonology

長谷川　史裕*　伊藤　仁志*　宮澤　利夫*　関口　優*
Fumihiro HASEGAWA　Hitoshi ITOH　Toshio MIYAZAWA　Yu SEKIGUCHI

要　　旨

カラー文書画像を視認性の低下を抑えながら高度に圧縮する．まず処理対象のカラー文書画像に文書領域識別処理を行い，文字だけの画像（以下，文字画像）と文字以外の画像（以下，背景画像）に分離する．文字画像は減色を行い，可逆圧縮する．背景画像は解像度を落として非可逆圧縮する．両者を重ね合わせ表示の表現が可能なPDFフォーマットでひとつのファイルにまとめる．生成されたPDFファイルは無料配布されているビューワーで閲覧可能である．文字と文字以外に適した圧縮手法を別々に施すことにより，従来の単一の圧縮手法に比べて高度な圧縮が可能となった．

ABSTRACT

Compact PDF technology enables color document images to be highly compressed without severe deterioration of images. First, an original color document image is separated into two text only images and non-text only images using the document segmentation process. Color reduction is achieved for text only images and compressed using a lossless method. Resolution of non-text only images is reduced and compressed with a lossy method. Both compressed images are overlapped in the PDF format, which allows images to be overlapped. A free viewer can display the produced PDF file. We achieved a higher compression rate by applying two different adequate compression methods for text only images and non-text only images than by applying one compression method.

* ソフトウェア研究開発本部 画像アプライアンス研究所
Image Appliance Lab., Software R&D Group

1. 背景と目的

近年，カラープリンタの普及や高解像度スキャナの登場，個人端末の高性能化によって，デジタル化されるオフィス文書は従来の白黒濃淡画像からフルカラー画像に移行しつつある．フルカラー画像では，フルカラーの写真や図の挿入，色文字による表現の強調など白黒画像では表現できなかった多彩な文章表現が可能となる．

しかし，多くの利点があるものの，一方でファイルサイズの増大が問題となる．というのは，オフィス文書がデジタル化される主な用途は，その蓄積，共有，配信であり，ファイルサイズが大きいほど，そのコストも増大するからである．そのため，一般的にフルカラー画像の画質を可能な限り維持したまま，ファイルサイズの削減を行う画像圧縮が望まれる．

デジタルカメラなどの画像を保存するために用いられる一般的な圧縮方式としてJPEGがある．JPEGは人間の視覚的特徴を生かした圧縮方式であり，写真などの階調が緩やかに変化する画像に対して有効である．しかし，文字のように階調が急激に変化するような画像に対してJPEGによる圧縮を行うと画質劣化が激しく，文字の視認性が低下する．オフィス文書は文字が多く含まれるためJPEGで圧縮すると文字の視認性が低下してしまう．

そこで高圧縮PDFでは，「文字の視認性が高く」かつ「ファイルサイズが小さい」フルカラー画像を実現する．そのために，画像の様々な特性を使って，文字と文字以外を分離し，文字と文字以外に適した圧縮技術を用いる．本報告では，高圧縮PDFの特徴を説明した後に，高圧縮PDFの有効性を実験的に示す．

2. 技術の特徴

高圧縮PDFの考え方は，領域特性に応じて複数の圧縮技術を適用するもので，この考え方の一例として，1990年に発表された「カラーファクシミリの構造化」などがある[1), 2)]．そこでは，フルカラー・ファクシミリ画像をフルカラー画像に適した圧縮技術を適用する領域，白黒濃淡画像に適した圧縮技術を適用する領域に分離し，それぞれをファクシミリで送信後，受信側でそれらを一枚のフルカラー・ファクシミリ画像に統合している．

高圧縮PDFでは，この考え方を大きく分けて以下の3つのステップに分離して，実現している（Fig.1参照）．

［分離過程］：対象とする画像を文字画像と背景画像に分離する．背景画像には文字以外の写真などが分類される．今後，背景と言う場合，対象画像の文字以外の部分を示すものとする．

