

表層的な情報を用いた画像の内容特定と分類

嶋田 和孝[†] 伊藤 哲郎[†] 遠藤 勉^{††}

[†]大分大学工学部知能情報システム工学科
〒 870-1192 大分県大分市旦野原 700 番地
E-mail: {shimada, ito}@csis.oita-u.ac.jp

^{††}九州工業大学情報工学部知能情報工学科
〒 820-8502 福岡県飯塚市大字川津 680-4
E-mail: endo@pluto.ai.kyutech.ac.jp

あらまし WWW 上には多くの画像データが存在するが、それら画像データは多くの情報が欠落しており、従来の画像解析手法による内容特定処理では十分な精度は望めない。本論文では、WWW 上の画像の内容を隣接する文章から特定、分類する手法について提案する。まず、画像サイズからの対象画像抽出処理について述べ、隣接する文章へのキーワードマッチングによる画像名特定処理および分類処理について述べる。最後に実験を通じて本手法の有効性を検証する。

キーワード 表層表現 内容特定 分類 WWW

Image Identification and Classification using Surface Expressions

Kazutaka SHIMADA[†] Tetsuro ITO[†] Tsutomu ENDO^{††}

[†] Department of Computer Science and Intelligent Systems, Oita University
700 Dannoharu Oita, 870-1192 Japan
E-mail: {shimada, ito}@csis.oita-u.ac.jp

^{††} Department of Artificial Intelligence Faculty of Computer Science and Systems Engineering,
Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
E-mail: endo@pluto.ai.kyutech.ac.jp

Abstract There are many images on WWW. Traditional image analysis, however, can not estimate the contents of images because they lost a lot of information. In this paper, we proposed a method for identification and classification of images using surface expressions. First, our method extracts images and relational sentences from a homepage on WWW. It then identifies and classifies the contents of images. Experimental results show the effectiveness of our method.

key words Surface expression, Identification, Classification, WWW

1. はじめに

近年の急速なインターネットの普及により、職場や家庭にしながら世界中から発信された情報にアクセスできる環境が整ってきた。これに伴い、今までは紙面で伝えられていた内容が電子化されてきている。ネットワークから入手できる情報はテキストが中心だが、WWW を考えると表や画像も多く含まれている。しかし、WWW 上に存在する画像データは表示を目的としたものであり、画像の解像度が低く多くの情報が欠落しており、従来のような画像処理に適しているとは言い難い。一方で、画像と隣接するテキストの内容には密接な関係がある。

テキストと画像との関係特定に関する研究では、新聞記事中の画像と本文中に存在する人名の対応付け[1]や TV ニュースと新聞記事の対応付け[2]、パターン情報と言語情報を用いた植物図鑑の図の理解[3]、テキストを用いた画像解析[4]などがある。しかし[1]では、画像が人物だけに限られている。また、[2][3][4]では、画像側からも何らかの内容情報が必要となる。

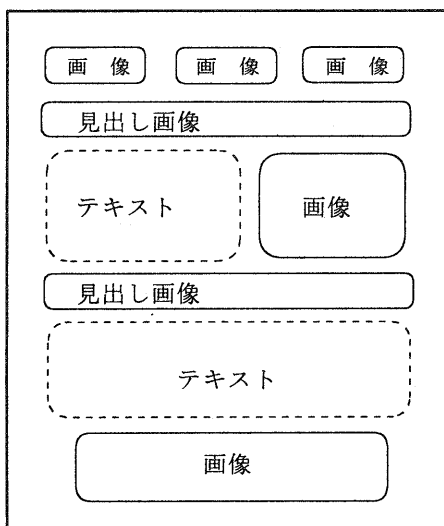


図1 製品紹介のホームページ

筆者らは現在、製品の性能などを記述した表からの特徴データの抽出[5]、要約処理[6]と画像とを統合したマルチメディア要約処理システムの構築を進めている。本稿では、企業の製品紹介に関するホームページ中に存在する画像内容を隣接するテキスト中に存在するキーワードなどの表層的な手掛かりを用いて推定する手法を提案する(図2)。そして提案された手法の有効性を示す。

