Vol.2015-CVIM-197 No.13
2015/5/18

文書領域検出のための人工生成サンプルを用いた教師あり学習

瀬川 雄太[1,a)]　川本 一彦[2,b)]　岡本 一志[3,c)]

概要：文書領域検出のための教師あり学習に関して，完全に人工生成した文書領域画像データを学習用サンプルとして用いる方法を提案し，一人称視点映像中の文書領域の検出に応用する．一般に教師あり学習では，学習用サンプルを大量に収集することが必要になるが，文書画像の電子的な収集や蓄積は著作権等の制度的な問題を抱えている．そこで本研究では，文字列，文字濃度，文字間隔，行間隔といった要素を文書領域生成のために決定し，回転，輝度変化，ノイズによる加工を行って学習用サンプルとなる文書領域画像を人工生成する．評価実験では，読書行為を含む一人称視点映像に対して，二つの識別器を用いて検出率を評価した．一つは文書領域検出によく利用されるガボール特徴を用いた最近傍識別器で，もう一つは深層畳み込みニューラルネットワークを用いた特徴学習および識別である．10 種類の読書シーンに対して平均識別精度を評価した結果, 前者の識別における誤検出率は 5.4%, 未検出率は 29.0% であり, 後者の識別においてはそれぞれ 3.7%, 19.5% であった．

キーワード：人工生成サンプル, 一人称視点映像, 文書領域検出

1. はじめに

近年のネットワーク環境の発達に伴い, 大学図書館では本や雑誌, 新聞などの物理的な情報源に加え, 電子ブックやOPAC などといった電子的な情報源も, お互いを補い合う形で提供されている．それらの情報源の活用のされ方「情報利用行動」の調査は, 図書館のより良い環境構築に必要な基礎研究である [11]. 図書館における情報利用行動とは, 物理的な情報源, および電子的な情報源を利活用する行動を指す．

従来, 情報利用行動の調査は動画の目視確認などによって行われている．そのため, 調査の大規模化には効率性が問題となる．そこで, 図書館利用者による一人称視点映像を用いた, 情報利用行動の機械的な識別を行うことでのコスト削減を考える．情報利用行動には「読む」「PC を使う」「ノートを書く」などといった「行動」[13] に加え, その対象となる「本」「PC」「ノート」のような「情報源」に関する情報が含まれる．物理的な情報源の種類を識別する取り組みには [10] があり, 画像に映っている情報源を学習によって「雑誌」「新聞」「漫画」などのクラスに分類している．本研究では物理的な情報源の利用行動を識別することに焦点を当て, 行動の対象となる本や雑誌といった情報源を一人称視点映像中から検出することに取り組む．

本や雑誌のような文書の検出は, 一般物体認識の手法を用いて教師あり学習を行うことで実現できる．一般に教師あり学習では, 学習用サンプルを大量に収集することが必要になるが, 文書画像の電子的な収集や蓄積は著作権などの制度的な問題を抱えている．そこで本研究では, 大量の文書領域画像の完全な人工生成を行い教師あり学習に用いる方法を提案し, 一人称視点映像中の文書領域の検出に応用する．

評価実験では, 二つの識別器を用いて検出率を評価する．一つはガボール特徴を用いた最近傍識別器である．ガボール特徴は文書領域検出によく利用されており, [8], [9] でもガボールフィルタによって得た文書領域の特徴を用いて教師あり学習を行っている．もう一つは深層畳み込みニューラルネットワーク [7] を用いた特徴学習および識別である．[4] や [5] では, 畳み込みニューラルネットワークを用いて

[1] 千葉大学大学院融合科学研究科
Graduate School of Advanced Integration Science, Chiba University
[2] 千葉大学統合情報センター
Institute of Management and Information Technologies, Chiba University
[3] 電気通信大学 大学院情報理工学研究科
Graduete School of Informatics and Engineering, The Universiy of Electro-Communications
a) segawa@chiba-u.jp
b) kawa@faculty.chiba-u.jp
c) kazushi@uec.ac.jp

表 1　文書領域画像の生成パラメータ.

パラメータ	値
画像サイズ	32 × 32 pixel
文字濃度	輝度 [32, 128]
文字間隔	pixel 数 [0, 24]
行間隔	pixel 数 [8, 16]
回転角度	$[-180°, 180°]$
輝度変化	輝度 [64, 192]
ノイズ	平均 0, 分散を $[5^2, 15^2]$ から選んだガウス分布.

