# PDFによる超高圧縮カラーコンテンツ作成効率化システムの開発

The development of the PDF based high compressed full color contents generation support system

本田 克己[1)]　芦塚 浩一[2)]
Katsumi HONDA　Koichi Ashizuka

1)株式会社ハイパーギア
（〒349-0123 埼玉県蓮田市本町２丁目１９番　渡辺ビル E-mail: hg-honda@po.iijnet.or.jp）
2)株式会社ハイパーギア　技術部
（〒349-0123 埼玉県蓮田市本町２丁目１９番　渡辺ビル E-mail: ashi@sb3.so-net.ne.jp）

**ABSTRACT.** Although it is important, converting paper information such as a catalog into the electronic media for the Internet is difficult because of the data size. The HC-PDF technology which is based on the de facto standard of an electronic document PDF, compresses character and picture portion separately, so that reduce data size dramatically, but the productivity is not good because of manual process of the graphics. We developed two Linux based systems, one for graphics workstation which separate character, picture portion and compensate scanned image, and one for ASP server which convert images to HC-PDF, so that at least twice productivity improvement can be gained

## 1. 背景

インターネットを用いた電子商取引の分野においては、ユーザに提供できる情報の量が取引の成否を決める鍵であるが、ウェブ上では要約された比較的容量の小さい情報しか提供できなく、このため、取引の機会を逸したり、ユーザは別途郵送などの手段で紙のカタログを送付してもらうなど不便を強いられたりしているのが現状である。

当社では、これらの問題を電子文書のデファクトスタンダードであるPDFを利用して解決するHyperCompact PDF（ＨＣ－ＰＤＦ）を開発した。これは標準のPDFフォーマット上で、画像部分と文字部分をPDFのマルチレイヤーの最適化した画像ファイルとして合成することにより、例えば旅行会社のパンフレットなど、FAXでもつぶれてしまう程細かい文字を含むカラーの印刷物でも、１ページ100-200KBと充分実用になるサイズに高圧縮する技術である。これにより高速の回線をもたない、一般のインターネットユーザでも、１ページ20-40秒程度で細かなパンフレットなどをいつでも取出し、安価なインクジェットプリンタでも１分程度で印刷することができる。しかしながら、高圧縮を実現するためには、自動処理だけでは変換できないという点があげられる。汚れ、ゆがみ、網点処理の補正や、文字部分と画像部分の分離など手作業による補正、調整が１ページあたり2,3時間かかり、またそのためには高度な画像処理ソフトウエアが使えるデザイナが必要など、限られた人しか作業できなかった。このように、ＨＣ－ＰＤＦ形式は、小さい容量で詳細な情報を提供できるデータ形式であるが、紙の印刷物をスキャンして得たデータからＨＣ－ＰＤＦ形式のデータを作成するツールとして、画像と文字を効率的に分離処理する機能を含むソフトウェアの開発が期待されている。

## 2．目的

本開発では、紙の印刷物をスキャンして得た画像データから、画像と文字を分離処理する機能を含む画像処理システムと、分離した各々のデータからＨＣ－ＰＤＦ形式のデータを生成するサーバーシステムを開発するものである。本開発では、画像の補正を極力自動化し、また文字部分や画像部分の分離もLinuxで、誰でも使える専用のツールを用意することにより、これらを在宅の主婦などで可能にするサービスビジネスとして展開できるようにして、世の中にある多くのカタログや、パンフレットなどを極めてリーゾナブルなコストで電子化できる仕組みを提供することを目的とする。また補正などのツールに加えて、多量の計算能力を必要とするPDF変換部分は、HC-PDF変換サーバをASP(Application Service Provider)として用意することにより、各作業者毎に高価な設備投資をしなくてもすむ仕組みを開発する。これにより、一連の作業を、家庭の主婦や障害者を含むＳＯＨＯ事業者と協業して実施することにより、新たなビジネスモデルの創出を目指すものである。

