動向情報の根拠探索のためのテレビ番組からの図表画像検出

梅澤　啓史†　　宮森　　恒†

† 京都産業大学コンピュータ理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1044129,miya}@cse.kyoto-su.ac.jp

あらまし　テレビ番組の映像データから図表画像を検出する手法を検討する．Web コンテンツ上の情報，特に，ある対象の時間的変化を記述した動向情報を対象に，その信頼性判断を支援する根拠の一つとして，テレビ番組で用いられた図表画像を提示する手法を提案している．先行研究では，クエリとして入力された動向情報の根拠となりうる図表画像がテレビ番組中に存在すれば，一定の良好な精度で関連図表を検出できるという結果が得られた．しかし，一部の図表画像の検出精度やカバレッジの向上が必要という課題が残った．本稿では，図表画像の統計的な性質やエッジに基づく特徴などを考察し，様々な検出手法を比較検討する

キーワード　Web コンテンツ, 情報信頼性, 動向情報, テレビ番組, 図表画像

1. はじめに

現在, ネット上では, 膨大な情報が発信・蓄積され続けている. 利用可能な情報量の増大は, 一見すると利便性が向上しているようにも見えるが, 実際は, それら情報は内容の検証や整理をされないままネット上に放置されているようなものであり, 情報が増えるほど, 内容の精査や正しい情報の選別はますます困難になるといわざるを得ない. ソーシャルメディア等の発達により, 情報を発信する手段は普及する一方, 信頼度の高い情報を選別する手段は十分に整備されているとは言い難い.

Web コンテンツの信頼性判断を支援する方法としては, これまでに, 情報の発信者や社会的意見, 情報外観 (参考 文献や連絡先の明記等) といったた多角的な観点からの情 報をユーザに提示する手法や, 評価対象の Web コンテ ンツとそれに関連する Web コンテンツをデータ対で表現し, その support 関係の強さで対象コンテンツの信憑性を 評価する手法などが提案されている.

筆者らは, これまで, ソーシャルネットワーク上の言説, 特に, ある対象の時間的変化を記述した動向情報を対象に, その信頼性判断を支援する根拠の一つとして, テレビ番組 で用いられた図表画像を提示する手法を提案している. 先行研究では, クエリとして入力された動向情報の根拠となりうる図表画像がテレビ番組中に存在すれば, 一定の良好な精度で関連図表を検出できるという結果が得られた. しかし, 一部の図表画像の検出精度やカバレッジの向上が 必要という課題が残った.

そこで, 本稿では, 図表画像の統計的な性質やエッジに基づく特徴などを考察し, 図表画像の検出手法について比較検討する.

2. 関連研究

Web コンテンツの信頼性判断支援の研究は活発に行われている.

その一つに,Web コンテンツの信頼性を判断する基準として,「情報内容, 情報発信者, 情報外観, 社会的評価」といった 4 つの観点から関連情報をユーザに提示することで,Web コンテンツの信頼性判断を支援する研究 [2] がある. 情報内容については,Web ページの本文に書かれている内容に着目しており, 情報発信者については, 発信者の身元に着目した所属の分類やその分野での専門性の有無に着目している. 情報外観については, 情報ソースやデザイン, 連絡先の明記などの Web ページの外観に着目しており, 社会的評価については, 他の利用者がその情報についてどのような見方をしているかに着目している.

また, データ対間の関係分析に着目した Web コンテンツの信憑性評価の研究が挙げられる [3]. ここでは, 評価対象の Web コンテンツとそれに関連する Web コンテンツのデータ対で表現し, その support 関係の強さで対象コンテンツの信憑性を評価するというモデルを導入している. 図表画像の検出については, 文書画像を対象に, 写真と描画の識別をするために, ヒストグラムに基づく特徴を用いた研究がある [4]. ここでは, ヒストグラムの特徴を表す 3 つの尺度を検討し, 写真と描画を区別する尺度の順位は,2 つの最大の山が占めるヒストグラム範囲の割合 (Pct2Pk) が最も高いことを示している.