識別器を構成し識別に有効な特徴の学習を行うことで, 文書領域の検出に応用している. これらの研究では, 文字そのもののデータセットによって学習を行い, ストリートシーンにおける文書領域の検出を行っている. 本研究では, 完全に人工生成した文書領域画像を用いた教師あり学習によって, 一人称視点映像中の 10 種類の読書シーンについて文書領域の検出を行う.

2. 文書領域画像サンプルの人工生成

一人称視点映像中の本の検出に応用するために, 本の表面の文書領域に関する教師あり学習を行う. 学習のために, ポジティブサンプルとして文書領域画像を, ネガティブサンプルとして文書領域を含まない画像を用意しデータセットを作成する. 一般に, 教師あり学習には大量の学習サンプルを要するが, 文書領域画像の収集は著作権などの制度的な問題により難しい. そこで, 実際の本が持つ文書領域が人工的な手続きによって生成されていることに注目し, 本研究では文書領域画像を完全に人工生成する.

文書領域画像を次のような手順で人工生成する. まず始めに, 文書領域画像の生成に関するいくつかのパラメータを**表 1** のような値に決定する. 画像サイズを除いた生成パラメータは, 表 1 にある範囲内から生成する画像ごとに無作為な値に決定される. 次に, 半角英数字の組み合わせでを用いて描画する文字列を無作為に決定し, 文字サイズ, 文字濃度, 文字間隔, 行間隔のパラメータに従って画像上に描画する. 最後に, 回転と輝度変化, ノイズ付加の処理を行う. 以上の手順を踏んで, **図 1**(a) のような文書領域画像が生成される. (b)-(g) は, (a) に関する画像サイズ以外の生成パラメータを変化させたものである.

3. 文書領域検出のための学習と識別

本研究では, 完全に人工生成された学習サンプルを用いた教師あり学習について, 二つの識別器を用いることで評価する. 一つはガボール特徴を用いた最近傍識別器で, もう一つは深層畳み込みニューラルネットワークを用いた特徴学習および識別である. ガボール特徴は, ガボールフィルタを用いて 2 次元画像から抽出される特徴であり, 文書領域検出によく利用されている. 深層畳み込みニューラルネットワークは, Deep Learning に用いる多層ニューラル

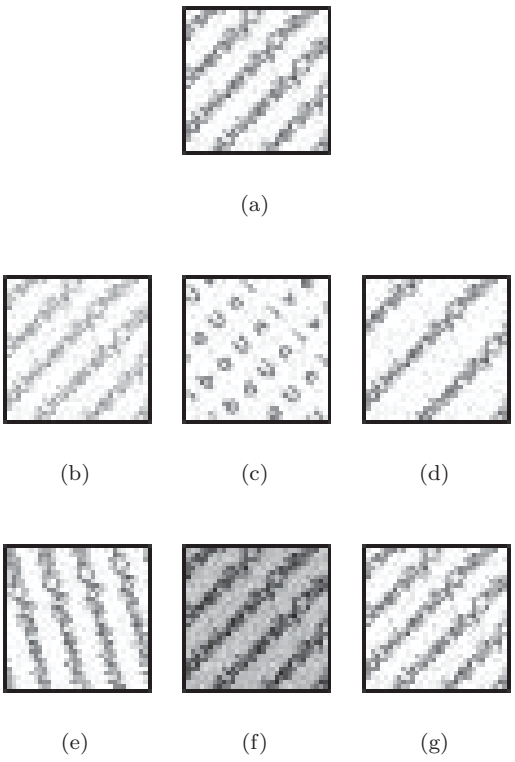


図 1　(a) 人工生成された文書領域画像と (b)-(g) 生成パラメータ変化による画像の見えの違い. (b)-(g) は (a) に対して (b) 文字濃度 (c) 文字間隔, (d) 行間隔, (e) 回転角度, (f) 全体の輝度, (g) ノイズの生成パラメータを変化させたもの.

ネットワークの一つである. Deep Learning を用いることで, 識別に有効な特徴量を学習によって獲得することができる.