## 3．開発内容

### 3．１　画像処理方式の開発

本開発では、スキャンされた画像データから、極力自動処理により、効率よく、画像と文字の分離を行い、また、スキャナ入力による品質の低下などを補正するために、以下の開発項目に対して、その実装方式を検討して、ソフトウェアとして実現し、その結果を評価するものとする。

（１）　文字画像分離方式の開発

カタログやパンフレットを含む印刷物をスキャンして

得た画像データから、文字部分を分離するために、画像データに最も多く含まれる色を背景色と判定し、背景色からある閾値以上色が異なる部分を文字部分であると認識する方式、画素の濃度が著しく変化した点を検出し文字部分を判定する方式などの方式を利用して、文字部分画像データと背景画像データを生成する方式を開発する。

（２）　裏映り補正方式の開発

両面印刷された印刷物をスキャンした際、おもて面の画像に裏面の画像が薄く映ってしまう現象を補正するために、画像データを統計処理し、裏面の画像による影響を推定する方式、実際にスキャンした裏面の画像データを減算処理する方式などの方式を利用して、裏映りを補正する方式を開発する。

（３）　網点補正方式の開発

カタログやパンフレットを含む印刷物をスキャンして得た画像データから、網点を除去し、画像データの容量を縮小させるために、２次元 FFT により特定の周波数成分を除去する方式、短い同濃度画素の連なり（以下、「ラン」という。）の出現密度から網点領域を判定し、ガウシアンフィルタをかける方式のいずれかの方式を利用して、網点を補正する方式を開発するものとする。

## ３．２　応用システムの開発

前項で開発した技術を実装する形で、印刷物のスキャン、手作業による補正などをＳＯＨＯ事業者側でおこなう画像処理システムと、そこで作成されたデータをインターネットを通じて受け取り、負荷の重い HC-PDF 変換などをＡＳＰサーバとして処理するサーバーシステムの二つに分けて開発をおこなった。