また, テレビ番組を対象に,HOG 特徴量 [5] と AdaBoost を組み合わせることで図表画像を識別する研究がある [1]. ここでは, 棒グラフ, 円グラフ, 折れ線グラフについて識別器を構築しており, 棒グラフ, 円グラフについては, 比較的良好な検出ができるものの, 折れ線グラフについては十分な精度が得られず, 改良が必要であるという課題が残った.

3. 動向情報の根拠探索

ここでは, 先行研究である, テレビ画像中の図表画像を利用した動向情報の根拠探索 [1] について概要を説明する .

まず, 動向情報とは, 一般に, 商品の売り上げ推移や内閣支持率の変化等, ある事柄や数値についての時間的変化を表現するデータあるいは言語情報を指す. ここでは特に後者の言語情報で表現されたものを対象とする.

また, 図表画像とは, その中に折れ線グラフや円グラフ等の図

表が主として出現しているテレビ番組中の画像フレー ムのことである. テレビ番組の図表は, 予め番組スタッフが政府統計等の複雑な統計データ表から, 最適なデータを選び出し, 視聴者の誰が見ても理解しやすいように噛み砕いた形に視覚化されているため, 一般ユーザが一目で内容を確認するという作業には非常に適している. また, テレビ番組は, プロの番組制作者によって作られており, 公共性を保つ必要があるため, 一定の信頼度が担保された情報 源と考えることができる.

全体の処理手順は以下の通りである.

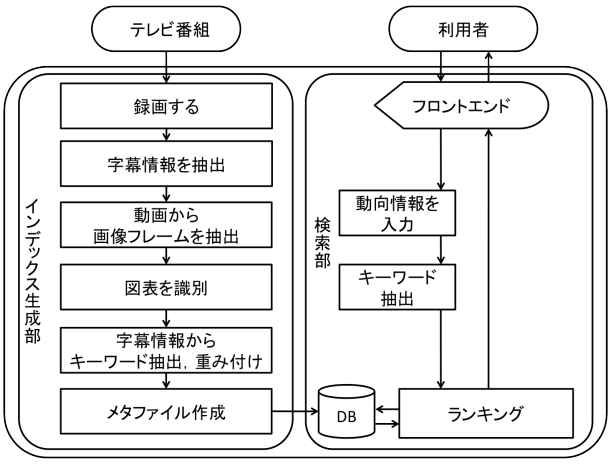


図 1　システム構成図

まず, インデックス生成部は, 図表画像, キーワードを抽出し, メタファイルを生成する. 以下の手順で処理を行う.

（1） テレビ番組を録画して動画，字幕情報（クローズドキャプション）を抽出する

（2） 動画から一定間隔おきに画像フレームを抽出する

（3） 図表画像識別器を用いて，抽出した画像フレームに図表が含まれるかどうか判定する（図表が含まれると判定されたフレームを図表画像と呼ぶ）

（4） 字幕情報から，図表画像の前後 N 秒に出現するキーワードを抽出し，重み付けする

（5） 抽出した図表画像，重み付けしたキーワードをインデックス化し，DB に格納する

検索部は，利用者が与えたクエリを受け取り，DB 内の図表画像をランキングし，結果を出力する．大まかな処理手順は以下の通り.

（1） 利用者が入力した動向情報クエリからキーワードを抽出する.

（2） クエリから抽出したキーワードと，字幕情報から抽出したキーワードの合致度によって，図表画像をランキングする

（3） ランキング結果を出力する

4. 図表画像の検出

4.1 正例，負例データの収集

録画したテレビ番組, および,Web から人手で図表画像を収集し, 正例とする. 図表が使用されている番組のほとんどがニュースであったため, 今回は, ニュース番組を中心に作業を進めた. 図表の種類は, 棒グラフ, 円グラフ, 折れ線グラフを対象とした. 同様に, 負例もテレビ番組から人手で収集した. また, ここでは, 色情報による差を落とすためにグレースケール化し,128x128 ピクセルに正規化した画像を学習データとした.