3.1 ガボール特徴を用いた最近傍識別

文書領域特徴の抽出のために, ガボールフィルタを用いる. 得た特徴量を用いて, 最近傍識別器を用いた学習と識別を行う.

3.1.1 ガボールフィルタ

ガボールフィルタは正弦波とガウシアンからなるフィルタである. 2 次元ガボールフィルタは, ある 2 次元点 (x, y) に対して,

$$h(x,y) = \exp\left\{-\frac{1}{2}\left[\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right]\right\}\cos(2\pi u_0 x' + \phi), \quad (1)$$

なる $h(x, y)$ を与える [6]. ここで, σ_x^2, σ_y^2 はそれぞれガウス関数の分散を, u_0, ϕ はそれぞれ正弦波の中心周波数, そして x' 方向の位相を表す. また, (x', y') はフィルタの回転角 θ を用いて

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \quad (2)$$

なる 2 次元点である. 2 次元ガボールフィルタのグラフは**図 2** のように表される.

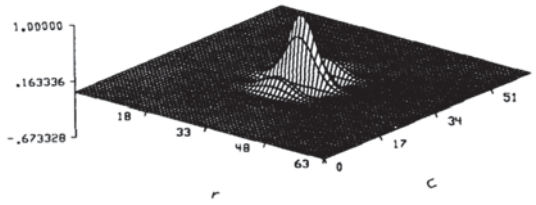


図 2 ガボールフィルタのグラフ [2].

ガボールフィルタは, 適用した関数に関してフィルタの方向に応じた方向の成分を強調する性質がある. この性質から, ガボールフィルタによって 2 次元画像から文書領域に関する特徴を獲得することが期待できる. 本研究では, $\theta = 0°, 45°, 90°, 135°$ の 4 方向のガボールフィルタを用いて特徴量を抽出する.

3.1.2 特徴量の抽出

ガボールフィルタによって得られた応答から文書領域の特徴量として texture energy [6] を抽出する. texture energy は平均絶対偏差 (AAD; Average Absolute Deviation) の計算により求められる. 平均絶対偏差は, ある窓内における値のばらつきを表す指標の一つである. ガボールフィルタによって得られた応答画像 r 上の点 (x, y) に関して, これを中心とする一辺 M の窓 $W_{x,y}$ を考える. 点 (x, y) における, 窓 $W_{x,y}$ 内での平均絶対偏差 $e(x, y)$ は,

$$e(x,y) = \frac{1}{M^2} \sum_{(a,b)\in W_{x,y}} |\psi(r(a,b))| \qquad (3)$$

のようにして得られる. ここで ψ は, スケーリングのための関数であり,

$$\psi(t) = \tanh(\alpha t) = \frac{1 - e^{-2\alpha t}}{1 + e^{-2\alpha t}} \qquad (4)$$

と表される. $\alpha = 0.25$ のとき, ψ はシグモイド関数のように閾値処理変換を行う関数となる. 本研究では $\alpha = 0.25$ を用いる. 画像 r 上のすべての点に関して計算した平均絶対偏差を, texture energy 特徴量として学習と識別に用いる.

3.1.3 最近傍識別器による識別

得られた特徴量を用いて, 最近傍法による学習と識別を行う. 最近傍法を用いた識別によって, ある入力データは予め与えられたデータ群のうちその最近傍にあるデータと同じクラスとして分類される.

3.2 深層畳み込みニューラルネットワークを用いた特徴学習と識別 [12]

深層畳み込みニューラルネットワークを用いて, 文書領域の識別に有効な特徴の学習とそれを用いた識別を行う. ネットワークは, 文書領域を含む画像をポジティブクラス, 含まない画像をネガティブクラスとした 2 クラス分類器として構成される. ネットワークの学習には確率的勾配法を用いる. 学習したネットワークを用いて, 画像中の文書領域の識別を行う.

表 2 用いるネットワークの構成.

層	種類	カーネルサイズ, 出力数
入力	画像	32 × 32
1	畳み込み	5 × 5, 32
2	プーリング	2 × 2
3	maxout	4
4	畳み込み	5 × 5, 32
5	プーリング	2 × 2
6	maxout	4
出力	Classify	2

ネットワークは入力層, 畳み込み層, プーリング層, 出力層からなる. 構成を**表 2** に示す.

