コンパクトPDFの開発

Development of Compact PDF

吉田宏樹＊ Yoshida, Hiroki
糀谷香美＊ Koujiya, Kagumi
正木賢治＊ Masaki, Kenji

要旨

企業向けのデジタル複合機市場では，各機種の付加価値を高めるため，様々な機能が搭載されている。この中でもスキャナで読み取ったドキュメント画像を，従来のJPEG圧縮での保存よりもファイルサイズをより小さくしてPDFファイルに変換する技術開発（コンパクトPDF変換）が注目されている。この機能によって，スキャンした画像をメールで送る“Scan to mail”の通信時間が短くなり，より使いやすくなる。

ファイルサイズを小さく（圧縮）することは，画質とのトレードオフの関係にあるが，本技術開発では，文字と写真を分離して，それぞれ圧縮することにより，画質を保持したまま，ファイルサイズの小さいPDFに変化する技術を確立した。本稿では，このコンパクトPDF変換を実現するための技術について述べる。

Abstract

Various functions are incorporated to raise the added value of each model in the digital MFPs for business. Technological development of the compact PDF conversion is promising, which makes image data size smaller than that of simple JPEG compression for document images captured via scanner. The compact PDF conversion makes the time of “scan to mail” shorter and improves the usability.

Although, compression of image generally degrades image quality, we have developed a technology to convert scanned image data into the compact PDF without visual image quality degradation, whereby images of characters and photographs are separated and compressed respectively.

This report describes the technology developed to realize such compact PDF conversion.

＊コニカミノルタテクノロジーセンター㈱
システム技術研究所　イメージシステム開発室

1　はじめに

企業向けのデジタル複合機市場では，各機種の付加価値を高めるため，様々な機能が搭載されている。この中でも，原稿のカラー化に伴い，スキャナで読み取ったドキュメント画像を，従来のJPEG圧縮した保存よりもファイルサイズを小さくしてPDFファイルに変換する技術開発（コンパクトPDF変換）が注目されている。

コンパクトPDFに変換するためには，まずドキュメント画像から，精度良く文字と写真を分離する必要がある。その後，文字には文字用の最適な圧縮を，写真には写真用の最適な圧縮を施して，両者を別レイヤーとして重ね合わせPDFファイルフォーマットで一つにまとめる（**Fig. 1** 参照）。これによって，従来の単一の圧縮（例えば、JPEG圧縮）よりもサイズの小さいファイルに変換できる。