表 1　正例，負例の件数

グラフの種類	正例数	負例数
棒グラフ	155 件	1300 件
円グラフ	371 件	1300 件
折れ線グラフ	194 件	1300 件

4.2 図表画像の検出に有効な特徴量と学習

本稿では以下の次元数の異なる 3 種類の特徴量を用い、これらの特徴量を特徴ベクトルとして抽出し, 機械学習により識別器を構築し, 比較を行う.

（1） ヒストグラムに基づく特徴量

一般の写真と図表の分類について検討した文献 [4] を参考にした.

図表を含む画像は, 一般に, 平坦なテクスチャで表されることが多く, 図表以外の画像は, 複雑なテクスチャが含まれ, 多くの色が混在しているという統計的な性質があることが指摘されている. そこで, 写真と図表のヒストグラムを利用した特徴量を考えることにより, テレビ番組を対象とした場合でも, 図表検出に有用である可能性があると考えた.

ヒストグラムに基づく特徴量 (PCT と呼ぶ) としては, 以下の 3 種類を用いることとする.

pct2pk　画像の輝度ヒストグラムにおける 2 つの最大の山が占めるヒストグラム範囲の割合. 山とは, 全画素数の 0.5%より大きい画素数を含み,2 つの最小値 (ヒストグラム範囲の 10%を超えない変曲点) の間にある連続したビンとして定義される.

pct0.5　画素数の 0.5%より大きいヒストグラムビンの割合.

absdiff　全ての連続したビンの頻度差の絶対値和. ヒストグラム頻度の変動度合を表す.

（2） K-MEANS による局所特徴量

画像を表現する特徴量を教師なし機械学習で設計する手法 [6], [7] が提案されている. この手法は，学習画像から一定の大きさのパッチをランダムにサンプリングし，パッチの k-means クラスタリングから得られる特徴量 (K-MEANS と呼ぶ) を用いることで，自然画像を高い精度で分類できることが報告されている.

特徴量学習の流れを示す. 本稿では [7] に習い, パッチサイズを 6×6 とする.

1. 学習画像をグレイスケール化し,128×128 画素にリサイズする.
2. 学習画像からパッチをランダムに 100 カ所取得する.
3. 得られたパッチに対して正規化および白色化を施す.
4. k-means クラスタリングによって特徴量を学習する.

ここで, 頻繁に出現する一定の色領域をフィルタリングで除外するために，ランダムに取得するパッチは, 分散が最大画素値の 10%より小さいときは除外し，パッチを再取得する．また，正規化とは，各パッチに対して画素値の平均を引き, 標準偏差

で割る操作であり，これにより，平均 0，分散 1 となり，コントラストが正規化される．また，白色化とは，パッチの共分散行列が単位行列となるような変換処理である．

k-means による学習では，クラスタリングの結果得られたクラスタ重心群をコードブックと呼び、コードブックとの類似度を特徴とする．

図表画像の識別

1. 入力画像からパッチを 1 画素ずつ移動させながら取得する．
2. 学習時と同様に各パッチの正規化を行う．
3. 各パッチを式 (1) により符号化する．
4. 入力画像を 2×2 の領域に分割し，各領域において 3. の総和をとる．
5. 4. で得られた特徴ベクトルを入力画像の特徴ベクトルとする．

$$f(x) = [f_1(x), f_2(x), \cdots, f_K(x)] \tag{1}$$

$$f_k(x) = \max(0, \mu(z) - z_k) \tag{2}$$

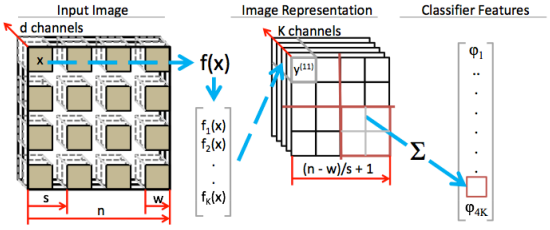


図 2　K-MEANS による局所特徴量の抽出手順 (注3)

ここで $z_k = ||x - c^{(k)}||_2$(k 番目のクラスタ中心 $c^{(k)}$ とのユークリッド距離),$\mu(z)$ は全クラスタ中心との距離の平均値を表す．

