

# 小型端末を用いた Web 閲覧のための画像分類方式

## An Image Classification Method for Web Browsing using Small Devices

前川 卓也<sup>▼</sup> 原 隆浩<sup>▲</sup> 西尾 章治郎<sup>▲</sup>

Takuya MAEKAWA Takahiro HARA  
Shojiro NISHIO

一般に、Web ページに含まれる画像は、それぞれが役割を持っており、例えば、サイトのメニューとして用いられる画像、箇条書きの行頭のアイテムとして用いられる画像などさまざまである。これらの画像の役割を正しく把握することで、Web 画像に対して適切な処理を行うことができる。そこで本研究では、その役割ごとに、11 の Web 画像のカテゴリを定義する。そして、定義したカテゴリへの画像の自動分類を実現するために、40 のサイトから収集した 3901 の画像を手動でカテゴリライズし、その結果から分類のための画像特性を抽出した。抽出した画像特性から画像の自動分類手法を提案し、実際の Web ページに含まれる画像を用いて評価を行ったところ、最高で 83.1% の分類精度を達成した。

In general, Web pages include images with various kinds of rolls, e.g., the site menu, a line head in itemization, and the title of the page. However, most conventional works for mobile Web browsing haven't given much attention to the rolls of Web images. In this paper, we define eleven Web image categories according to their rolls for a proper Web image handling. Then, we manually categorize collected 3,901 Web images from forty real Web sites, and extract image features of each category according to the classification result. By making use of the extracted features, we propose an automatic Web image classification method. Furthermore, we evaluate the proposed automatic classification method using real Web pages, where we show that up to 83.1% accuracy is achieved.

### 1. はじめに

近年、モバイル端末の小さな画面による Web 閲覧の制約を改善するために、Web ページをモバイル端末の画面サイズに合うよう再構成することを目的とした研究や製品開発が数多く行われている [1, 2, 3]。これらにおいては、ページのレイアウトなどに用いられるような画像を削除したり、大きな画像を縮小したりする処理が一般的であるが、画像の検出方法は単純であるため多くの検出誤りが発生する。例えば、サ

イト内のメニューのための画像が削除されたり縮小されたりする誤りなどは非常に深刻である。この問題を解決するために、画像の役割を把握することが有効であり、これにより画像の処理方法を適切に決定することができる。

本研究では、Web 画像の正確な取り扱いを可能とするため、また、画像の役割を利用した新しいアプリケーションの構築を支援するため、11 の Web 画像のカテゴリを定義する。そして、様々な種類の 40 の Web サイトから収集した 3901 の画像を、それらの 11 のカテゴリに手動で分類する。手動分類の結果から、ページ内の画像を効果的に分類するための 37 の画像特性を定義し、それらの特性から作成した決定木を用いて画像の自動分類を実現する。

ここで画像から抽出可能な特性には、HTML ソースの解析から抽出できるもの、Web サーバに通信で問い合わせ取得できるもの、ページを実際にレンダリングした際に DOM 木のレイアウト情報から取得できるもの、画像処理から取得できるものがある。モバイル環境では、端末の性能や通信環境から取得できる特性はさまざまである。また、端末を取り巻く環境によっても取得できる特性は異なる。例えば、プロキシサーバを利用できる環境なら、全ての特性を取得することも不可能ではない。さらに、アプリケーションによっては 11 のカテゴリへの分類を求めないものや、特定のカテゴリへの分類に対する精度のみを要求するものもあり、全ての特性を利用する必要がないこともある。そこで本研究では、利用できる特性ごとに 11 のカテゴリへの分類精度を評価した。そして、11 のカテゴリに対して最高で 83.1% の分類精度を達成した。

### 2. 画像のカテゴリ

本章では、定義した 11 のカテゴリについて説明する。また、収集した画像群に見られたカテゴリごとの特徴についても簡単に説明する。ここで、以降の理解を容易にするために、主に文字列のみを含む画像が属する 4 つのカテゴリをまとめて文字列画像と呼び、小さい画像が属する 2 つのカテゴリをまとめて小画像と呼ぶ。以下の説明は、本研究で収集した 40 の Web サイトに含まれる画像を例として用いる。