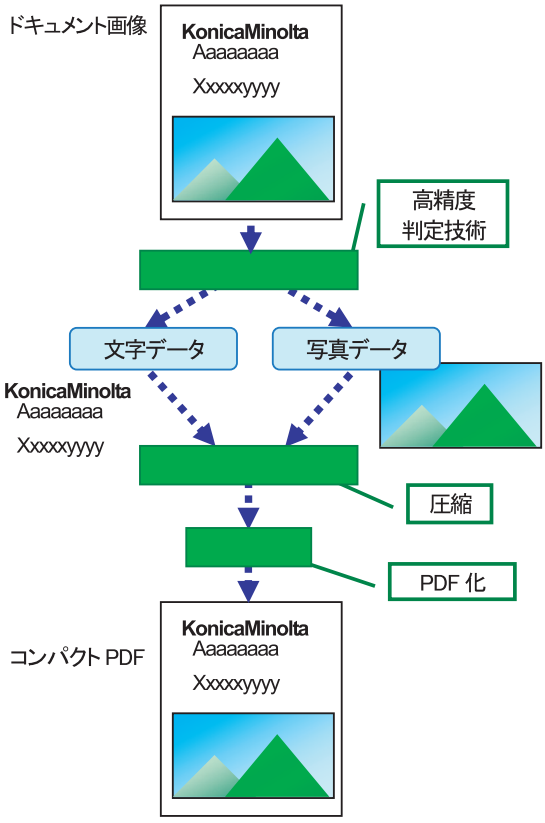

Fig.1 Concept of Compact PDF

本稿では，この技術開発のうち，文字と写真を分離する技術及び評価を中心に述べる。

2　構成の特徴

コンパクトPDF変換技術の特徴は，以下の２点である。

⑴スキャナ等で読みとったフルカラードキュメントイメージを“文字”と“写真”の領域に自動的に分離

⑵文字領域，写真領域を，それぞれに適した圧縮を施し，ファイルサイズの小さいPDFに変換

本変換技術は，Fig. 2 の手順で処理を行う。

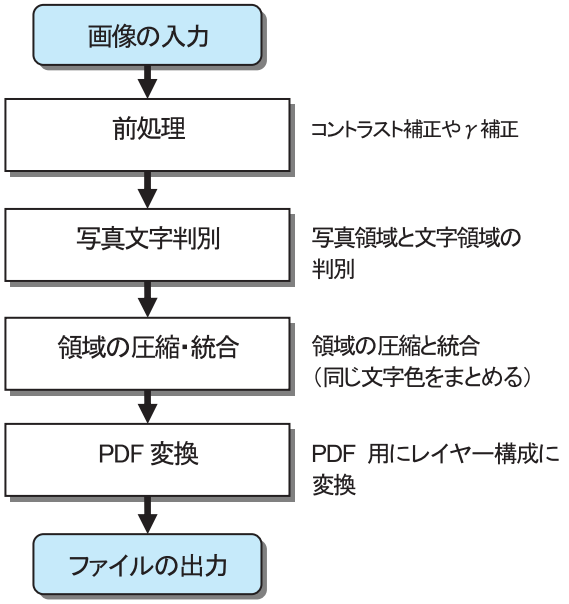

Fig.2 Processing flowchart

以降では，写真と文字の判別，及び領域の圧縮統合のそれぞれについて述べる。

3　写真文字複合判定

ドキュメント画像中から，写真及び文字を判別する方法は，現在多くの方法が，研究及び開発[*1)*2)]がなされている。

この文字判別と写真判別は，コンパクトPDF変換において最も重要な技術であり，性能を大きく左右する。判別アルゴリズムには，当社が保有する写真・文字画像に対する豊富な研究から生まれた技術を利用して，写真・文字それぞれを判別する技術を確立した。

本技術の特徴は，従来一つの基準で分離されていた写真と文字の領域に対してそれぞれ別々の技術より抽出し，それぞれの結果を統合し，重複領域に対してその相関関係から動的に基準を作成して判定し，より詳細に文字と写真の分離を行う所にある。

特に写真判別技術については，写真画像データの小領域の詳細な特徴量解析技術により，従来の技術と比較して小さな写真までも正確に抽出する所にある。従来では，小さな写真領域は，文字として誤判別するケースが多々あったが，この技術によって，正確に写真領域を限定することが可能となる。また文字判別技術については，コピー機の開発で培った文字特性を抽出する技術をベースにして，下地のあるなしに関わらず高精度に文字のみを抽出できるところにある。これは文字のレイアウトの情報から，より厳密に写真の分離を行うものである。これによって，従来の技術で問題となっていた，写真内の不自然な文字の抽出を抑え，写真に隣接した文字の未抽出を抑えることが可能となった。

Fig. 3 に写真・文字の分離処理のサンプル画像を示す。Fig. 3 -⑴は，スキャナで取り込まれたドキュメントイメージである。Fig. 3 -⑵は，通常の写真文字分離処理を行った結果であり，通常の文字部分は正確に抽出して２値化されている。しかし，写真領域の一部を文字として誤判別しており，誤って２値化されている。これに対し，本技術を適用したFig. 3 -⑶では，写真と文字が正確に分離できている。

⑴Original Image（Photo）

⑵Conventional way

⑶Proposal way

Fig.3 Result of photo and character separation

更に，従来は，領域分離そのものが正確でも，色文字が混在する箇所では，文字毎に正確に分離ができず文字色を間違って解釈することがあった。本技術では，文字として抽出した箇所に対して，文字の色相と文字サイズから，領域を限定し，より正確な文字色へと設定する。この技術によって，従来問題点となっていた，文字を２値化した後の文字色の誤りがなくなり，自然な変換が可