［圧縮過程］：文字画像と背景画像をそれぞれに適した圧縮技術で圧縮する．

［統合過程］：圧縮された文字画像と背景画像を一つのファイル上で統合して高圧縮PDFとする．

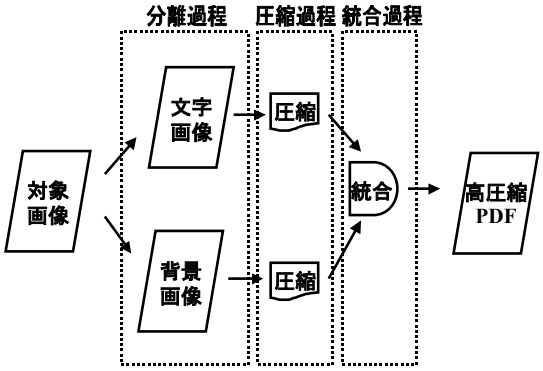

Fig.1　Procedure for Compact PDF.

下記各節では，ステップ毎の概要を説明する．

2-1　分離過程

分離過程では，対象画像を文字画像と背景画像とに分離する（Fig.2参照）．この分離によって文字画像と背景画像に対して，それぞれに適した圧縮技術を用いることができる．逆に言うと，この過程で誤分離が発生すると，圧縮過程のファイルサイズ面での利点を失うだけでなく，圧縮技術に適さない領域には画質面での悪影響が出てしまう．そのため，分離技術は高圧縮PDFの要素技術の中で最も重要な技術と言え，現在も盛んに研究が行われている[3), 4)]．

ソフトウェア研究開発本部では，OCR製品（Ridoc Document System[5)]）や文書画像処理ツールキット「リコードキュメントSDKシリーズ」[6)]の開発で多くの経験と技術を蓄積してきた．特に文字の分離を目的とした領域識別技術を保有しており，そのノウハウを十分生かすことで高圧縮PDFを実現することが可能となった．

Fig.2　An example of page segmentation.

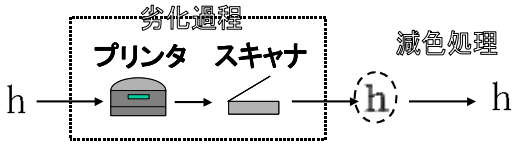

Fig.3　Image defect and recovery process.

Fig.4　An example of subtracted colors.

2-2　圧縮過程

圧縮過程では，分離過程で分離された文字画像と背景画像に対して，それぞれの特性に適した処理を施し，文字画像と背景画像のファイルサイズを削減しつつ，視認性の高い画像を生成する．

2-2-1　文字画像の圧縮

文字画像では，文字画像の特性に応じたファイルサイズ削減処理を行う．一つは文字の単色性を活かした減色処理，もう一つは文字に適した圧縮技術の適用である．

オフィス文書は通常，黒文字を主体とした数種類の色の文字で構成されることがほとんどである．また，文字の色は単色で一文字が複数の色で書かれることは，ほとんどない．

しかし，画像をプリンタで印刷することによって，画像内の色はシアン，マジェンタ，イエロー，ブラックの4色のドットで表現され，ドットのズレなどの問題が発生する．また，プリンタで印刷された原稿をスキャナで読み込むというアナログデータからデジタルデータへの変換などの問題もある．その結果，文字の単色性は失われ，文字の色数が多くなりファイルサイズが増加してしまう（Fig.3参照）．

そこで高圧縮PDFでは，劣化した色情報から劣化前の色を推定し，色数をN色まで減色する．この減色効果によってファイルサイズを，画像をそのまま圧縮する場合に比べて格段に小さくすることができる．

次に上記処理で得られたN色を用いて，文字の単色化を行い全体としてN色の多値画像を作成する．そして，色数の限定された多値画像に適したFLATE圧縮[7]と呼ばれる方式を用いて圧縮を行う（Fig.4参照）．

2-2-2　背景画像の圧縮

背景画像では，ダウンサンプリングとJPEGによる圧縮を行う．

ダウンサンプリングとは，複数のピクセルの色を統合して，サイズが小さい画像を生成する処理をいう．この処理を背景画像に適用しても視認性はあまり損なわれない．これに対して文字画像では，この処理を行うと影響が大きいため行っていない．