本稿では [7] と同様の手順を行い，コードブックサイズを 200 とし，特徴ベクトルの要素数を 800 とした．

（3） エッジに基づく特徴量

先行研究 [1] で用いたエッジに基づく特徴量 (HOG とよぶ) [5] を用いることとする.HOG 特徴量 [5] は，一つの局所領域におけるエッジ情報に着目した特徴量であり，輝度勾配の方向についてヒストグラムをとったものである．人物の検出等に広く利用されているが，高次元ベクトルとなる傾向があり，計算時間がかかるという性質をもつ．本研究では先行研究 [1] にならい，図 3 のように特徴量の計算をおこなった．

5. 実験と考察

本実験では，テレビ番組の動画から図表画像をどの程度正確に抽出できるかを比較する．

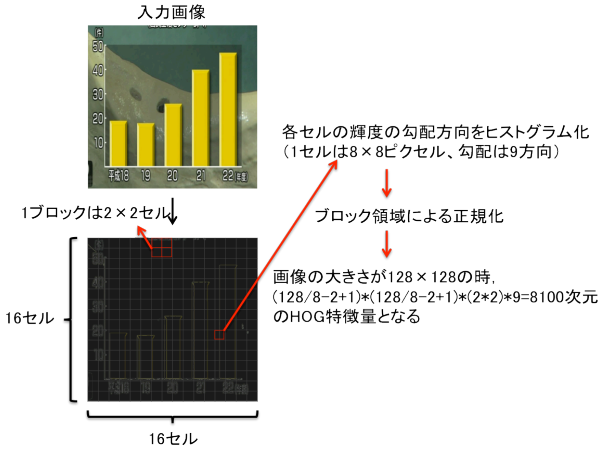


図 3　棒グラフの正例，負例から HOG 特徴量抽出，ヒストグラム作成 (注1)

5.1 実験データ

4.1 節で述べたように録画したテレビ番組,Web から人手で収集した正例，負例 (表 1) を用いて,PCT 特徴量,K-MEANS 特徴量,HOG 特徴量を抽出する．また，収集した正例，負例のデータ数には偏りがあり，十分な分類精度が得られない恐れがあるため,SMOTE [8] を用いて人工的に正例データを生成し，両者が均衡したデータ数 (表 2) となるように配慮した．

表 2　SMOTE をかけた正例，負例の件数

グラフの種類	正例数	負例数
棒グラフ	1240 件	1300 件
円グラフ	1484 件	1300 件
折れ線グラフ	1552 件	1300 件

5.2 評価方法

棒グラフ，円グラフ，折れ線グラフそれぞれの場合について，各特徴量で学習した識別器を用いて,10-fold 交差確認を行い，精度を適合率 (Precision)，再現率 (Recall)，F 尺度 (F-measure) によって評価する．適合率，再現率は，以下の式で算出する

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

だだし，tp は正しく抽出できた正解画像数，fp は間違って抽出してしまった不正解画像数，fn は抽出できなかった正解画像数である．

5.3 結果・考察

PCT 特徴量,K-MEANS 特徴量,HOG 特徴量の適合率，再現率,F 尺度の比較をそれぞれ表 3,4,5 に示す．

5.4 エラー分析

(注3)：文献 [6] より引用

(注1)：画像は 2011 年 10 月 31 日 NHK 総合 NHK ニュース 7 から得られたもの

(注4)：画像は左図:web，右図:web から収集したもの

(注5)：画像は左図:2011 年 06 月 14 日 NHK 総合 NHK ニュース 7，右図:2011 年 06 月 14 日 NHK 総合 NHK ニュース 7 から得られたもの

表 3　適合率の比較

手法	棒グラフ	円グラフ	折れ線グラフ
PCT	0.393	0.593	0.277
	(44/112)	(220/371)	(33/119)
K-MEANS	0.898	0.842	0.811
	(97/155)	(256/304)	(77/95)
HOG	0.913	0.975	0.907
	(116/127)	(347/356)	(127/140)
PCT+SMOTE	0.844	0.848	0.840
	(1140/1351)	(1373/1619)	(1476/1757)
K-MEANS+SMOTE	0.972	0.933	0.946
	(1236/1272)	(1458/1562)	(1538/1626)
HOG +SMOTE	0.990	0.987	0.877
	(1237/1250)	(1472/1492)	(1546/1565)