3.2.1 畳み込み層

畳み込みは近接画素とのみ結合を行うことで, 局所的な応答を表現する. 本研究では, 畳み込み層の活性化関数として maxout [3] を用いる. maxout は,

$$h'_i = \max_{j\in[1,k]} h_{ij} \qquad (5)$$

のように k 個の特徴マップから各画素の最大値を出力マップの画素値とする手法であり, シグモイド関数や打ち切り線形関数より高い表現力をもつ.

3.2.2 プーリング層

プーリングはパターンの幾何学的変動, 照明条件などに対して不変な特徴を取り出すことを目的とする. 本研究では, プーリングとして max pooling を用いる. max pooling とは,

$$h'_i = \max_{j\in K_i} h_j \qquad (6)$$

のようにプーリングの対象領域 K に含まれる各画素 h_i の最大値を出力とする手法である. max pooling を用いることで汎化性を向上させることができる [1] .

3.2.3 出力層

出力層では, ポジティブ, ネガティブそれぞれのクラスの確率を求める. クラス y^i の確率 $P(y^i)$ は, softmax によって

$$P(y^i) = \frac{\exp(h_i)}{\sum_j^2 \exp(h_j)} \qquad (7)$$

のように求める.

4. 一人称視点映像中の文書領域の検出

実際に読書行為を含む一人称視点映像を撮影し, 人工生成サンプルを用いた教師あり学習に関する評価実験を行う. 本研究では, ガボール特徴を用いた最近傍識別器による識別と, 深層畳み込みニューラルネットワークを用いた特徴学習と識別の, 二つの方法によって学習を評価する. 最近傍識別では OpenCV の FLANN ライブラリ [*1] を用いた近

[*1] http://opencv.jp/opencv-2.2_org/cpp/flann_fast_approximate_nearest_neighbor_search.html

(a) 一人称視点カメラ. (b) 装着した状態.

図 3 評価用データ採取のための実験環境.

似最近傍探索を行う. インデックス生成にはランダム kd ツリーと階層型 k 近傍ツリーを組み合わせて用いた. ネットワークの学習には Deep Learning ライブラリ Pylearn2 [*2] を用いた. ネットワークの更新回数は 1 万回である.

4.1 実験環境

用いる一人称視点映像は, デスクシーンにおける読書行為を撮影したものである. 撮影には, Panasonic® 社の一人称視点カメラ, HX-A500 [*3] (**図 3**(a)) を図 3(b) のように装着し使用した. 評価するシーンはまず, デスク上が整頓されているシーンと, デスク上に読書対象以外の物が雑多に置かれているシーンの二つに大別される. これら二つのシーンそれぞれにおける, 「本」「雑誌」「文庫本」「辞典」「原稿」を対象とした 5 つの読書行為のシーンに関して評価実験を行う. 評価用データとして, 映像中のそれらのシーンから**図 4** のような 10 種類の静止画を抽出し用意した. 抽出された静止画のサイズはいずれも 960×540 pixel である.