- 文字列画像：以下の四つのカテゴリに分類される。
  - MENU: サイト内のメニュー画像を含む。図 1(a)における“HOME”、“SCHEDULE”などの画像や、図 1(b)における“THE COLLECTION”、“EVENTS & PROGRAMS”がこれに当たる。ページの上部や下部に横並びで配置されることが多い。ページの左部に縦並びで配置されることも多い。収集した 3901 の画像を調査したところ、MENU に属する画像のうち、67.6% の画像において 3 つ以上横に並び同じ高さの画像があった。11.5% の画像において 3 つ以上縦に並び同じ幅の画像があった。
  - SECTION: ページ内のセクションの表題画像を含む。図 1(c)における、“U.S.”の画像がこれに当たる。この画像の下に文字列が配置されていることが多い。実際、92.8% の画像の下に文字列が配置されていた。アスペクト比が小さいことが多く、平均で 0.142 であった。
  - DECORATION: 装飾文字列の画像を含む。図 1(d), (e) の画像がこれに当たる。HTML 言語では表現できないような文字列を利用するとき用いる。ハイパーリンクをもたない。
  - BUTTON: ハイパーリンクをもつ画像を含む。図 1(f),

<sup>▼</sup> 正会員 日本電信電話株式会社 コミュニケーション科学基礎研究所 [maekawa@cslab.kecl.ntt.co.jp](mailto:maekawa@cslab.kecl.ntt.co.jp)

<sup>▲</sup> 正会員 大阪大学大学院情報科学研究科 [fhara.nishio@ist.osaka-u.ac.jp](mailto:fhara.nishio@ist.osaka-u.ac.jp)

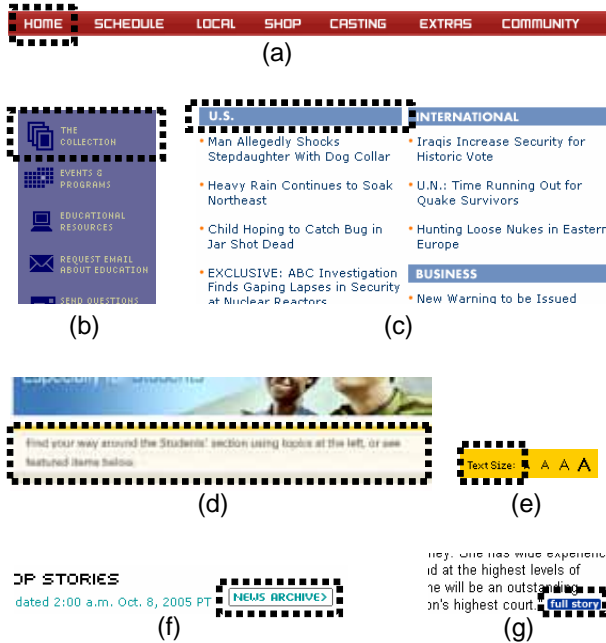


図1 文字列画像

Fig.1 String Images

(g)の画像がこれに当たる。文字列が周辺に配置されていることが多く、周辺の文字列に関連するページへのハイパーリンクとなっている。BUTTONに属する画像のうち、16.1%が上に文字列をもち、8.0%が下に文字列をもち、36.8%が左に文字列をもち、13.8%が右に文字列をもち、25.3%が文字列を周辺にもたなかった。

- 小画像：以下の二つのカテゴリに分類される。
  - ITEM: 箇条書きにおける行頭のアイテムの画像を含む。図2(a)の画像がこれに当たる。同じ幅の画像が縦並びで配置されることが多く、74.6%が3つ以上縦に並び同じ幅の画像をもっていた。また、文字列が右に配置されていることが多く、99.4%がそうであった。それ以外は、画像が右に配置されていた。周辺に配置されている文字列の平均は31.7文字であった。アスペクト比が1に近いことが多く、その平均は1.052であった。
  - ICON: あるものを表現する画像を含む。図2(b),(c)の画像がこれに当たる。文字列が右や左に配置されていることが多く、58.3%が右に文字列をもち、22.0%が左に文字列をもっていた。アスペクト比が1に近いことが多く、その平均は0.942であった。
- TITLE: ページのタイトル画像を含む。図3(a)の画像がこれに当たる。ページの上部に配置され、サイトのインデックスページや、そのページ自身へのハイパーリンクであることが多い。
- MAP: マップメニューの画像を含む。図3(b)の画像がこれに当たる。サイトのメニューとして用いられることが多い。
- AD: パナール広告の画像を含む。図3(c)の画像がこれに当たる。外部サイトへのハイパーリンクであることが多く、25.5%がそうであった。また、アスペクト比が小さいことが多く、その平均は0.459であった。少量の文字列が下に配置されていることがある。14.0%が下に文字列をもち、78.7%が周辺に文字列をもたなかった。下に配置する文字