### ３．２．１　システムの構成

以下の図１に全体のシステム構成図を示す。

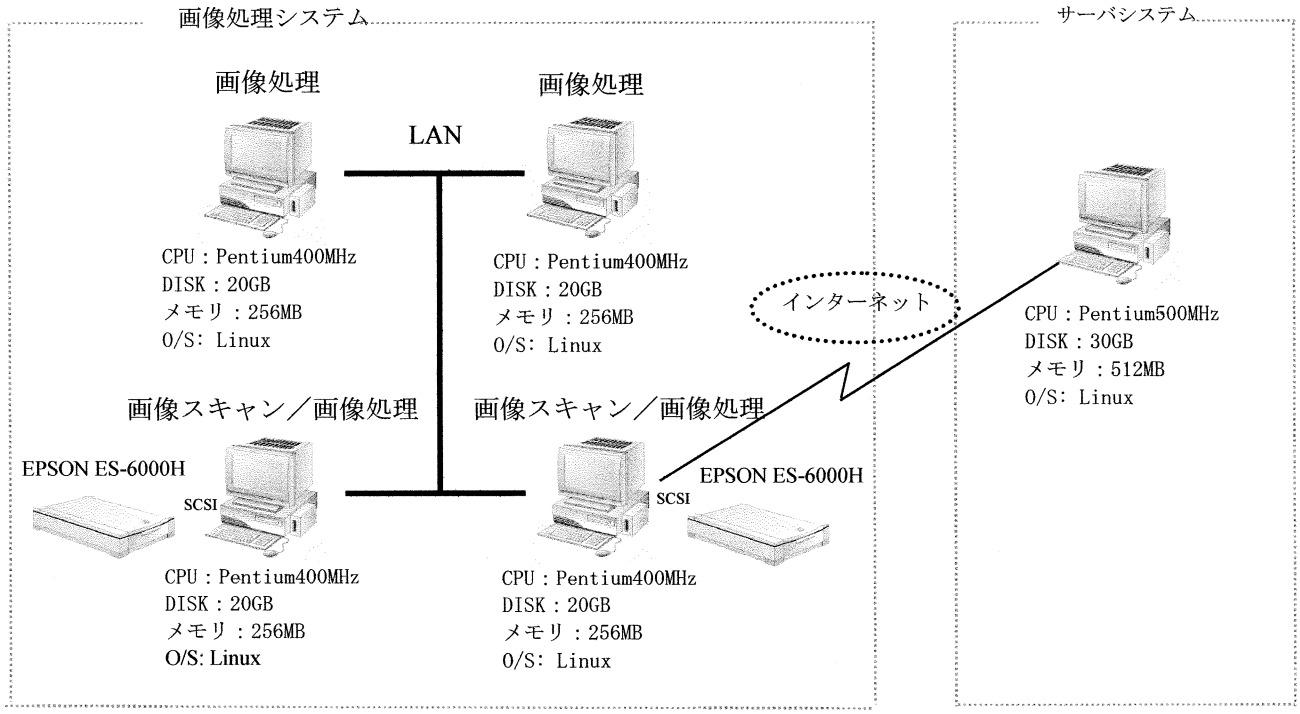


図１　システム構成図

### ３．２．２　システムの機能ブロック図

前項のシステム構成は、図２に示されるような機能ブロックに分割され、以下の機能を実現している。

#### ３．２．２．１　会話型画像処理機能

画像処理システムにおいて、画像データを読み込み、画像汚れ補正処理、画像傾き補正処理、文字部分分離処理を行い、文字部分画像データと背景画像データを生成し、これらを画像圧縮し、ASP サーバ上に格納する機能を提供し、Linux 上の画像ソフトウエアのプラグインとして実装される。

（１）画像汚れ補正機能

ユーザが入力した情報により、同一色判定の閾値以内の色の差をもつ画素を同一色であると認識し、その画素の数が最も多い色を背景色と判定し、また、汚れ判定の閾画素数より小さい連続した画素に対して、その画素の色を背景色とし、連続した画素の外接四角形を算出し、文字領域判定の閾画素数より大きい四角形は文字領域であると判断し、その領域以外の領域について画像汚れ補正処理をおこなう。

（２）画像傾き補正機能

ユーザが入力した情報により、文字列の方向に対して直角方向に補正画像データを複数に分割し、各領域ごとに濃度のヒストグラムを算出し、そのモードを結んで傾き量を算出し、傾きの補正画像傾き補正処理を行う。

（３）文字部分分離機能

ユーザが入力した情報と判定された背景色の情報により、文字部分分離処理を行う。グラデーションを含む背景も処理を可能とする。また、分離、処理された画像情報は圧縮され、さらにユーザ名、パスワードを含むＩＤ

情報が埋め込まれ、ＡＳＰサーバへ転送される。

### ３．２．２．２ HC-PDF 変換機能

ASP サーバではディレクトリを監視し、ユーザ認証、データの正当性チェックを含む確認を行い、HC-PDF 合成変換ソフトウェアにより、HC-PDF 形式のデータに変換する機能を提供する。