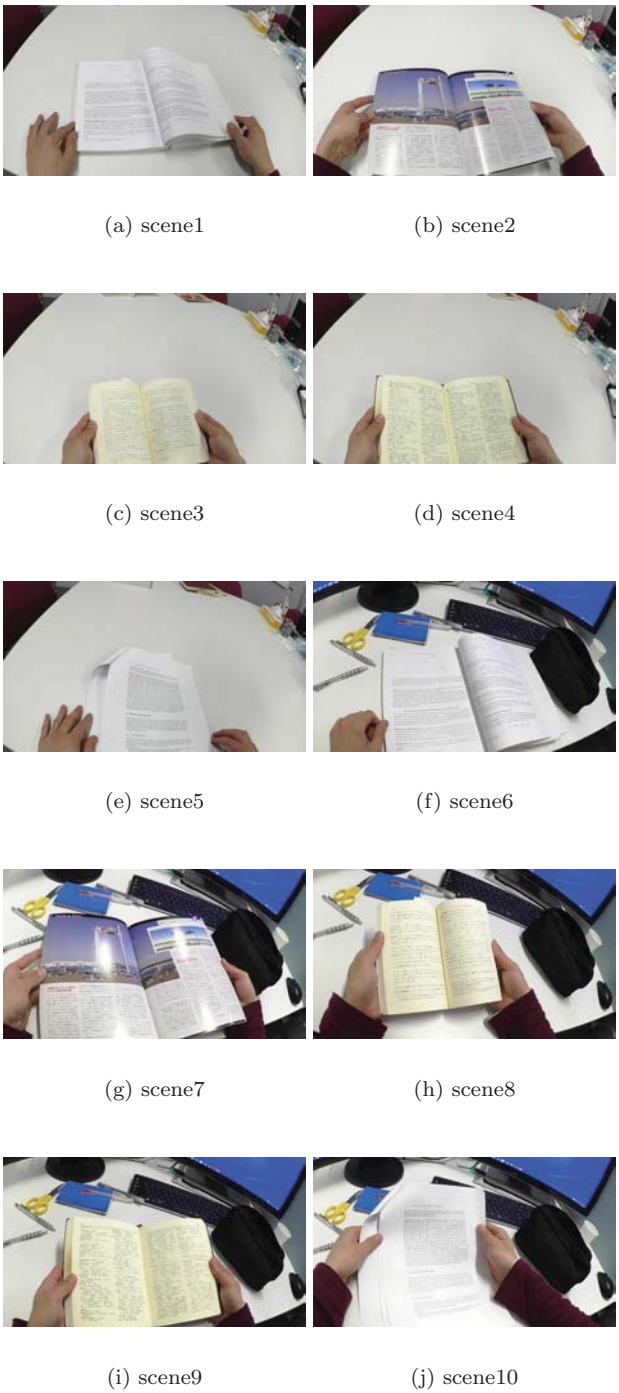
4.2 サンプルの人工生成とデータセットの作成

学習用データセットとして, ポジティブサンプル, ネガティブサンプルをそれぞれ 5 万枚ずつ作成し, 計 10 万枚のサンプルを用意する. ポジティブサンプルは, 2 節で述べた方法を用いて人工生成し用意する. ネガティブサンプルは, 文書領域を含まない一人称視点映像中から回転を加えて無作為に切り出し**図 5** のように用意する. サンプルのサイズはいずれも 32×32 pixel である.

4.3 文書領域の検出

表 3 のように用意したデータセットを用いて, 二つの方法による文書領域の学習と識別を行う. 評価用画像に対して検出窓を 50% の移動率で走査し, ポジティブ領域との重

[*2] http://deeplearning.net/software/pylearn2/
[*3] http://panasonic.jp/wearable/a500/



(a) scene1 (b) scene2
(c) scene3 (d) scene4
(e) scene5 (f) scene6
(g) scene7 (h) scene8
(i) scene9 (j) scene10

図 4 異なるデスク環境と読書対象について採取された評価用データ. (a)-(e) は整頓されたデスクシーン, (f)-(j) は読書対象以外が多く映るデスクシーンにおける読書行為. 二つのシーンにおいて, それぞれ「本」「雑誌」「文庫本」「辞典」「原稿」の 5 つを読書対象としている.

なり率が 80% 以上のとき, 正しく検出できているとする. **図 6** は二つの方法における文書領域の検出結果の一例である. ポジティブ領域は手作業で各フレームに与えた.

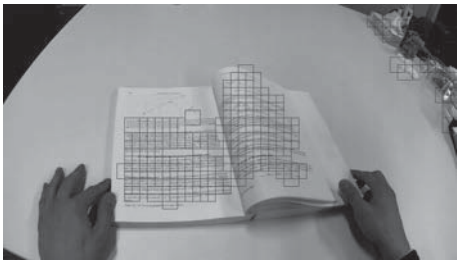
4.4 実験結果

ガボール特徴を用いて最近傍識別を行う方法と, 深層畳

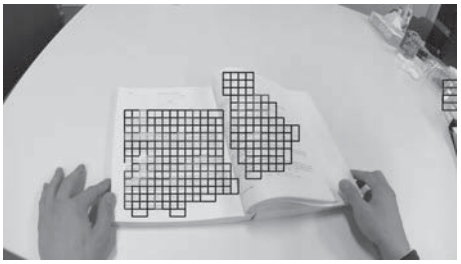
図 5　実験に用いるネガティブサンプル.

表 3　実験に用いた学習用, 評価用データセット.