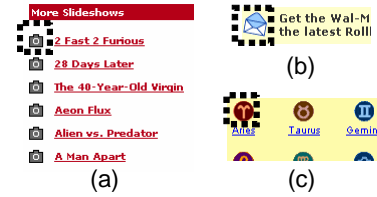


図2 小画像

Fig.2 Small Images



図3 TITLE, MAP, AD, CONTENT, LAYOUTER

Fig.3 TITLE, MAP, AD, CONTENT, LAYOUTER

列の平均は45.2文字であった。

- CONTENT: Webページ内のメインのコンテンツに関連付けられている画像を含む。図3(d)の画像がこれに当たる。アスペクト比が1に近いことが多く、その平均は0.951であった。文字列が右や下に配置されていることが多く、35.1%の画像が右に文字列をもち、51.7%の画像が下に文字列をもっていた。周辺文字列の平均は99.7文字であった。JPEG形式であることが多く、55.4%がそうであった。一方、CONTENT以外に属する画像では6.6%がJPEG形式であった。
- LAYOUTER: ページのデザインやレイアウトを調整するための画像を含む。空白や単一色の画像であることが多く、ページ内の画像や文字列の表示位置を調整するために用いられることが多い。1つの画像がページ内で複数回利用されることが多く、平均で10.7回用いられていた。図3(e)のように、他の重要な画像のレイアウトを調整するために用いられることもある。

### 3. データセットの収集と検証

#### 3.1 画像の収集と分類

まず、ポータルサイト、オンラインショッピングサイト、ニュースサイト、製造会社のサイトなど、ランダムに選んだ40のWebサイト(20の米国サイトと20の日本サイト)において、インデックスページおよびその他の2つのページから画像を収集した。2つのページはインデックスページから互いにリンクしており、そのサイト内で重要と思われるものを選んだ。そして、40のサイトから収集した計3901の画像を、

2章において定義した11のカテゴリに手動で分類した。

### 3.2 自動分類に用いる画像の特徴

画像の自動分類を実現するため、抽出する画像の特徴を以下に示す。特性F1~F20はHTMLソースの解析により取得できる値である。なお、幅や高さなどの値がソースのIMGタグに明示されていないときは欠損値となる。F21, F22は画像をダウンロードする際や、WebサーバにHEADメソッドを用いてリクエストする際に取得できる値である。F23~F30は、Webページを受信してレンダリングする際に、DOM木などから取得できる値である。本研究ではMSHTML DOM木を用いてこの値を取得した。F31~F37は、画像処理を用いて取得できる値である。

- **F1-面積**
- **F2-幅**
- **F3-高さ**
- **F4-アスペクト比**
- **F5-マップを利用しているか否か**{TRUE, FALSE}: MAPに属する画像がTRUEとなる。
- **F6-ハイパーリンクか否か**{TRUE, FALSE}: SPACERに属する画像の多くや、DECORATEに属する画像はFALSEとなる。
- **F7-アウトリンクか否か**{TRUE, FALSE}: 他のドメインへのリンクか否かを示す。
- **F8-ループバックリンクか否か**{TRUE, FALSE}: サイトのインデックスページや、その画像が含まれるページ自身へのハイパーリンクか否かを示す。
- **F9-ALTの有無**{TRUE, FALSE}: 文字列画像に属する画像などの多くはTRUEとなる。
- **F10-ALTの文字数**: CONTENTに属する画像はこの値が大きく、ALTをもつものの平均は26.8文字であった。MENU, SECTION, DECORATION, BUTTON, ICON, ITEM, TITLE, ADの平均はそれぞれ8.5, 11.3, 19.6, 9.18, 3.8, 9.9, 19.0, 19.7文字であった。
- **F11-周辺文字列の文字数**: 画像の周辺に配置されている文字列の文字数を示す。MENUに属する画像はこの値が小さく、その平均は2.7文字であった。一方、全ての画像における平均は69.8文字であった。
- **F12-JPEGか否か**{TRUE, FALSE}: CONTENT以外に属する画像はほとんどがFALSEとなる。
- **F13-HTMLソースにおけるインデックス**: ソースにおいてその画像のIMGタグが現れるインデックスを示す。TITLEはこの値が小さいことが多く、その平均は48.4であった。一方、全ての画像における平均は424.7であった。
- **F14-同じページ内で画像の現れる回数**
- **F15-同じページ内の面積が同じ画像の数**: CONTENT, ICON, ITEMに属する画像はこの値が大きいために多く、その平均はそれぞれ7.5, 4.3, 4.0であった。
- **F16-同じページ内の幅が同じ画像の数**: CONTENT, AD, ICON, ITEM, SECTIONに属する画像はこの値が大きいために多く、その平均はそれぞれ8.1, 3.5, 4.3, 4.5, 4.4であった。AD画像は、縦揃えでページの端に並べられていることが多いため、この値が大きくなった。
- **F17-同じページ内の高さが同じ画像の数**: CONTENT, MENU, SECTION, ICON, ITEMに属する画像はこの値が大きいために多く、その平均は、8.1, 8.5, 4.8, 4.4, 4.8であった。
- **F18-同じページ内の近くに位置する面積が同じ画像の数**