能となる。

Fig. 4 に文字処理部のサンプル画像を示す。Fig.4-(2)の従来技術を利用した処理結果では，“W”と“r”がそれ以外の文字と接触しているため，単一の文字として誤認識されてしまい，文字色が不自然となっている。

これに対し本技術を適用したFig.4-(3)では，正しい文字色を付与している。

(1)Original image

(2)Conventional way

(3)Proposal way

Fig.4 Result of character

4 領域の圧縮・統合

文字と写真を分離した後，それぞれの領域毎にまとめて，PDF用のレイヤー構成に最適な形に変更する。

4．1 文字圧縮

文字領域では，以下のステップで文字特性に応じた処理を行い，視認性を高めた圧縮を行う。

(1)文字色の 2 値化（原色）

(2)G4圧縮

文字領域では，多階調整を保ったままJPEG圧縮を行うよりも，2 値化した方が視認性は高い。

また 2 値化して，可逆圧縮のG4圧縮を行うことによって，ファイルサイズをより小さくすることが可能となる。

4．2 写真画像の圧縮

写真領域には，適切な解像度に変換した後，JPEG圧縮を行う。

写真の場合，解像度を落としたとしても，視認性には大きく影響せず，また高速に高圧縮が可能なJPEG圧縮を用いた。

4．3 統合処理

文字及び写真をそれぞれレイヤーとして，再構成を行う。文字画像は，それぞれの文字が 2 値化されているため，文字色ごとにまとめてレイヤー化する。これらのレイヤーを，重ね合わせてPDFファイルに変換する。

5 評価

ここでは，評価方法とその結果について示す。

評価において，圧縮効果を検証するために，ファイルサイズの測定を行った。また画質評価においては，できるだけ定量的な評価を実現するため，以下の 3 項目を測定した。

(1)文字の可読性　処理前と処理後の文字に対して，OCRを行い，認識率を算出した

(2)文字の誤判別・欠損チェック　文字・写真分離後に誤った箇所の面積（誤判別している箇所を囲む最小の四角形の面積）を算出した

(3)写真の誤判別チェック　文字・写真分離後に，写真を誤って文字として判別した箇所の面積（誤判別している箇所を囲む最小の四角形の面積）を算出した

文字の可読性評価におけるOCRは，市販のOCRソフトを利用した。

5．1 実験データ

実験に使用した画像は，当社MFP“bizhub C350”を使用して，300dpi，A4，フルカラーの条件でスキャンした，カタログ原稿 4 枚，オフィス原稿（文字中心） 3 枚，オフィス原稿（表） 3 枚，雑誌原稿2枚の計12枚である。また文字の可読性評価については，専用の評価画像を 6 枚用意した。それ以外の評価は，上記の画像を使用している。

比較対象としては，Adobe製，Acrobat5.0を用いて圧縮されたPDFを用いた。Acrobat5.0を使用する際には，標準的に文字の視認性が高い基準として，Acrobat5.0 Distillerのデフォルト・オプションを指定した。

5．2 ファイルサイズの評価

ファイルサイズの比較結果をTable 1 に示す。

Table 1 Results of file size (unit: KByte)

ファイル名	オリジナル	Acrobat	本技術
カタログ(1)	25654	680	114
カタログ(2)	25655	627	130
カタログ(3)	25658	903	154
カタログ(4)	25654	954	177
オフィス・文字(1)	25647	285	66
オフィス・文字(2)	25651	410	84
オフィス・文字(3)	25649	325	85
オフィス・表(1)	25653	561	150
オフィス・表(2)	25650	458	121
オフィス・表(3)	25654	566	116
雑誌(1)	25657	1134	173
雑誌(2)	25657	1052	167
平均	25653	663	128