2. 対象画像と文章の抽出

2.1 対象画像の抽出

製品紹介のホームページには、製品やそれに関する画像の他にも、見出しとして使われている文字画像やリンク用のボタン画像など、内容推定の必要のない画像も多く存在する。そこで、まずホームページ中の画像から内容推定を行うべき画像の抽出方法について述べる。対象画像の抽出には、画像のサイズ情報を用いる。対象画像と非対象画像を分類するための条件は、337 個の画像から(形式は GIF と JPEG)人手で抽出した。対象画像抽出処理で用いる条件は以下の3つである。

- (1) 画像サイズが 55×55 (pixels)以上。
- (2) 縦横比が 3.8 倍以下。

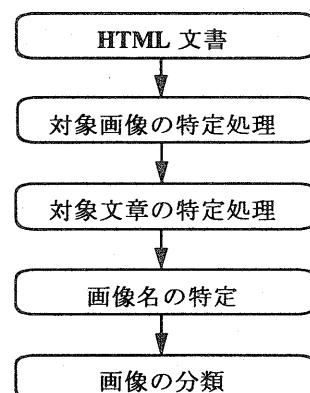


図2 処理の流れ

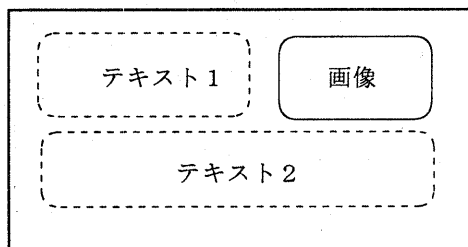
(3) 隣り合う画像の高さが同じなら，1つの画像としてあつかう。

高さの等しい隣り合う画像を1つの画像として扱うのは，WWWの性質上，1つの画像をあまり大きなサイズにせず，分割して保存している場合があるためである。以上の条件を用いて対象画像を抽出する。

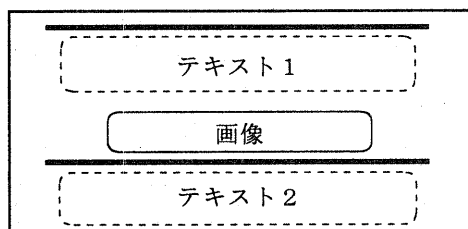
2.2 文章の抽出

次に隣接する文章の抽出処理について述べる。どの文章が隣接しているかはHTMLのタグから判定する。以下に処理の手順を示す。

- (1)HTML文書から対象画像名(タグ)部分を抽出。
- (2)対象画像前後のタグ・文章を抽出。
- (3)タグにALTデータがあれば，そのALTデータの内容を含む隣接テキストを対象文章として抽出。
- (4)タグからレイアウト情報を抽出。
- (5)抽出されたデータから最も画像に近く，



(a)



(b)

図3 レイアウトとそのHTML

優先順位の高いタグでレイアウトされたテキストを対象文章として抽出。

現在使用しているタグは<TABLE>，
，<HR>などである。優先順位は最も使用頻度の高い<TABLE>タグが最も高い。図3にレイアウトとHTML表記の例を示す。図3では，共に「テキスト1」が画像に関連のある文章だとして抽出される。

3. 画像名の特定と分類

WWW上の画像は表示が目的であるため，多くの情報が欠落しており，従来の画像解析の手法では，十分な精度が得られるとは言い難い。本章では，隣接するテキスト中に存在するキーワードとHTMLのタグの情報などの表層的な手掛かりを用いて画像内容を特定する手法について述べる。

3.1 画像名の特定

まず，表層の手掛かりとタグを用いて画像名を特定する。特定処理では，初めに位置情報から紹介されている製品の画像を特定する。前章で抽出された対象画像のうち，位置

```
<TABLE>
<TR>
  <TD>テキスト1</TD>
  <TD><IMG src="画像"></TD>
</TR>
<TR>
  <TD colspan="2">テキスト2</TD>
</TR>
</TABLE>
```

```
<HR>
テキスト1<BR>
<IMG src="画像">
<HR>
テキスト2<BR>
```