- **F19-同じページ内の近くに位置する幅が同じ画像の数**
- **F20-同じページ内の近くに位置する高さが同じ画像の数**
- **F21-バイトサイズ**
- **F22-面積あたりのバイトサイズ**: CONTENTやAD画像などの複雑な画像はこの値が大きいために多く、その平均は0.83, 0.71byte/pix<sup>2</sup>であった。また、小さい画像はこの値が大きいために多い。ICON, ITEM, LAYOUTER画像は、その平均がそれぞれ1.2, 1.0, 8.9であった。これは、画像ファイルのヘッダによる影響であると考えられる。
- **F23-画像の左上のX座標**: TITLEに属する画像はこの値が小さいことが多く、その平均は46.2であった。一方、全画像の平均は314.1であった。
- **F24-画像の左上のY座標**: MENUやTITLEに属する画像はこの値が小さいことが多く、その平均は216.3, 20.0であった。一方、全画像の平均は603.4であった。
- **F25-同じページ内のF23が同じ画像の数**: ページにおいて特定のURLをもつ画像が複数回使われている場合、それぞれを1つの画像としてカウントする。SECTION, ITEM, TITLE, CONTENT, AD, LAYOUTERに属する画像はこの値が大きいために多く、その平均は13.8, 15.4, 6.1, 6.2, 6.8, 8.4であった。LAYOUTERが大きいために、縦に並べられている画像や文字列間の空白を作るために用いられることが多いからである。
- **F26-同じページ内のF24が同じ画像の数**: ページにおいて特定のURLをもつ画像が複数回使われている場合、それぞれを1つの画像としてカウントする。MENUに属する画像はこの値が大きいために多く、平均は5.8であった。一方、全ての画像の平均は2.0であった。
- **F27-同じページ内の幅とF23が同じ画像の数**: ページにおいて同じ画像が複数回使われている場合、それぞれを1つの画像としてカウントする。SECTION, ITEM, CONTENT, ADに属する画像はこの値が大きいために多く、その平均は4.4, 11.2, 3.1, 2.9であった。TITLEとLAYOUTER画像はF25が大きかったが、この値は小さくなった。
- **F28-同じページ内の高さ**と**F24が同じ画像の数**: ページにおいて同じ画像が複数回使われている場合、それぞれを1つの画像としてカウントする。MENUに属する画像はこの値が大きいために多く、平均は5.6であった。一方、全ての画像の平均は1.6であった。
- **F29-ページの下部から画像の下部までの距離**: MENUに属する画像の一部はこの値が小さいことが多い。
- **F30-周辺文字列の位置**{UP, DOWN, LEFT, RIGHT, NONE}
- **F31-使用色数**
- **F32-同色領域数**
- **F33-近くにある画像との類似度**
- **F34-アニメーションGIFか否か**{TRUE, FALSE}: ADはアニメーションGIFであることが多く、実際ADに属する画像のうち、14.29%の画像がアニメーションGIFとなっていた。一方、AD以外に属する画像のうち0.36%がアニメーションGIFであった。
- **F35-角が丸まった矩形か否か**{TRUE, FALSE}
- **F36-文字列領域の占有率**: 画像内の文字列を検出することで、文字列画像を区別することができる。例えば、LAYOUTER画像はさまざまな形で至るところに存在するため、SECTIONやDECORATION, TITLE画像と混同しやすい。また、AD画像は複数の位置に文字列をもつ

表1 分類精度  
Table 1 Classification Accuracy

| C1    | C2    | C3    | C4    | C5    |
|-------|-------|-------|-------|-------|
| 0.749 | 0.768 | 0.796 | 0.766 | 0.831 |

ことが多い。例えば、図3(c)では、複数の領域に異なるフォントの文字列が描かれている。一方で、MENU や SECTION は1つの領域に文字列が描かれていることが多い。例えば、図1(a)では、1つの領域に同じフォントの文字列をもつ。

#### ● F37-文字列領域の数

### 4. 評価実験

本章の評価実験では、トレーニングセットの画像を11のカテゴリに分類する決定木を3章で示した画像特性から作成し、作成した決定木を用いてテストセットの分類を行った。決定木の作成にはC4.5 [4]を用いた。40のサイトから収集した画像から、1つのサイトの画像をテストセットとし、残り39のサイトの画像をトレーニングセットとするような評価実験を、全てのサイトがそれぞれテストセットとなるよう40回行った。ここで、利用する画像特性を以下の条件のように変化させて実験を行った。