表 4　再現率の比較

手法	棒グラフ	円グラフ	折れ線グラフ
PCT	0.284	0.593	0.170
	(44/155)	(220/371)	(33/194)
K-MEANS	0.626	0.842	0.397
	(97/155)	(256/371)	(77/194)
HOG	0.784	0.935	0.655
	(116/155)	(347/371)	(127/194)
PCT+SMOTE	0.919	0.925	0.951
	(1140/1240)	(1373/1484)	(1476/1552)
K-MEANS+SMOTE	0.997	0.982	0.991
	(1236/1240)	(1458/1484)	(1538/1552)
HOG +SMOTE	0.998	0.992	0.996
	(1237/1240)	(1472/1484)	(1546/1552)

表 5　F 尺度の比較

手法	棒グラフ	円グラフ	折れ線グラフ
PCT	0.330	0.593	0.211
K-MEANS	0.738	0.842	0.533
HOG	0.823	0.955	0.760
PCT+SMOTE	0.840	0.885	0.892
K-MEANS+SMOTE	0.984	0.957	0.968
HOG +SMOTE	0.994	0.989	0.992

表 6　PCT の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	75 件	105 件
円グラフ	152 件	137 件
折れ線グラフ	95 件	153 件

表 7　PCT+SMOTE の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	218 件	91 件
円グラフ	250 件	111 件
折れ線グラフ	291 件	92 件

PCT 特徴量

PCT 特徴量の誤識別結果を表 6,7 に示す.

誤検出としてイラスト系の画像 (図 4) があげられる.

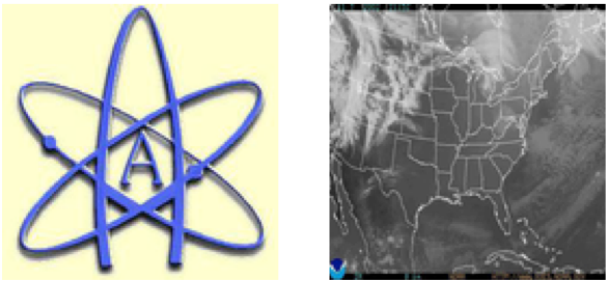


図 4　PCT 誤検出例 (注4)

図 5　PCT 検出漏れ例 (注5)

PCT 特徴量は写真とイラストを識別するために考えられている特徴量であることから, イラストやイラストらしいもの, 写真の用に複雑な色 (濃淡) が使用されていないものが識別されたと考えられる.

検出漏れとして色の変化が激しいものや色の濃淡があるもの (図 5) があげられる.

誤検出とは反対に, グラフの背景色の変化が大きいものや, 複数の色を使用していることによる色の濃淡の変化が大きいものが写真と識別されたものだと考えられる.

K-MEANS 特徴量

K-MEANS 特徴量の誤識別結果を表 8,9 に示す.

誤検出として識別する対象と類似した形状をしたもの (図 6) があげられる. K-means クラスタリングによる学習ではサンプリングした局所領域を対象としているため, 棒グラフであれば直線や直角などの特徴を, 円グラフであれば曲線や丸みといった特徴などをクラスタリングし, クラスタ重心をコードブックとしている. 局所領域がコードブックに類似していた場合検出してしまうものだと考えられる.

検出漏れとしてグラフの背景が複雑なものや, グラフに文字等が重なっているもの (図 7) が上げられる.

局所領域の類似度から特徴量を算出しているので図表に対して文字が重なっていたり, 無地の背景でなく特徴を持ったものであった場合, 作成されたコードブックとの類似度が下がる要因

となると考えられるので検出漏れの原因になったと考えられる.

K-MEANS による特徴量では局所領域ごとのパターンを利用しているため, 似た形状のものを誤検出しやすく, 形状を乱すような位置に文字や背景が存在すると検出漏れの原因となると考えられる.

