Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

Sobel オペレータと OCR を用いた多重画像 spam フィルタの提案

万　鵬[†]　　上原　稔[†]

Spam メールは，受信者の時間を浪費させだけではなく，また不良メッセージ及びウイルスを伝播している． Spam を防ぐには spam フィルタが用いられている．それを避けるために，spam 送信者は spam メールを画像化し，大量送信している，これを画像 spam という．従来な画像 spam フィルタは，文書のスキャン及び写真に弱みが存在している．一方，情報を伝えるために，多くの場合，文字が必ず 画像 spam に存在する．そこで，本研究では，我々は画像メールを文字と画面の構成により分類し，水平垂直 sobel オペレータと 45 度 sobel オペレータを用いたエッジ検出の結果の違いより，spam を検出する手法を提案し，そのフィルタと OCR を利用し，多重フィルタを構築する．

Spam Detection using Sobel Operators and OCR

PENG WAN[†]　MINORU UEHARA[†]

Spam mails do not only lead to the waste of time to the recipients but also spread the bad message and virus. To avoid the spam filters, spam senders usually send image instead of text. the traditional image spam filter have weaknesses in scanning documents and photographs. However, in order to transmit information, the letters are always used in the image spam. Therefore, in this study, we classified mail image by the configuration of letters and the drawing. And we propose a spam detection method that using horizontal vertical and diagonal sobel operators for edge detection to detect the spam and a multiple filter using sobel operator and OCR.

1. はじめに

ここ数十年，インターネットは，我々の仕事や生活の中で，益々重要な役割を演じていく．その中，コストが低い，効率的で，便利な交流手段として，メールは既に我々の生活から離せなくなった．しかしながら，2007 年 Kaspersky が発表した報告によると，2001 年から 2007 年までの６年間，spam（迷惑メール）が全メールに占めている比重は 5%から 95%まで上昇した．2011 年２月中国インターネット協会 spam 対策センターの報告[1]によると，中国人が毎週 13.5 つの spam を受けているという．

従来の spam 対策は，来源検査及び内容検査二つの分野に分かれている．来源検査は，オープンリレーの禁止[2]，ブラックリストの設定[3]，送信者ドメインの検査[4]などの対策を行っている．内容の検査には，キーワードの検査[5]が普通だが，spam 送信者（spammer）が spam を画像にし，画像 spam を送ることによって，キーワードの検査は効能低下になってしまう．

画像 spam の特徴を抽出し，複数の特徴によって多重フィルタの連続処理が多数提案されている．テキストと画像情報を用いたフィルタ[6]や，画像 spam の類似性によるフィルタ[7]や，テキスト抽出とか[8]，時間をかかって，複雑な処理を行っても，画像 spammer の簡単な作業によって，効能低下になる．また，新聞や雑誌の写真及びスキャンは，判定し難い，それは，テキスト部分が実物に存在するのか，送信者が添付したのか，完全判別できないからである．

そこで，本研究では，我々まずは画像メールを四種類に分類する：

- 画面のみ ham
- 文字あり ham
- 文字のみ spam
- 画面あり spam

そして，我々が提案したフィルタは，閾値の調整によって独立フィルタとして処理しても，多重フィルタの一環として作業しても対応できる．また，本研究の重要な特長は以下の２つがある．

- spam の中で多数を占めている文字のみ spam を完全検出である．
- 他のフィルタに弱い文書の写真及びスキャンを，誤認せず検出できる．

手法の方，我々は画像メールを水平垂直 sobel オペレータでエッジ検出し，その結果を，45 度 sobel オペレータでエッジ検出した結果と比較し，その比率により画像 spam を検出するフィルタを提案し，そのフィルタの上，光学文字認識を用いた多重フィルタを提案する．

本論文の構成を以下に示す．まず２章で sobel 演算子及びエッジ検出について述べ．

† 東洋大学工学研究科情報システム研究科
Graduate School of Open Information Systems Engineering,Toyo University

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

３章では提案方式の詳細について示す．４章では実験評価を行う．５章ではまとめと今後の課題について述べる．

2. 関連研究

2.1 エッジ検出

エッジ検出は，数学的手法を使い，画像のピクセルから，輝度値（グレー）を持つ，大規模スペース勾配方向があるエッジ及び線を抽出することである．エッジ検出の目的は、デジタル画像の明るさの変化は明白なポイントを特定することである．画像のエッジは、しばしば重要な性質と大きな変化を反映している．その計算方法は以下の様に示す．

$$L_x(x,\ y) = -\frac{1}{2} * L(x-1,\ y) + 0 * L(x,\ y) + \frac{1}{2} * L(x+1,\ y) \quad (1)$$

$$L_y(x,\ y) = -\frac{1}{2} * L(x,\ y-1) + 0 * L(x,\ y) + \frac{1}{2} * L(x,\ y+1) \quad (2)$$

