

高階調古文書画像による IDP データベースの形成

村山 健二 坂本 昭二 岡田 至弘
龍谷大学 古典籍デジタルアーカイブ研究センター

国際敦煌プロジェクトは世界各国に分散している古代の中央アジアの資料を各国の機関と連携してデジタルアーカイブ化を進めている。このプロジェクトで用いるアーカイブデータは高解像度で高階調な特性を持つ高品質な画像データであり、このデータが単なる記録用のデータとしてだけでなく、資料の解析用データとして利用できるかについて検討する。

IDP database for ancient document images with high depth resolution

Kenji Murayama Shouji Sakamoto Yoshihiro Okada
Digital Archives Research Center
Ryukoku University

Ancient materials from central Asia were dispersed to worldwide. The materials are digitized by the International Dunhuang Project (IDP). The digitized image is a high quality data with high resolution and high depth. Then, for ancient document analysis, we investigate the availability of the image data.

1. はじめに

近年、文化資産をデジタル技術によって高精度に記録、保存、蓄積し、後世に伝えていくデジタルアーカイブが盛んに行われている。龍谷大学は中央アジア伝来の仏典を中心とした古文書を保管しており、資料保存と活用を目的として高解像度、高階調特性を持つ古文書画像データとして入力、蓄積、配信を行っている。本報告では、このデータがもつ高階調特性を利用し、濃淡情報と色彩情報による紙質の分類、及び墨書の解析の可能性について検討した。

2. IDP

2.1 IDP について

1900 年初頭に実施された中央アジアへの 3 回にわたる探検隊（大谷探検隊）が収集した資料のうち、古文書類（卷子本、冊子本、貝葉、木簡）のほか、印本、帛画、染織品、植物標本、古銭、拓本、考古資料などの約 9,000 点（大谷コレクション）が龍谷大学に保管されている。一方、大谷探検隊と期を同じくして、1900 年頃に世界各国の探検隊（イギリスのスタイン、ドイツのグリェンウェーデルとル・コック、フランスのペリオ、ロシアのオルデンブルグ等）が中央アジア遠征を行っており、このときに膨大な数の資料の収集が行われた。現在、これら収集資料は大谷コレクションも含め、残念ながら世界中に分散しており、すべての資料へのアクセスは困難になっている。また、資料の多さが保存機関に対して、その管理と保護財源に重圧をかけている。この問題に取り組むために、国

際敦煌プロジェクト (International Dunhuang Project: IDP) が 1994 年に設立され、その事務局がイギリスの大英図書館に置かれた。プロジェクトの目的は敦煌とその他のシルクロード遺跡から発見された 11 世紀以前の古文書等に関する研究とそれらの保存を促進することである。[1] 現在、このプロジェクトに参加している機関は、イギリス、中国、ロシア、日本そしてドイツの 5 カ国に存在し、龍谷大学が大谷コレクションを保管していることなどから IDP の日本支部として 2004 年から各国の機関と連携して活動を開始している (図.1 参照)。



図.1 IDP-Japan のホームページ
(<http://idp.afc.ryukoku.ac.jp>)

IDP は 1997 年にインターネット上に資料を公開する目的で資料のデジタル化を開始、1998 年にはウェブサイト (<http://idp.bl.uk>) を公開、そして敦煌資料のデジタル画像が登録されているデータベースも公開され、研究者や一般ユーザ

一が自由に利用できるようになっていて、デジタル画像をインターネット上に公開することによって世界中に分散している資料のデジタル画像がインターネット上に集約されることになり、資料画像に容易にアクセスできるようになっている。ここで使用されているデータベースはIDP データベースと呼ばれ、現在、IDP データベースは上述の5カ国に設置されたサーバーから構成されており、各国のサーバーが同期を取る事によってシームレスな検索を可能にしている。これまでに十数万点の画像データが登録されている。

IDP-Japan では、現在、大谷文書資料を大判カメラとデジタルカメラバック（Phase One H25、画素数：5,436×4,080 画素（2,200 万画素）、色深度：各色 16bit）を用いて撮影し、デジタル画像を Web 公開用に処理、そしてデータベース登録を行っている。IDP-Japan が登録する資料の多くは古文書の断片であり、各断片は墨を使って漢字や古代のウイグル語、ソグド語等、中央アジアで用いられた様々な言語で、麻をはじめとする各種植物繊維で漉かれた紙に書かれている。