表 8　K-MEANS の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	11 件	58 件
円グラフ	48 件	115 件
折れ線グラフ	18 件	117 件

表 9　K-MEANS+SMOTE の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	36 件	4 件
円グラフ	104 件	26 件
折れ線グラフ	88 件	14 件

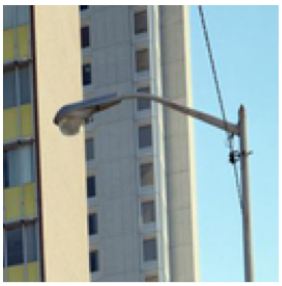


図 6　K-MEANS 誤検出例 (注8)

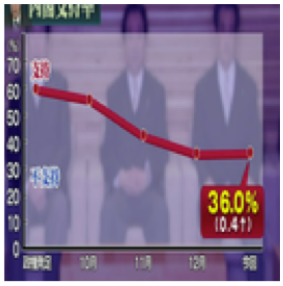


図 7　K-MEANS 検出漏れ例 (注9)

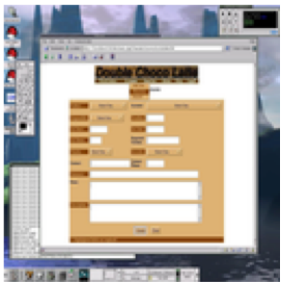
表 10　HOG の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	11 件	51 件
円グラフ	13 件	35 件
折れ線グラフ	8 件	65 件

(注8)：画像は左図:web, 右図:web から得られたもの
(注9)：画像は左図:2011 年 06 月 07 日 ABC テレビ ANN ニュース, 右図:2009 年 03 月 13 日 NHK 総合 NHK ニュース 7 から得られたもの
(注6)：画像は左図:web, 右図:web から得られたもの
(注7)：画像は左図:2011 年 08 月 10 日 NHK 総合 NHK ニュース 7, 右図:2011 年 06 月 14 日 NHK 総合 NHK ニュース 7 から得られたもの

表 11　HOG+SMOTE の誤識別

グラフの種類	誤検出	検出漏れ
棒グラフ	30 件	2 件
円グラフ	20 件	5 件
折れ線グラフ	18 件	3 件



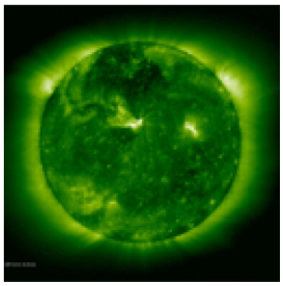
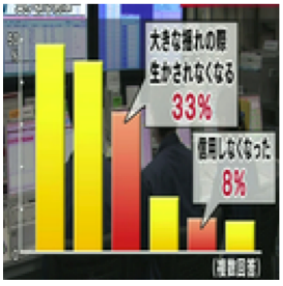


図 8　HOG 誤検出例 (注6)



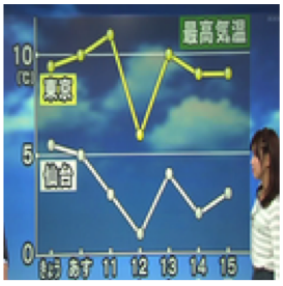


図 9　HOG 検出漏れ例 (注7)

HOG 特徴量

HOG 特徴量の誤識別結果を表 10,11 に示す.

誤検出として識別する対象のエッジの勾配に似たもの (図 8) が上げられる. HOG 特徴量は局所領域について着目した特徴量であるが,K-MEANS 特徴量とは異なり, 局所領域の勾配に着目しているので局所領域で見た場合に類似していなくても全体を通してみた場合に類似している箇所が多ければ検出できる. 誤検出として現れるものは勾配方向が類似しているものだと考えられる.

検出漏れとしてグラフの背景が複雑なもの (図 9) が上げられる. K-MEANS 特徴量と同様にこちらも図表の局所領域の特徴を乱すような背景や文字などによって勾配方向が複雑になり正しく識別できなかったものだと考えられる.