（１）　ディレクトリ監視機能

ASP サーバにおいて、ディレクトリを監視し、データ正当性確認し、正当であれば、変換機能を起動しする。

（２）データ正当性確認機能

ユーザ認証は､圧縮画像ファイルに含まれるユーザ名、パスワードを照合することにより行うこと。

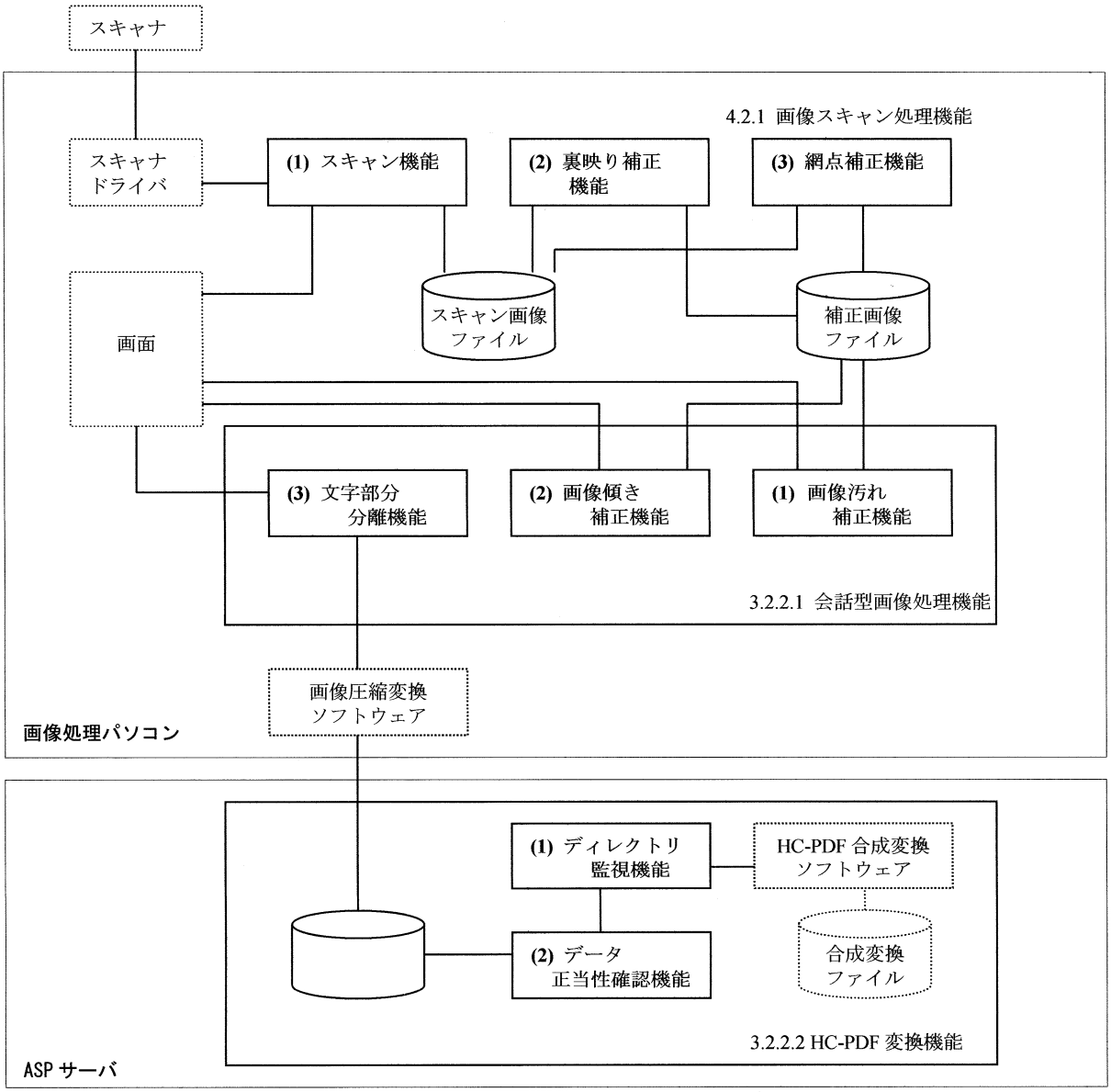


**図２　システムブロック図**

## ４．開発評価

各技術と総合評価の結果を以下に示す。

### ４．１　文字画像分離方式の開発の評価

画像データに最も多く含まれる色を背景色と判定し、背景色からある閾値以上色が異なる部分を文字部分として認識する方式の場合、グラデーションのかかったような背景色画像データまたは、閾値判定前の背景色画像の平滑化処理を行い、的確に処理することができれば、その背景色画像データから確実に文字部分画像データを抽出し分離することができる。図３，４