2.2 大谷文書資料の解析

IDP-Japan では、様々な専門分野の研究者によって活動が行われており、以下に述べるようなグループによって様々な角度から研究活動が行なわれている。

まず第一のグループとして、大谷文書資料をはじめとする古文書資料に記述されている内容について研究を行うグループがある。大谷文書資料の中には制作された当時の役所等で使用された紙片や死者の埋葬時に使用された紙片等が含まれており、このような資料からは古代の中央アジアの生活、文化、経済、社会情勢など様々な事柄に関する記述が見受けられ、当時の状況が明らかにされる。例えば、当時の律令制のもとで、耕作地に関して記述されているものや、税金、労役、兵役等に関して書かれているもの、そして、詩などの文学作品までであり、その内容は多岐に渡っている。この他にも、資料の中には仏教や儒教等の様々な教典も含まれており、例えば、断片資料がどの經典のどの部分であるかを同定したり、仏教の伝来等に関する研究も進められている。また、IDP データベースによって世界各国に存在する断片資料の画像がインターネット上に集約されるので、散逸している断片の結合を期待できる。

また、大谷コレクションには漢字資料の他にも 13 種の文字と 15 の言語（インド・ヨーロッパ語族（サンスクリット、トカラ、ソグド、ホータン語）、古代チュルク語（突厥、ウイグル語）、西夏語、モンゴル語、チベット語などの諸言語、ブラーフミー文字、カローシュティー文字、ソグド文字、マニ教文字、ウイグル文字

など）が使用されており、これらを利用した言語学的な研究も進められている。例えば、漢字で記述された資料からは、文字形状の変化を調べることによって、それが書かれた年代や地域の特定をすることができるとと思われる。このような文字の変化を調べるために文字データベースの構築も行われている。これによって文字や言語の時空間的な広がりをも明らかにすることができると期待される。さらにこれに伴って、筆者の特定などに利用できる可能性もある。また、写経の中には偽物が多く含まれていることが知られており、写経の真贋判定にも応用が試みられている[2,3]。

次に第二のグループとして、科学的な分析を行っているグループがある。例えば、蛍光X線装置を使用して、非破壊で古文書に含まれている元素の定量分析を行っている。これによって、使用されている顔料等の特定が可能となる。この他にも、紙を制作する際に使用した水に含まれる成分が紙を漉く段階で繊維に附着して残留しており、このような水に含まれていた成分から紙が制作された産地や時代の推定等も可能となる[4,5]。

また、走査型電子顕微鏡による観察と撮影により、資料の微細な構造まで調べることができる。例えば、どのような繊維から紙が作られたのかが分かる。また、繊維上の墨の附着具合や紙の表面の凹凸具合まで調べることができ、新たな知見が示されてきている[6]。

最後に第三のグループとして、情報を専門とするグループがある。上述のように大谷コレクションには巻物、冊子本や木簡等、様々な形態をした資料があり、これらをどのようにアーカイブしてどのようにに活用するかについて研究している。

また、上述した科学分析による結果や書誌学的な結果とを合わせることによって、資料が制作された時代の推定や資料の真贋判定等の新たな知見が得られると期待される。将来的にこのような各分野からの様々な研究データを統合するプラットフォームとして IDP データベースは設計されている。

次節では、資料のデジタルアーカイブの一環として、IDP 基準で撮影された高階調、高解像度の画像データが単なる記録データとしてではなく、資料の解析用のデータとしてどの程度まで利用できるかについて検討する。

3. 高階調古文書画像の活用

古文書画像の入力では、標準化された照明光源のもとで画面下部のカラーチャートと共に撮影し、後 CMS のための色補正を行う。ここで得られる画像データは、RGB 各 16 bit 階調である。

これだけの解像度と階調で記録する理由として、現存の状態を記録するためにできる限り詳細な情報記録するためということがあげられる。しかし、その記録されている詳細な情報を活用するには至っていない。そこで、高階調古文書画像を対象として高階調の優位性とその特徴を示し、その適用例を示す。その際に、階調と解像度との関係についても述べる。