2.2 Sobel オペレータ

sobel オペレータとは，離散的な差分演算子であり，主にエッジ検出に使用され，画像の明るさの関数の勾配の近似値を計算する画像処理の演算子の１つである．画像内の任意の点でこの演算子を使えば，対応する勾配ベクトルを生成することができる．sobel オペレータは式（3）のフィルタに基づいている．

$$L_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * L \quad と \quad L_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * L \quad (3)$$

1 次微分の見積もりがこのように与えられると，勾配の大きさは式（4）のように計算できる．

$$|\nabla L| = \sqrt{L_x^2 + L_y^2} \quad (4)$$

このとき勾配の方向は式（5）のように見積もられる．

$$\theta = atan2(L_y,\ L_x) \quad (5)$$

3. 提案手法

3.1 メール画像の分類

spam は，送信者の様々な目的で，大量送信されている．その目的を達成するには，情報を伝えないといけない．すなわち，spam には，必ず文字を含めていること．それは，画像 spam も同じである．

そこで，本研究では，この特徴によって，メールに添付される画像を，図 1 の様に分類する．

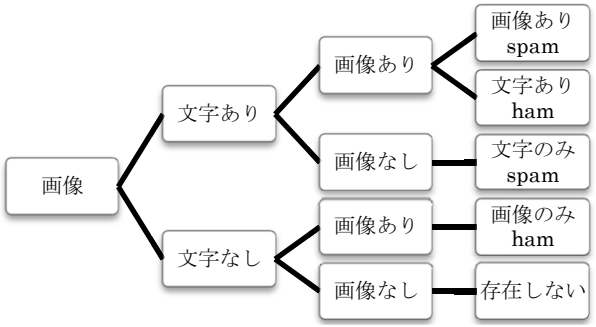

図 1　メールに添付する画像の分類

図 2　分類例

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

図 2 のように，画面のみ ham とは，文字全く付かずの自然画像のことである．文字あり ham とは，新聞や雑誌の写真及びスキャンなど，実物に文字が付いている自然画像など，また合法的な広告のことである．一方，画面がなく文字のみのイメージファイルが ham だった場合，撮影，スクリーンショットまたスキャンしたものしか考えられない.その理由は合法的なテキストをイメージ化するより，PDF ファイルに転換した方が最も使用されている.撮影とスキャンに関する検討は 3 章で示す.スクリーンショットについて，インターネットから収集した 3299 枚画像 spam と 2027 枚 ham に，合法的なテキストのスクリーンショット一つもないのことで，本稿は検討しないとする..

画面あり spam とは，spam に文字だけではなく，背景に何かのパターンや図面が存在する spam のことである．文字のみ spam とは，背景の色にも関わらず，全画像に文字のみ存在する spam のことである．そこで，インターネットから収集した 3299 枚画像 spam の中，画面あり spam と文字のみ spam の比率は図 3 のように示す.

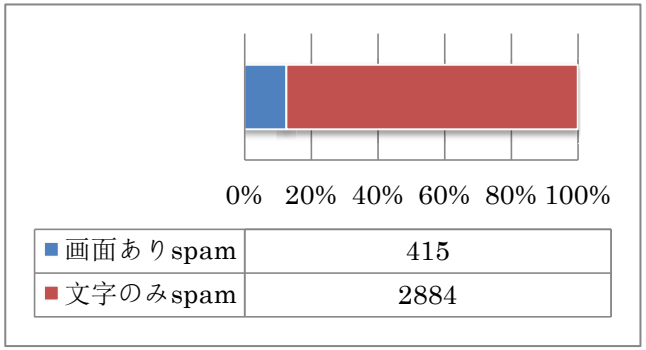

図 3　spam 種類の割合

3.2　45 度 sobel オペレータ

Sobel オペレータは，普通水平と垂直二つのフィルタがあるが，式（6）の様な 45 度フィルタを使うこともできる.

$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & +1 \\ 0 & +1 & +2 \end{bmatrix} \tag{6}$$

この 45 度 sobel オペレータは，エッジに敏感過ぎで，普段は殆ど使われていないが，本研究では，45 度と 0/90 度 sobel オペレータによるエッジ検出の結果を比較し，spam を検出する.

3.3　基本方針

自然画像に，通常は大量の自然的ノイズが存在している．色，光度，形などの小さな，頻繁な変化はエッジとして検出されることができる．一方，画像にある文字は，色と形が変化しても自然的ノイズの様なエッジが起こらない．そこで，45 度 sobel オペレータは自然的ノイズを含めるエッジ曲線を検出することできる．一方，水平垂直 sobel オペレータは単なる画像の輪郭を検出することができる．これらの結果を比較することによって文字は全画面に占めている比率が得られる．そこで，閾値を設定し，画像メールを分類することができる．図 4 にフィルタの構成を示す.