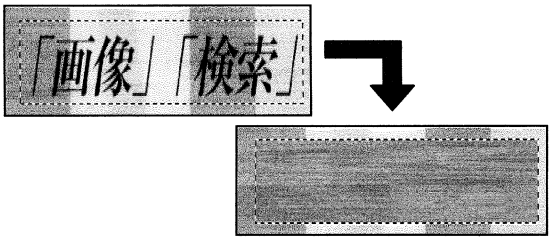


図３　グラデーションを考慮しない文字背景分離

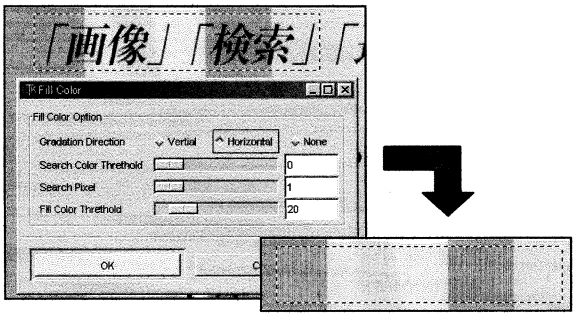


図4　グラデーションを考慮した分離

### 4．2 裏映り補正方式の開発

画像データの裏映りを判定すために、画像データを統計処理し、裏面の画像による影響を推定する方式では、充分な効果は得られなかった。おもて面画像データから裏面画像データの情報を減算する方式による裏映り補正では、完全には補正できないまでも、おもて面の品質を損なうこと無く補正することができ、効果的であると判断できた。

### 4．3 網点補正方式の開発

2次元FFTによる網点補正を行う方式は、画素に含まれる高周波数成分の除去のみで有効な補正処理が行えないと判断する。ガウシアンフィルタのフィルタを使用して網点補正を行う方式は、有効で効果的であると判断する。

### 4．4 システム全体の総合評価

システム全体の総合評価を行うために全体の作業の流れを従来の手作業による方法と、本システムを用いた作業した場合の、品質、作業時間などの評価を行った。図5この結果から本システムで一定の生産性向上をはかれたと考えられる。

### 5．まとめ

本開発におけるシステムの構築で、最も重要視していたのは、HC-PDF形式のデータをいかに、簡単に、効率的に作成することができるかであった。特に画像としてもっとも情報の欠落しやすい文字部分のデータの品質を落とすこと無く、画像部（背景画像）のデータの品質を保持し、データを作成する工程を簡素化することが目標でもあった。こうして、実際に構築されたシステムを評価してみた結果、細部の処理がまだ、完全に自動化されている訳ではないが、画像処理そのものに知識のない作業者が、効率的に処理を行えるシステムとして構成できたものと考える。また、本システムを利用し効率的にPC-PDF形式のデータを作成することで、幅広いシステム利用者と顧客の拡大を期待したいと考える。

### 6．参加企業及び機関

本プロジェクトへの参加企業及び機関は以下の通りである。

株式会社ハイパーギア

### 7．参考文献

[1] Adobe Systems Inc.：Portable Document Format Reference Manual Ver.1.3, March 11, 1999
[2] 長尾真訳：デジタル画像処理，近代科学社，(1978) Azreil Rosenfeld, Avinash Kak
[3] 田村秀行編：コンピュータ画像処理：応用実践編，総研出版，(1991)

| | 実験項目 | | 手作業 | | HC-PDF システム | |
|---|---|---|---|---|---|---|
| | | | 平均時間評価（分） | 品質評価（ｻｲｽﾞ:KB） | 平均時間評価（分） | 品質評価（ｻｲｽﾞ:KB） |
| 機能 | 画像スキャン処理機能 | 画像スキャン機能 | 0.62 | 4 | 0.6 | 4 |
| | | 裏映り補正機能 | 35 | 3 | 12 | 3 |
| | | 網点補正機能 | 15 | 3 | 7 | 4 |
| | | | | | | |
| | 会話型画像処理機能 | 画像汚れ補正機能 | 25 | 5 | 5 | 3 |
| | | 画像傾き補正機能 | 5 | 3 | 4 | 4 |
| | | 文字部分分離機能 | 35 | 3 | 15 | 4 |
| | | | | | | |
| | HC-PDF 変換機能 | ディレクトリ監視機能 | — | 4（560） | — | 4（380） |
| | | ファイル正当性確認機能 | — | — | — | — |
| | | | | | | |
| | | 合　　計 | 115.62 | | 46.6 | |

＊ 1ページに関しての測定値

図5　総合評価