3.1 高階調古文書画像の色による分類

3チャンネルともに高階調である画像からは、その色がより細かな変化を捉えていると考えられる。その分布の様子を墨書の文字の部分为例として8 bit および16 bit についてそれぞれ図.2, 図.3 に示す。同じ文字の領域を同じ点数で、CIE $L^*a^*b^*$ 色空間に変換してプロットしたものである。図.2 の8 bit 階調では、量子化によるプロット位置の偏りがあることがわかる。16 bit 階調では、色空間上のプロット位置が細かくなっていることがわかる。より詳細な色を扱うには、高階調である必要がある。

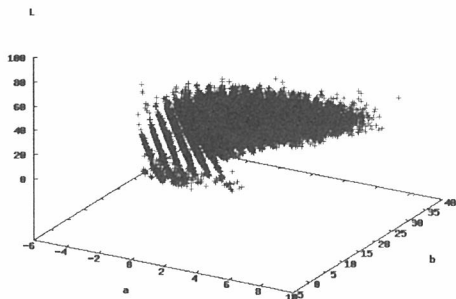


図.2 8 bit階調での分布例

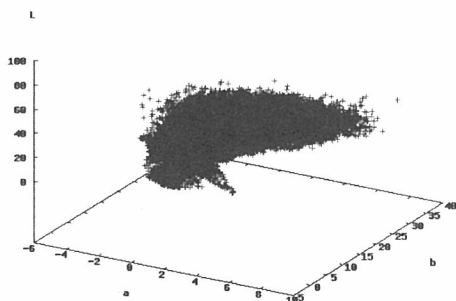


図.3 16 bit階調での分布例

ここでは CIE $L^*a^*b^*$ データを用い、微細な色差の違いを表す色空間上で色分布パラメータを導出し、分類する手法について述べる。まず、色分布パラメータを求めるために、特徴量として

- 1) 色情報である L^*, a^*, b^* ,

- 2) 注目画素の所属している領域の色分布の複雑さ
- 3) 注目画素が単一領域にあると小さくなり複数の領域にまたがっていると大きくなる値

の5つを特徴量として導出する。次に、画像平面上で大まかな領域分割は事前に行っているものとして、前述の特徴量を用いて構成される特徴空間でクラスタリングを行う。そして、各クラスを5次元の超楕円体によって近似する。これより、色分布パラメータとして、データ点数、分散共分散行列の逆行列、平均値の3つのパラメータを求める。以下に、詳細を述べる。

まず、CIE $L^*a^*b^*$ の変換を示す。AdobeRGB [8] を前提としているため、取得した RGB 値を下式により変換を行う。

$$R' = R^{2.19921875}$$

$$G' = G^{2.19921875}$$

$$B' = B^{2.19921875}$$

CIE $L^*a^*b^*$ へ変換する前に、XYZ 色空間へ変換する。

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.57667 & 0.18556 & 0.18823 \\ 0.29734 & 0.62736 & 0.07529 \\ 0.02703 & 0.07069 & 0.99134 \end{pmatrix} \begin{pmatrix} R' \\ G' \\ B' \end{pmatrix}$$

D65 を光源として、CIE $L^*a^*b^*$ への以下の変換を用いる。

$$\begin{cases} L^* = 116 \left[f\left(\frac{Y}{Y_n}\right) \right] - 16 \\ a^* = 500 \left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right) \right] \\ b^* = 200 \left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right) \right] \end{cases}$$

ただし関数 $f(X/X_n)$ は、

$$\begin{cases} f\left(\frac{X}{X_n}\right) = \left(\frac{X}{X_n}\right)^{1/3} & \left(\frac{X}{X_n} > 0.008856\right) \\ f\left(\frac{X}{X_n}\right) = 7.787\left(\frac{X}{X_n}\right) + \frac{16}{116} & \left(\frac{X}{X_n} \leq 0.008856\right) \end{cases}$$

である。 $f(Y/Y_n)$ および $f(Z/Z_n)$ も同様である。

色の複雑さ X_1 は、注目画素 (x_0, y_0) を中心とする $N \times N$ の矩形領域に対して、領域内の画素、全てについて評価値 $Val(x, y)$ を求め、その中央値をとる。 Val は以下の手順で求める。半径 r の円形フィルターを間隔 s で適用したと