	ポジティブ	ネガティブ	評価用
サイズ	32×32	32×32	960×540
枚数	50000	50000	10



(a)



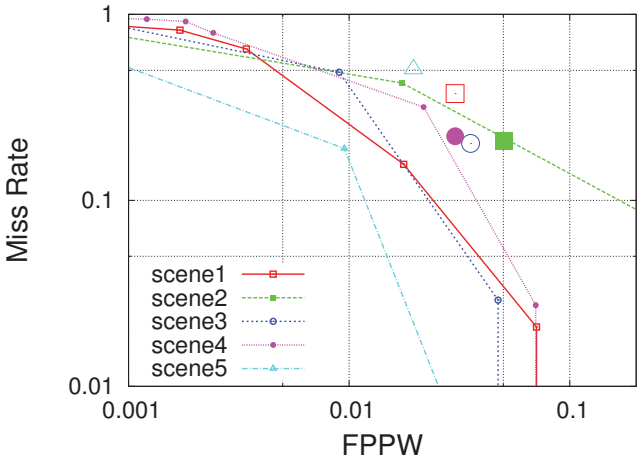
(b)

図 6　二つの方法による文書領域の検出結果の例. 矩形は検出窓. (a) ガボール特徴を用いた最近傍識別, (b) 深層畳み込みニューラルネットワークによる識別.

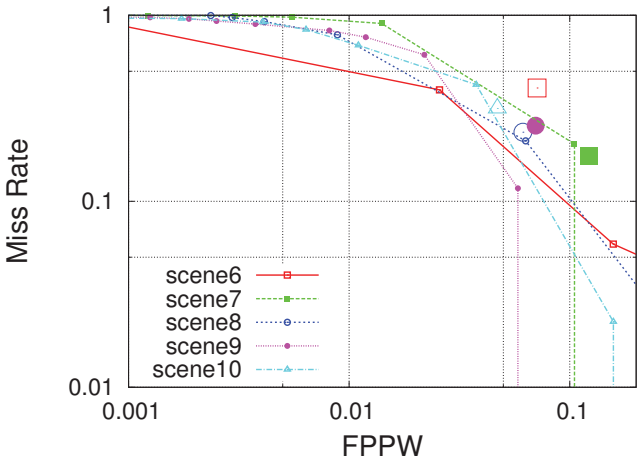
表 4　ガボール特徴を用いた最近傍識別の結果.

シーン	誤検出率	未検出率
scene1	3.0%	37.5%
scene2	5.0%	20.9%
scene3	3.6%	20.2%
scene4	3.0%	22.1%
scene5	2.0%	50.6%
scene6	7.1%	40.7%
scene7	12.2%	17.6%
scene8	6.1%	23.5%
scene9	7.0%	25.5%
scene10	4.7%	31.5%

み込みニューラルネットワークを用いる方法のそれぞれについて, 実験結果を**表 4**, **図 7** に示す. 図 7 には, 表 4 の結果についても対応する色と印で示している.



(a) 整頓されたデスクシーンでの結果.



(b) 乱雑に物が置かれたデスクシーンでの結果.

図 7　深層畳み込みニューラルネットワークを用いた特徴学習と識別の結果を表す DET 曲線. 横に窓単位での誤検出率 (FPPW), 縦に未検出率 (Miss Rate) の対数軸を取る. 図中の印は, 対応するシーンについて, ガボール特徴を用いた最近傍識別の結果を表す.

4.5　考察

図 7 に注目して識別性能を評価する. 図 7 は横軸に窓単位での誤検出率 (FPPW), 縦軸に未検出率 (Miss Rate) を取るため, グラフが原点に近いほど性能が良いことを示す. まず初めに二つの識別器の性能に注目すると, 深層畳み込みニューラルネットワークによる識別はガボール特徴を用いた最近傍識別に比べ, ほぼ全てのシーンにおいて高い識別精度であることがわかる. これはネットワークの学習によって有効な特徴を学習, 獲得しているためだと考えられる.