的に最も上部に存在する画像をそのホームページで紹介されている製品の画像とする。これは製品紹介の場合、必ず上部にパソコン画像が存在するためである。同列に複数画像がある場合には、より画像サイズの大きい方をパソコン画像とする。

続いて、その他の画像の画像名を特定する。使用する表層的手掛かりはテンプレートと括弧表現、位置情報、タグはに付属するALT 情報である。以下にその手順を示す。

- (1) テンプレートを用いて対象文章の第1文目を解析する。

テンプレートの例：[画像名]を装備。

- (2) 対象文章全文から括弧表現を抽出する。使用する括弧表現は「」と『』である。括弧表現が文章中に複数存在する場合、出現がはやい括弧表現の内容を画像名とする。

- (3) タグのALT 情報を抽出する。

ALT 情報が存在すれば、その内容を解析する。

- (3-1) ALT 情報中に助詞や句点、読点などが存在するか解析し、存在すればそのALT 情報を棄却する。存在しなければ、それを画像名とする。

それぞれの手順で複数画像名が抽出された場合、以下の優先度で固有名を特定する。

優先度 (1)>(2)>(3)

3.2 画像の分類

次に画像の内容が何であるかをキーワードやテンプレートなどを用いて分類する。キーワードはあらかじめ人手により抽出している。キーワードには重みが付けられており、分類される項目に大きく依存する語は重みを大きく、比較的多く文章に出現する語は小さくしている。この重みはキーワードを抽出お

よび分類する際に著者の主観で決定した。また、複数の分類項目にキーワードとして定義されている語も存在する。画像が分類される項目は全てで11種類存在する。表1に分類項目を示す。これらの分類項目は、表からの要約文生成処理[6]で用いられる話題にもあたり、文章はこの分類項目に合わせて生成される。また、これらの分類項目は、パソコンの製品紹介に合わせて作成したものであり、どの項目にも分類されるべきではない画像も抽出された対象画像中には存在する。例えば、パソコンの製品紹介記事で希に存在する周辺機器のプリンタやスキャナの画像などがこれにあたる。

分類処理はまず第1文目のキーワードマッチングとテンプレートマッチングを行う。ここで使用するテンプレートは画像名特定処理で使用したものと同じである。キーワードがテンプレートの「を格」とマッチングした場合、そのキーワードが分類されている項目を

表1 分類項目とその内容の例

分類項目	分類される内容
基本性能	CD-ROM や CD-R/RW CPU や DVD-ROM
拡張性	拡張ベイや USB
画像処理	IEEE 端子や画像編集ソフト
表示能力	液晶ディスプレイ
操作性	特殊ボタン付きキーボード ポインティングデバイス
サイズ	他モデルとの比較
通信	携帯・PHSとの接続機器
音楽関連	オーディオ端子 音楽関連ソフト
情報	ユーザ登録やサービス
他モデル	関連モデル
ソフト	分類されなかった 添付ソフト

表2 タグによる重みのパラメータ

	n
文字色の変化	1.5
文字の強調	1.5
文字の拡大	1.75
文字の縮小	0.75
文字の拡大 + {色 or 強調}	2.0

画像の分類項目として決定する。また、テンプレートの「を格」とマッチングしない場合でも、第1文でマッチしたキーワードはそれ以降の文でマッチングした場合の2倍の得点を加算する。また、フォントサイズの変更や文字の強調などのタグに囲まれたキーワードがあった場合は、そのタグに合わせて得点が n 倍される。タグごとの n の値の一覧を表2に示す。第1文のテンプレートとキーワードのマッチング終了後、それ以降の文章もキーワードのマッチングを行う。タグが存在すれば、同様に得点を n 倍する。全ての文章とマッチングを終えた後、総得点の最も多い項目をその画像の分類項目とする。総得点が同値の場合、第1文でマッチしたキーワードの多い方をその画像の分類項目とする。

4. 実験・考察

2. および3. で提案したアルゴリズムを基に実際のホームページ上の画像と文章を用

いて実験を行った。実験対象はパソコンに関する製品紹介のホームページである。以下、その結果を示し、考察する。

まず、対象画像の抽出処理についてである。条件作成および修正のために337画像（訓練データ）、テストデータとして450画像を用意した。その結果を表3に示す。

対象画像抽出処理は、画像内容を特定するための前処理であり、できるだけ多くの正解画像を取り出せる方が望ましい。訓練データ、テストデータの双方とも97～98%の高い再現率を得ており、条件はサイズの比較だけとシンプルであるが、本手法の有効性が確認された。