JPEGは写真などを圧縮するのに適用しており，背景画像から文字が取り除かれているため，さらなる高圧縮が可能となる．

2-3　統合過程

統合過程では，分離過程，圧縮過程で分離，圧縮した文字画像と背景画像を一つの画像ファイル上で統合する．画像の統合には，画像を重ね合わせて表示できる画像フォーマットなら何でも可能である．例えば，PDF以外にはJPM[8]などがある．

高圧縮PDFでは，PDF形式で統合を行っている．PDF[9]は官公庁などを中心にオフィスで広く使われている汎用フォーマットであり，無料配布されている専用リーダーをインストールすればOSの種類を問わずPDF文書を閲覧可能である．

3. 実験と結果

この章では，実験とその結果について示す．実験には9つの画像を用い，高圧縮PDFの有効性を画質とファイルサイズの面から評価する．

3-1　実験データと比較対象

実験に用いたのは，300dpiのフルカラー，ファイルサイズが25.5MByteとなる9つの画像である．画像の内容は，雑誌記事が4枚，カタログが2枚，地図が1枚，新聞記事が1枚，事務書類が1枚である．

比較対象としてAdobe Acrobat5.0を用いて圧縮された画像（以下，標準PDF）とJPEGによって圧縮された画像（以下，JPEG）を用いた．標準PDFは「文字の視認性が高い」画像の代表として，JPEGは「ファイルサイズの小さい」画像の代表として比較対象に選んだ．ただし，標準PDFはAdobe Acrobat5.0 Distillerのデフォルト・ジョブ・オプションを[CJKScreen]で圧縮した場合，JPEGは高い圧縮率を指定した（1～99まで選べる圧縮率をファイルサイズが最も小さくなる99に指定した）場合である．

3-2　ファイルサイズの比較

ファイルサイズの比較結果をTable1に示す．

Table 1　File size（KByte）．

原稿名（No）	オリジナル	標準PDF	JPEG	高圧縮PDF
雑誌A(1)	25,500	3,412	222	209
雑誌B(2)	25,500	1,220	253	275
雑誌C(3)	25,500	3,412	228	227
雑誌D(4)	25,500	1,536	327	309
カタログA(5)	25,500	7,338	275	249
カタログB(6)	25,500	937	294	235
地図(7)	25,500	12,302	310	488
新聞記事(8)	25,500	1,356	290	372
事務書類(9)	25,500	721	237	181

標準PDFと高圧縮PDFを比較した場合，高圧縮PDFの方が明らかにファイルサイズが小さいことがわかる．

次にJPEGと高圧縮PDFを比較した場合，高圧縮PDFが高い圧縮率を指定したJPEGと同等以下のファイルサイズになっていることがわかる．

この結果から，高圧縮PDFが「ファイルサイズが小さい」画像を実現できることがわかる．

3-3　画質の比較

画質の比較は定量的な評価が困難なため，主観的な評価を行った．評価は，9人の被験者に標準PDFとJPEG，高圧縮PDFを文字と背景の2つの項目で比較し，順位をつけてもらうことで行った．ただし，順位は品質が同等ならば，同一順位もありとした．

評価結果をTable2，Table3に示した．Table2，Table3はそれぞれ，標準PDFと高圧縮PDF，JPEGと高圧縮PDFを比較して，順位が高かった人数を示している．同等は順位がどちらも同じだった場合である．また，画像の一部をFig.6～Fig.8に示す．

Table 2　Image Quality（PDF vs. Compact PDF）．

原稿名（No）	文字			背景		
	標準PDF	同等	高圧縮PDF	標準PDF	同等	高圧縮PDF
雑誌A(1)	9	0	0	4	5	0
雑誌B(2)	4	5	0	6	3	0
雑誌C(3)	1	5	3	2	7	0
雑誌D(4)	8	1	0	3	4	2
カタログA(5)	5	4	0	4	5	0
カタログB(6)	4	4	1	9	0	0
地図(7)	6	1	2	9	0	0
新聞記事(8)	9	0	0	8	1	0
事務書類(9)	8	1	0	8	1	0