- C1: HTML ソースの解析により取得できる特性を用いる。つまり、F1~F20を用いる。
- C2: HTML ソースの解析により取得できる特性および Web サーバから取得できる特性を用いる。つまり、F1~F22を用いる。
- C3: HTML ソースの解析により取得できる特性、Web サーバから取得できる特性、および、DOM から取得できる特性を用いる。つまり、F1~F30を用いる。ただし、DOM から Web サーバから取得できる特性を取得できるため、実際には Web サーバにアクセスする必要はない。
- C4: HTML ソースの解析により取得できる特性、Web サーバから取得できる特性、および、画像処理により取得できる特性を用いる。つまり、F1~F22, F31~F37を用いる。画像処理を行うためには、画像をサーバからダウンロードする必要があるため、Web サーバから画像のサイズを個別に取得する必要はない。
- C5: 全ての特性を用いる。つまり、F1~F37を用いる。

C1 から C5 における分類精度を表1に示す。ソースの解析による特性のみを用いたときでも、75%もの画像が正しく分類されている。また、基本的に利用する属性が増えるほど、その精度は増加するが、C4はC2に比べて微小ではあるが逆に精度が低下している。C4に比べてDOM木から得た特性を利用するC5では、その精度が0.831と大幅に向上している。一方、C5はC3に対して画像処理から得た特性を追加しており、C3に比べても大幅に精度が向上している。これにより、画像処理はDOM木と組み合わせることで、その効果を発揮することが分かる。

### 5. まとめ

モバイル端末による Web ページ閲覧において、Web ページ内の画像の役割を適切に把握することにより、小さいスクリーンサイズや貧弱な入力インタフェースによる制約を克服

するさまざまなアプリケーションを提供できるものと考えられる。そこで、本研究では Web ページに含まれる画像の11のカテゴリを役割ごとに定義した。そして、定義したカテゴリへの自動分類を実現するために、40のサイトから収集した3901の画像を手動で分類し、その結果から画像の37の特性を抽出した。それらの特性を用いて、最高で83.1%の分類精度を実現した。

#### [謝辞]

本研究の一部は、文部科学省21世紀COEプログラム「ネットワーク共生環境を築く情報技術の創出」の研究助成によるものである。ここに記して謝意を表す。

#### [文献]

- [1] T. W. Bickmore and B. N. Schilit, "Digester: device-independent access to the world wide web," Proc. World Wide Web Conference (WWW6), pp. 655-663, April 1997.
- [2] W. Y. Ma, I. Bedner, G. Chang, A. Kuchinsky, and H. J. Zhang, "A framework for adaptive content delivery in heterogeneous network environments," Proc. SPIE Multimedia Computing and Networking 2000, pp. 86-100, Jan. 2000.
- [3] OPERA Software, "Opera for mobile," <http://www.opera.com/products/mobile/>.
- [4] WEKA Machine Learning Project, "WEKA 3," <http://www.cs.waikato.ac.nz/~ml/weka/>.

#### 前川 卓也 Takuya MAEKAWA

2003年大阪大学大学院工学部情報システム工学科卒業。2006年同大学院情報科学研究科博士後期課程修了。情報科学博士。現在、日本電信電話株式会社コミュニケーション科学基礎研究所勤務。日本データベース学会、情報処理学会の各会員。

#### 原 隆浩 Takahiro HARA

1995年大阪大学工学部情報システム工学科卒業。1997年同大学院工学研究科博士前期課程修了。同年、同大学院工学研究科博士後期課程中退後、同大学院工学研究科情報システム工学専攻助手、2002年より同大学院情報科学研究科マルチメディア工学専攻助手、2004年より同大学院情報科学研究科マルチメディア工学専攻助教授となり、現在に至る。工学博士。IEEE、電子情報通信学会、日本データベース学会の各会員。

#### 西尾 章治郎 Shojiro NISHIO

1980年京都大学大学院工学研究科博士後期課程修了。工学博士。京都大学工学部助手、大阪大学基礎工学部および情報処理教育センター助教授、大阪大学大学院工学研究科教授を経て、2002年より同大学大学院情報科学研究科教授となり、現在に至る。2000年より大阪大学サイバーメディアセンター長、2003年より大阪大学大学院情報科学研究科長を併任。データベース、マルチメディアシステムの研究に従事。現在、Data & Knowledge Engineering等の論文誌編集委員、本学会理事、電子情報通信学会、情報処理学会の各フェローを含め、ACM、IEEEなど8学会の会員。