次に, 二つのデスク環境におけるシーンについて識別精度を比較する. (a) に比べ (b) は全体的に誤検出率が大きく

なっている．このことから, デスク上にある読書対象以外の物が誤検出を増やす要因となっていることがわかる．

最後に, 映像中の文書領域検出を本の検出へ応用することに関して, 人工生成サンプルを用いた教師あり学習を評価する．文書領域の有意な集合を用いて本の領域の識別を行うためには, 文書領域の未検出率が小さいことが期待される．したがって, 少なくとも 50% を上回るポジティブ領域が検出されている, すなわち, 未検出率が 50% 以下となるような識別精度を実用的な値として考えることができる．加えて, 本実験で使用した一人称視点映像におけるポジティブ領域が, フレーム全体に対して 10%〜20% 近くを占めていることから, 誤検出率が 10% 以下となることが実用的精度のための条件として期待される．ここで, これら実用的精度の条件の未検出率 50% 以下, 誤検出率 10% 以下における識別結果について注目する．(a), (b) 二つのデスク環境においても, DET 曲線および点で示される二つの識別器は, 概ねこの条件を満たす性能で識別が行えていることがわかる．このことから, 人工生成サンプルを用いた文書領域の教師あり学習に関して, 実用的な文書領域検出への応用が期待できる．

5. おわりに

本研究では, 文書領域検出のための教師あり学習に関して, 完全に人工生成した文書領域画像を用いる方法を提案している．文字の濃度や間隔, 行の間隔, 画像の回転角度, 輝度変化, ノイズといった生成パラメータを無作為に決定し, 大量の文書領域画像の人工生成を行い教師あり学習に用いた．これによる文書領域検出の評価のために, ガボール特徴を用いた最近傍識別と, 深層畳み込みニューラルネットワークによる特徴学習および識別の, 二つの方法を用いて, 一人称視点映像中の読書行為を映したシーンに対する評価実験を行った．実験結果から, 文書領域検出のための人工生成サンプルを用いた教師あり学習が実用的な精度で行えることを示した．しかし, 人工生成のために与えたパラメータは, 一定の範囲から無作為に決定したものである．文書領域の学習に有効なサンプルの人工生成を追究するにあたって, 生成パラメータの種類や値の決定の方法には検討の余地がある．

謝辞

本研究は JSPS 科研費 25330186 の助成を受けたものです．

参考文献

[1] Boureau, Y.-L., Ponce, J. and LeCun, Y.: A theoretical analysis of feature pooling in visual recognition, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 111–118 (2010).
[2] Farrokhnia, F. and Jain, A. K.: A multi-channel filtering approach to texture segmentation, *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, IEEE, pp. 364–370 (1991).
[3] Goodfellow, I., Warde-farley, D., Mirza, M., Courville, A. and Bengio, Y.: Maxout Networks, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1319–1327 (2013).
[4] Huang, W., Qiao, Y. and Tang, X.: Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees, *Computer Vision–ECCV 2014*, Springer, pp. 497–511 (2014).
[5] Jaderberg, M., Vedaldi, A. and Zisserman, A.: Deep features for text spotting, *Computer Vision–ECCV 2014*, Springer, pp. 512–528 (2014).
[6] Jain, A. K. and Bhattacharjee, S.: Text segmentation using Gabor filters for automatic document processing, *Machine Vision and Applications*, Vol. 5, No. 3, pp. 169–184 (1992).
[7] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).
[8] Yi, C. and Tian, Y.: Text Detection in Natural Scene Images by Stroke Gabor Words, *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 177–181 (online), DOI: 10.1109/ICDAR.2011.44 (2011).
[9] Yi, C. and Tian, Y.: Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification, *Image Processing, IEEE Transactions on*, Vol. 21, No. 9, pp. 4256–4268 (online), DOI: 10.1109/TIP.2012.2199327 (2012).
[10] 志賀優毅，内海ゆづ子，岩村雅一，黄瀬浩一ほか：読書活動の自動的記録のための文書画像の識別，電子情報通信学会論文誌 D， Vol. 97, No. 12, pp. 1733–1736 (2014).
[11] 寺井仁：大学図書館における情報探索活動に関する研究: われわれはいかに異なる情報源を活用しているのか?，名古屋大学附属図書館研究年報，Vol. 5, pp. 69–82 (2007).
[12] 浅沼 仁，川本一彦，岡本一志：Deep Convolutional Neural Network による全方位画像からの人検出，情報処理学会研究報告，Vol. 2015-CVIM-195, No. 59, pp. 1–4 (2015).
[13] 堀内麻由，川本一彦，岡本一志：一人称視点カメラと加速度センサを用いた情報利用行動の識別，HCG シンポジウム講演論文集，pp. 282–285 (2013).