Table 3　Image Quality（JPEG vs. Compact PDF）．

原稿名（No）	文字			背景		
	JPEG	同等	高圧縮PDF	JPEG	同等	高圧縮PDF
雑誌A(1)	0	0	9	0	0	9
雑誌B(2)	0	0	9	0	0	9
雑誌C(3)	0	0	9	0	0	9
雑誌D(4)	0	0	9	0	0	9
カタログA(5)	0	0	9	0	0	9
カタログB(6)	0	0	9	0	0	9
地図(7)	0	0	9	0	0	9
新聞記事(8)	1	1	7	0	0	9
事務書類(9)	0	0	9	1	1	7

標準PDFと高圧縮PDFを比較した場合，雑誌やカタログなどの原稿に関しては半数近くの人が標準PDFと高圧縮PDFの画質が同等と答えており，高圧縮PDFがファイルサイズの大きな標準PDF並みの画質を実現できていることがわかる．また，標準PDFの方が画質が良いと答えた人も，文字を背景として分離してしまった部分に注目した評価であり，高圧縮PDFとの差は僅差であると答えていた．

ファイルサイズがほぼ同等のJPEGと高圧縮PDFを比較した場合，ほとんどの原稿で高圧縮PDFの方が画質が良いという結果になった．

この結果から，高圧縮PDFが「文字の視認性が高い」画像を実現できることがわかる．

3-3　総合評価

ファイルサイズと画質の評価から，高圧縮PDFは「文字の視認性が高い」かつ「ファイルサイズが小さい」画像を実現できていることがわかる（Fig5参照）．

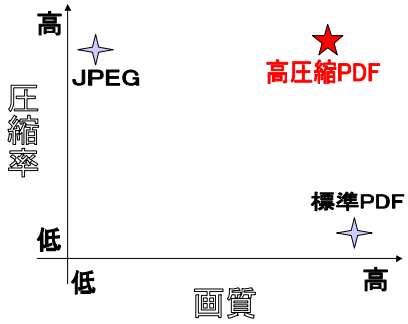

Fig.5　Performance.

4. まとめ

フルカラーのオフィス文書画像を「文字の視認性が高い」，「ファイルサイズが小さい」画像にする圧縮技術の開発に関して，高圧縮PDFを開発し，実験的に有効性を示した．

今後は，高圧縮PDFのコア技術である文字画像と背景画像の分離技術について，さらなる精度向上を目指す．また，一部機器ではすでに実装されているように，高圧縮PDFにOCR結果のテキストを貼り付けることで，検索を可能にするなどの機能面からの差別化を図る予定である．

参考文献

1) NTT：Color extension for Group 4 facsimile，CCITT Com VIII-80，(1990).

2) 松木 真：カラーファクシミリへの文書構造化，1993年電子情報通信学会春季大会予稿，D418，(1993).

3) 堀田 裕弘，など：均等色空間に基づくカラー画像の領域分割，信学論（D-II），vol.J74-D-II，no.10，pp.1370-1378，(1991).

4) 高橋 圭子，など：ISODATAクラスタリング法を用いたカラー画像の領域分割，信学論(D-II)，vol.J82-D-II，no.4，pp.751-762，(1999).

5) Ridoc Document System：http://ext.ricoh.co.jp/ridoc_ds/rds/.

6) 文書画像処理ツールキット「リコードキュメント SDKシリーズ」：http://ext.ricoh.co.jp/document-sdk/.

7) P.Deutsch：DEFLATE Compressed Data Format Specification version 1.3，RFC1951，(1996).

8) Information Technology-JPEG2000 Image Coding Standard-Part6:Compound Image File Format，ISO/IEC FDIS 15444-6.

9) アドビシステムズ：PDFリファレンス，(株)ピアソン・エデュケーション，(2001).

注1) Adobe Acrobat5.0はadobe社の製品です。
http://www.adobe.co.jp/

Fig.6　PDF sample.

Fig.7　JPEG sample.

Fig.8　Compact PDF sample.