きの値を、 $(Rv_r(x, y), Gv_r(x, y), Bv_r(x, y))$ とする。この値について、それぞれ差分を求める。下式は、赤(R)に対するものである。

$$Rdf_r(x, y) = |(Rv_r(x, y) - Rv_{r-s}(x, y)) / r|$$

R, G, B ごとに求めた差分値を統合して $Val_l_r(x, y)$ をもとめる。

$$Val_l_r(x, y) = \left| \frac{Rv_r(x, y) \cdot Rdf_r(x, y) + Bv_r(x, y) \cdot Bdf_r(x, y) + Gv_r(x, y) \cdot Gdf_r(x, y)}{Rv_r(x, y) + Gv_r(x, y) + Bv_r(x, y)} \right|$$

ここで、 $Val_l_r(x, y) < 1.0$ となる最小の半径 r のときの $Val_l_r(x, y)$ を画素 (x, y) での値 $Val_l(x, y)$ とする。

色分布の複雑さ X_2 は、注目画素 (x_0, y_0) を中心とする $N \times N$ の矩形領域について、領域内の画素についてそれぞれ評価値 $Val2(x, y)$ を求めて、その中央値をとる。 $Val2$ は以下の手順で求める。半径 r の円形フィルターを間隔 s で左上・右上・左下・右下の4つの領域に分けて適用する。そのときの値をそれぞれ、 $V_{1r}(x, y), \dots, V_{4r}(x, y)$ とする。これら4つの値の分散を求める。

$$D_r(x, y) = \sum_i \{V_{ir}(x, y) - \overline{V_r(x, y)}\}^2 / 4$$

R, G, B ごとに求めた分散値を統合して $Val2_r(x, y)$ をもとめる。

$$Val2_r(x, y) = \left| \frac{Rv_r(x, y) \cdot RD_r(x, y) + Bv_r(x, y) \cdot BD_r(x, y) + Gv_r(x, y) \cdot GD_r(x, y)}{Rv_r(x, y) + Gv_r(x, y) + Bv_r(x, y)} \right|$$

ここで、すべての半径 r の $Val2_r(x, y)$ について最大値を $Val2(x, y)$ とする。

ここで、 $Rv_r(x, y)$ を求める円形フィルターは、ガウス分布を用いた重みを用いた下式より得られる。

$$Rv_r(x, y) = \sum_{i=-r}^r \sum_{j=-r}^r \frac{1}{r} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{2(i^2 + j^2)}{r}\right)$$

以上、これらの5つのパラメータを領域内の点について求め、超楕円体として近似するため、サンプル数、平均、分散共分散行列の逆行列をクラス毎にもとめる。同じカテゴリに属する対象の超楕円から代表となる超楕円を生成し、そこからカテゴリとしてのパラメータも抽出する。2つの色分布の類似度は、ウイルクスの Λ 統計量を用いる。色分布パラメータから逆生成した点群から必要な統計量として、各パラメー

タの総和、各パラメータの二乗和、パラメータの積和を求めてウイルクスの Λ 統計量を得る。

$$\Lambda = \frac{|\text{グループ内の平方和積 和行列}|}{|\text{全グループの平方和積 和行列}|}$$

高階調であるため、色もより細かく分解できるようになっているが、分類の場合ではよほど詳細な変化を捉える必要がない限り、高階調は必要ではない。しかし、カラーマッチングなどで色の対応を正確に行える可能性がある。その際、標準の色の理論値がより正確でなければならないといった問題を解決する必要がある。今後、より正確な色のマッチング手法を検討する必要がある。

3.2 運筆解析のための補正フィルター

対象の全体を捉えつつ、詳細な画像情報を得るためには解像度が必要となる。解像度を高めると、古文書画像では表面の繊維のような細かな部位の特性が現れてくる。このとき、低階調では画像全体の階調を表現するために、他の明るさをもつ部分の表現に階調が割り当てられることがあり、細かな部分の表現が不十分になることがある。高階調であれば、対象全体としての階調を捉えつつ、細かな部分の変化も表現できると考えられる。

実際に、古文書画像を高解像度で撮影すると、墨書の表面における細かな反射の変化や繊維の形状が現れるようになる。墨による文字の部分では、理想的には運筆に関係する墨の濃さが変化して現れると思われる。しかし実際には、紙の影響が非常に大きく本来の墨の変化以上に現れてしまう。図.4に墨書の一部を拡大したものを示す。