いずれの画像についても，Acrobatの標準的なPDFファイルよりもファイルサイズは小さく，平均20%弱のファイルサイズを実現している。

Table 1 では，オフィスで一般的に利用されるA4サイズに対する結果であるが，それ以外の原稿サイズ，B5サイズ，A3サイズの原稿に対しても同様に，Acrobatの標準的なPDFファイルよりも小さく，20%弱のファイルサイズを実現している。

5．3　画質の評価

・文字の可読性

文字の可読性評価では，処理を行う前と処理後に対してOCR処理を行い，文字の認識率の変化を測定した。評価結果を**Table 2** に示す。

Table 2 のデータは，1000文字に対して，1000文字正しく認識した場合を100%としている。

変換前と比較すると，認識率に低下はなく，逆にあがるケースが発生する。これは，文字の2値化によってより判別しやすくなった結果と考えられる。

Table 2 Results of OCR processing (unit: rate)

ファイル名	変換前	変換後
原稿(1)	97%	98.6%
原稿(2)	88.8%	99.2%
原稿(3)	87.2%	99.4%
原稿(4)	97.7%	97.7%
原稿(5)	99.6%	99.6%
原稿(6)	95.2%	98.2%

・文字の誤判別・欠損チェック

文字の誤判別チェックには，ファイルサイズの評価で使用した原稿を用いて評価を行った。

評価結果を**Table 3** に示す。

Table 3 Results of extraction missing of characters (unit: rate)

ファイル	最悪値	平均値
雑誌･オフィス原稿 12 枚	0.1%	0.01%

評価値は，誤判定・欠損している面積を，画像全体の面積で割った値である。

評価においては，誤判定が目立つ基準として，“0.4%”を設定したが，全ての原稿において，基準値を下回り，誤判定，欠損は少なかった。

・写真の誤判別チェック

写真の誤判別チェックには，文字の誤判定チェックの際と同様の原稿を使用して評価を行った。評価結果を**Table 4** に示す。

Table 3 Results of extraction missing of characters (unit: rate)

ファイル	最悪値	平均値
雑誌･オフィス原稿 12 枚	0.01%	0.0%

評価した中では，目立つ誤判定は見受けられなかった。

5．4　評価まとめ

今回のファイルサイズの評価，及び画質の評価より，我々のコンパクトPDFは，ファイルサイズが小さく，文字の視認性が高く，誤判別も少ないことが確認された。

6　まとめ

スキャナで読み取ったフルカラードキュメント画像に対して，従来よりもファイルサイズがコンパクトで，誤判別が少なく高画質なPDFファイルへと変換する技術を確立した。また評価によって，その有効性を証明した。

今回の技術開発において，文字と写真それぞれに独立した分離基準値の導入とその結果に対しての動的な総合判別処理により，問題となっていた写真と文字の分離性能を向上させ，更に文字色判別処理により，文字色の抽出ミスを抑えることが実現できた。また，分離性能向上の副次的効果として，写真領域に対して思い切った圧縮を行うことが可能になり，全体の圧縮率向上に大きく寄与した。

この結果，スキャンした画像をメールで送る“Scan to mail”において従来サイズの点で敬遠されていたカラーの複数枚原稿送信時も実用的な利用が可能となった。更にOCR認識率の低下がないため，アーカイブやコピーログなどセキュリティ用途への展開も期待される。

今後は，この技術を様々な機器やソフトウェアに搭載し，同時にこの技術のコア部分である，写真・文字分離技術について更なる精度向上・性能向上を目指していきたい。

謝辞

本技術開発は，コニカミノルタビジネステクノロジー㈱制御開発本部　ソリューションセンター　ソリューション開発部の森　俊浩さん，廣岡　義昭さんの協力により成しえたものであり，ここに感謝の意を表する。

●参考文献

1）前田純治，南條英一，電子情報通信学会論文誌　D-Ⅱ vol.J78-D-Ⅱ No.**11** pp.1726 - 1729 (1995/11).

2）黄瀬浩一，大町真一郎，内田誠一，岩村雅一，電子情報通信学会技術研究報告，PRMU2004 - 246, (2005/3)