続いて、画像名の特定および分類処理について述べる。画像名の特定および分類処理では、対象画像抽出処理で抽出された画像のうち、前後に文章がある182画像を用いて実験を行った。それらの画像とテキストをランダムに2つにわけ、76個をキーワードとテンプレートの抽出用に、106個をテストデータとした。

まず、画像名特定処理について考察する。画像名特定処理の結果を表4に示す。失敗し

表4 画像名特定処理の実験結果

	訓練データ	テストデータ
データ数	76 個	106 個
正解数	57 個	72 個
精 度	75.0 %	67.9 %

表3 対象画像抽出処理の実験結果

	訓練データ	テストデータ
正解画像数	154 画像	211 画像
抽出された画像数	174 画像	218 画像
抽出された画像中の正解画像数	153 画像	205 画像
再現率	98.0 %	97.2 %
精 度	91.5 %	94.0 %

表5 画像の分類処理の実験結果

	訓練データ	テストデータ(修正前)	テストデータ(修正後)
データ数	76 個	106 個	106 個
正解数	59 個	71 個	83 個
精 度	77.6 %	67.0 %	78.3 %

たもののうち、マッチするデータがなく特定できなかったものは訓練データで 15 個、テストデータで 11 個、誤った特定をしたものは訓練データで 4 個、テストデータで 23 個であった。誤った特定をした原因は、実際には文章中に画像名を指す語はないが、関係のない「」や ALT 情報とマッチングしたものや正しい画像名が文章の後の方に出現し、前にある違う語にはじかれてしまうものなどである。優先順位やテンプレートの追加などが今後の課題として挙げられる。

次に画像の分類処理についてである。その結果を表5に示す。キーワードの総数は 165 個である。テストデータを用いた実験後、キーワード不足のため失敗したものを修正するためにさらに 19 個加えた。その結果、訓練データとテストデータの精度がほぼ同等になったことから、この値が本手法の精度といえ、本手法の有効性を確認できた。また、範囲が広いとキーワードの定義が難しい[分類項目:ソフト]に対応づけられるデータを対象外とした場合、訓練データで 84.4 %、テストデータで 87.3 %という高い精度を得ることができた。

5. おわりに

本稿では、製品紹介のホームページ中の画像名を隣接する文章のキーワードマッチングで特定し、その内容を分類する手法について提案した。画像自体を解析せず、単純なパターンマッチのみで比較的高い精度を得ることができ、本手法の有効性が確認された。

今後の課題としては、表の特徴抽出・要約処理と本手法を統合したマルチメディア要約処理への拡張、本手法の精度向上のためのキーワードの追加やキーワードの重みの再考察などが挙げられる。

参考文献

- [1] 山田剛一, 杉山一成, 中川裕志, “新聞記事における写真と言語表現の対応の学習”, 信学技報, NLC97-63, pp.65-70, 1998.
- [2] 渡辺靖彦, 岡田至弘, 角田達彦, 長尾真, “TV ニュースと新聞記事の対応付け”, 人工知能学会誌, Vol.12 No.6, pp.113-119, 1997.
- [3] 渡辺靖彦, 長尾真, “パターン情報と自然言語情報の統合による植物図鑑の図の理解”, 人工知能学会誌, Vol.11 No.6, pp.888-895, 1996.
- [4] 渡辺靖彦, 長尾真, “画像の内容を説明するテキストを利用した画像解析”, 人工知能学会誌, Vol.13 No.1, pp.66-74, 1998.
- [5] 嶋田和孝, 遠藤 勉, “製品性能表からの特徴データの抽出”, 情処学自然言語研報, 99-NL-133, pp.107-113, 1999.
- [6] 嶋田和孝, 遠藤 勉, “特徴化された表データからの要約文生成処理”, 信学技報, TL99-29, pp.25-31, 1999.