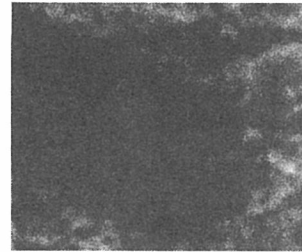


図.4 墨書の拡大図

このような墨の濃さの変化は、紙の繊維による影響が顕著に現れ、墨がまだらになったり線のような濃淡変化が生じる。このような変化を生じさせる要因となる繊維に関係する墨書表面の大まかな状態を図.5に示す。ここでは濃淡の変化を、それぞれの状態によって分類している。

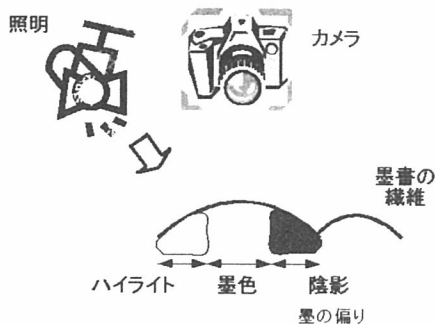


図.5 墨書の繊維による濃淡の分類

濃淡変化の要因は、繊維の形状と照明に関係する。その影響として、繊維上で反射によるハイライトが生じ、また逆にくぼんだ部分では陰影が生じる。それ以外の部分が、墨色をもった部分となる。また、繊維の形状によって墨の塗布状態のムラが生じる。このような変化は、運筆を調べる上でノイズとなり解析を困難にさせる。一般的な平滑化フィルターを適用すると、明るい部分が拡散し、また墨の濃淡構造が損なわれてしまう。そこで、墨書における濃淡構造を保存しつつ、墨のムラを抑制するフィルターを提案する。ここで、ムラの原因としてシワも大きな影響を及ぼすが今回は対象から外す。また、墨色をもった部分での色には、下地となる紙の色が含まれているが、ここでは濃淡の変化として扱う。

墨書の文字領域に対して明度ヒストグラムをとると多くの場合、図.6 のようになる。分布の多くがよく似た明るさをもった部分であり、より明度の低い部分と高い部分で構成されている。そのヒストグラムを3つに分類し、高い値をもつ部分をハイライト、低い値をもつ部分を陰影、その他を墨色とする。この分類によって、繊維の形状や照明に影響する部分を墨色部分で補正する。

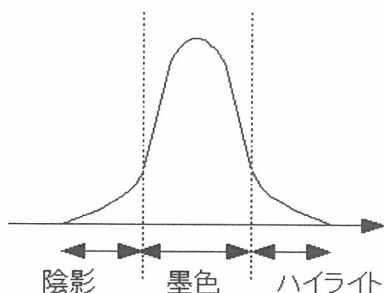


図.6 文字領域内の明度ヒストグラム

ハイライト部分や陰影部分の画素について、その点から8方向それぞれでもっとも近い墨色部

分の画素値と距離の逆数を重みとして、補正値を求める。補正値 $I(x, y)$ は下式によって求める。

$$I(x, y) = \sum_{i=0}^7 \omega_i \cdot V_i$$

ここで、 i は8つの方向を示し、 ω_i と V_i はそれぞれ方向 i における重みと画素値である。 ω_i は、下式によって求める。 r_i は、注目画素と方向 i での墨色部分までの距離を示す。

$$\omega_i = \frac{1}{r_i} / \sum_{j=0}^7 \left(\frac{1}{r_j} \right)$$

ただし、フィルターは注目画素から半径 l までの範囲で適用し、方向によって墨色領域に到達しない場合はその方向を無視する。

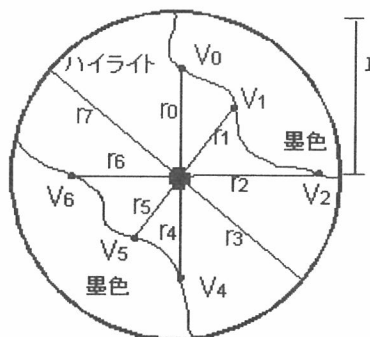


図.7 補正フィルター

この補正フィルターと高階調であることの優位性を示すために、評価実験を行った。一例として、対象画像を図.8 と図.9 に示す。筆で縦線を引いた状態を模した画像(図.8)とその上から陰影とハイライトに相当する汚れを付けた画像(図.9)を用意し、16 bit 階調と8 bit 階調で補正フィルターを適用して比較を行う。