5.4.1　考　　察

PCT 特徴量,HOG 特徴量,K-MEANS 特徴量を比較すると, 正例負例が不均衡であるにもかかわらず, どのグラフについても HOG で高い識別結果が得られている. 特に, 適合率では比較的良好な結果が得られているものの, 再現率は低くなる傾向が

あることがわかった.

また,SMOTE により, データ数を均衡させた場合でもそれぞれの結果を比較すると, いずれのグラフについても HOG で高い識別結果が得られている. これは,HOG のようなエッジに基づく特徴の方がグラフの特徴をよく捉えており, 識別に有効であることを示している.

識別結果は特徴量が多いものから順に高い結果を得ることができた.

ただし,PCT が 3 次元の特徴量であるのに対し, 本稿における HOG は 8100 次元の特徴量であることを考慮すると,PCT のような単純な特徴量を用いても, 正例が十分に収集できれば, 良好な精度を達成しうる能力をもっていることがわかる. 一般に,HOG 特徴量は高次元で計算時間がかかるため, より低次元の特徴量で高い精度を達成することは有用である.

図表画像の識別では, 不均衡データで円グラフ, 棒グラフ, 折れ線グラフの順に精度が高く, 均衡データでは大きな差は見受けられなかった.

円グラフや棒グラフは丸みや直線, 直角といった定まった形状をしているため特徴として有用に働いたと考えられる. しかし, 折れ線グラフは定まった形をもっているとはいえず, 線の角度だけでなく長さや太さ, 各頂点の有無や形状などによって特徴が変化しやすいため識別精度が悪いものになったと考えられる.

6. ま と め

本稿では, テレビ番組の映像データから図表画像を検出する手法を検討した. 図表画像のヒストグラムに基づく特徴,K-MEANS による局所特徴量, および, 局所領域のエッジに基づく特徴についてどのような検出性能の差があるかを比較した. その結果, 次元数の高いエッジに基づく特徴でより良好な結果が得られるものの, ヒストグラムを利用した単純な特徴でも, 正例が十分に収集できれば, 良好な精度を達成しうることがわかった.

今後, 実際のテレビ番組に対して各識別器を用いて識別精度を比較すること, および, 特徴量の組み合わせや別の特徴量を用いた識別精度の向上を検討していく予定である.

文 献

[1] 佐伯 隆太, 宮森 恒: テレビ番組に基づく Web コンテンツ の信頼性判断支援システムの提案, 第 4 回データ工学と情 報マネジメントに関するフォーラム (第 10 回日本データ ベース学会年次大会) (DEIM2012), B3-5, 2012

[2] Miyamori Hisashi, Akamine Susumu, Kato Yoshikiyo, Kaneiwa Ken, Sumi Kaoru, Inui Kentaro, Kurohashi Sadao, ”Evaluation data and prototype system WISDOM for information credibility analysis,” Internet Research, Vol.18, No.2, pp.155—164, 2008.

[3] 山本祐輔, 田中克己: データ対間のサポート関係分析に基 づく Web 情報の信憑性評価, 情報処理学会論文誌 Vol.3 No.2, pp.61—79, 2010

[4] Simske, S.J., ”Low-resolution photo/drawing classification: metrics, method and archiving optimization,” IEEE International Conference on Image Processing 2005(ICIP 2005), Vol.2, II,534-7, pp.11—14, 2005

[5] Navneet Dalal, Bill Triggs, ”Histograms of Oriented Gradients for Human Detection”, International Conference on Computer Vision & Pattern Recognition, Vol. 2, pp.886—893, 2005

[6] A. Coates, H. Lee, and A.Y. Ng, “An analysis of singlelayer networks in unsupervised feature learning,” International Conference on Artificial Intelligence and Statistics, pp.215—223, 2011.

[7] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, Jeffrey Heer ”ReVision: Automated Classification, Analysis and Redesign of Chart Images”

[8] Chawla, Nitesh V. and Bowyer, Kevin W. and Hall, Lawrence O. and Kegelmeyer, W. Philip, ”SMOTE: synthetic minority over-sampling technique”, Journal of Artificial Intelligence Research, Vol.16, No.1, pp.321—357, 2002