図.8 評価用画像(16 bit) 図.9 汚れ画像

補正結果は、見た目において明るい汚れ部分が微かに残っている程度で補正されている。実際の対象ではより狭い範囲の明るさでの変化になるため、より目立たなくなると考えられる。そ

それぞれの補正結果について、評価用画像と汚れ画像の誤差に対して、補正後の誤差がどの程度残っているかを表.1に示す。

表.1 評価用画像と補正画像の比較

比較する画像の組み合わせ	誤差残存率
評価用画像と補正画像(16 bit)	1.58%
評価用画像(8 bit)と補正画像(8 bit)	8.34%

画素値は、8 bit と 16 bit を比較するために共に 0～65535 までの値をとるようにしてある。16 bit で補正した場合の誤差残存率が、8 bit と比較して低い結果が得られた。これは、補正を行った値が高階調によってより元の値に近づいていることを示している。このような微細な変化のある対象を補正する場合には、量子化誤差が軽減されている高階調に優位性があると言える。

次に、補正フィルターを実際の古文書画像に適用した例を示す。文字領域は、事前に二値画像として用意しており、その範囲内でフィルターを適用する。



図.10 対象画像



図.11 フィルター適用結果

図.10 では、墨の濃さによる違いが含まれているものの、紙の繊維の影響で白っぽくなっている部分がある。図.11 にフィルターを適用した結果を示す。ストロークによる墨の濃い部分と紙表面の繊維によって生じた濃淡変化を軽減してい

る。運筆による筆の痕跡は残っていることがわかる。

これらの結果から、高階調であることでこのような画像処理において優位な面があるといえる。低い階調では、補正した値も低い階調で割り当てられることになり、結果そのものに量子化誤差が影響することになる。高階調では、補正後に生じる中間的な値も少ない量子化誤差で対応させることができることから、このような前処理での優位性が言える。しかし、このフィルターは陰影やハイライトなどの分離に際し、ヒストグラムの分離位置に依存するため、有用な情報を残すことができる分離位置を自動的に求める必要がある。今後、補正フィルターを改良し、運筆の解析を行う。

4. おわりに

IDP の国際連携データベースで使用されている高階調・高解像度の古文書画像を活用した古文書解析の可能性について検討した。特に、色彩情報と墨書の濃淡構造に注目して、階調と解像度の関係について述べた。

今後は、高階調の特性を活かしたカラーマッピングや運筆の解析を行う。

参考文献

- [1] <http://idp.afc.ryukoku.ac.jp/>
- [2] Imre Galambos: ORTHOGRAPHY OF EARLY CHINESE WRITING: EVIDENCE FROM NEWLY EXCAVATED MANUSCRIPTS, 2006.
- [3] Harumichi ISHIZUKA: Japanese manuscripts and forgeries intermixed among Dunhuang manuscripts, Scientific Analysis, Conservation and Digitization of Central Asian Cultural Properties, Ryukoku University, pp.31-36, 2005.
- [4] Masuchika KOHNO, et al.: Trace elements in paper ---Database for classifying old manuscripts---, Scientific Analysis, Conservation and Digitization of Central Asian Cultural Properties, Ryukoku University, pp.107-114, 2005.
- [5] 加藤 雅人, 正司 哲朗, 村山 健二, 岡田 至弘, 江南 和幸, 池田 和彦, 坂田 雅之: 簀の目の測定方法の開発および応用例, 文化財保存修復学会 第 25 回大会要旨集, pp.70-71, June 2003.
- [6] Anna-Grethe RISCHEL: Analysis of fibres used for oriental papermaking compared to botanical descriptions aiming at establishing an identification key, Scientific Analysis, Conservation and Digitization of Central Asian Cultural Properties, Ryukoku University, pp.21-30, 2005.
- [7] AdobeRGB(1998) Color Imaging Encoding (<http://www.adobe.com/digitalimage/pdfs/AdobeRGB1998.pdf>)