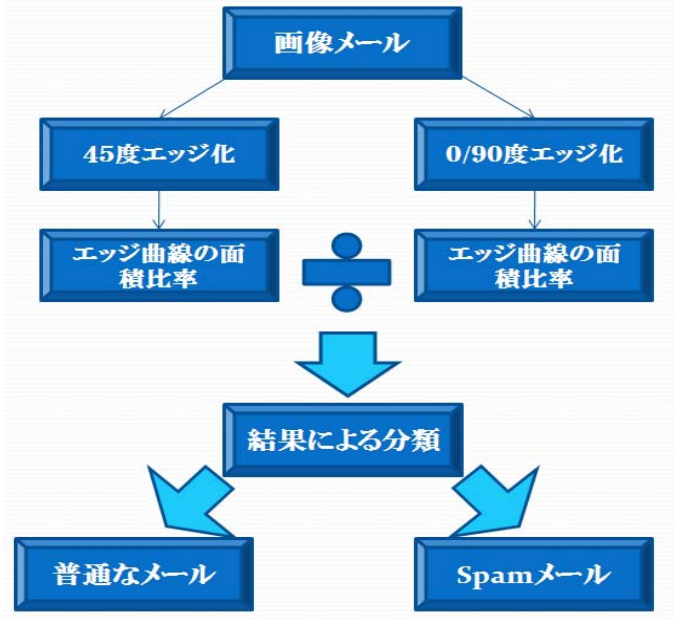

図 4　フィルタの構成

3.4　実験例

風景や人物は内容としての自然画像は自然的ノイズが多いので，45 度エッジ化の結果によるエッジ曲線の面積比率は通常 0/90 度エッジ化のエッジ曲線の 3 倍以上であ

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

る．一方，画像 spam は文字数が多ければ多いほど，45 度エッジ化のエッジ曲線は 0/90 度エッジ化のエッジ曲線に近くなる．それは，文字にノイズがないので，45 度エッジ検出と 0/90 度エッジ検出に近い結果が出てくるからである．

図 5　実験例

図 6　実験データ

図 5 の様に，自然画像では，45 度 sobel オペレータエッジ検出は画面の輪郭だけではなく，天井や部屋のコーナーに光度及び材質などの変化をエッジとして検出した．一方，水平垂直 sobel オペレータエッジ検出は，輪郭のみをエッジとして検出した．Spam の方では，材質及び光度の変化は少ないので，45 度と水平垂直 sobel オペレータエッジ検出の結果がほとんど同じである．

実験例の実際のデータを図 6 に示す．自然画像の 45 度エッジ検出のエッジ曲線の面積は，全画面の 20.57%を占めていて，水平垂直エッジ検出のエッジ曲線は，全画面の 5.24%を占めている．45 度の結果は，水平垂直の結果の 3.9247 倍である．一方，画像 spam の 45 度エッジ検出のエッジ曲線は，全画面の 12.08%を占め，水平垂直のエッジ曲線は，8.66%を占めている．結果の比率は，0.1208/0.0866=1.3951 である．

3.5　実験結果及び分析

本研究では，前文に述べた四種類の画像：画面のみ ham，文字あり ham，文字のみ spam と画面あり spam により，各種類に数十枚を選び，実験を行った．その結果を表 1 のように示す．

表 1　4 種類画像の実験データ

画面のみ ham			文字あり ham			画面あり spam			文字のみ spam		
水平垂直	45°	比率	水平垂直	45°	比率	水平垂直	45°	比率	水平垂直	45°	比率
0.0586	0.2648	4.5220	0.0599	0.1119	1.8668	0.0408	1.9079	1.9079	0.0318	0.0345	1.0103
0.0546	0.2993	5.4849	0.0565	0.1181	2.0899	0.1059	0.1870	1.7649	0.0780	0.0748	0.9591
0.0234	0.1321	5.6409	0.0495	0.2382	4.8169	0.0734	0.0938	1.2775	0.0481	0.0443	0.9219
0.0988	0.2660	2.6922	0.0984	0.1660	1.6862	0.1150	0.1457	1.2674	0.0721	0.0706	0.9795
0.0648	0.2272	3.5051	0.0559	0.1193	2.1338	0.0787	0.1757	2.2325	0.0756	0.0755	0.9978
0.0763	0.2165	2.8360	0.0316	0.0781	2.4705	0.0951	0.1770	1.8610	0.0352	0.0328	0.9330

表 1 はデータの一部である．図 7 に，45 度と水平垂直の比率の結果による分析を示す．

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

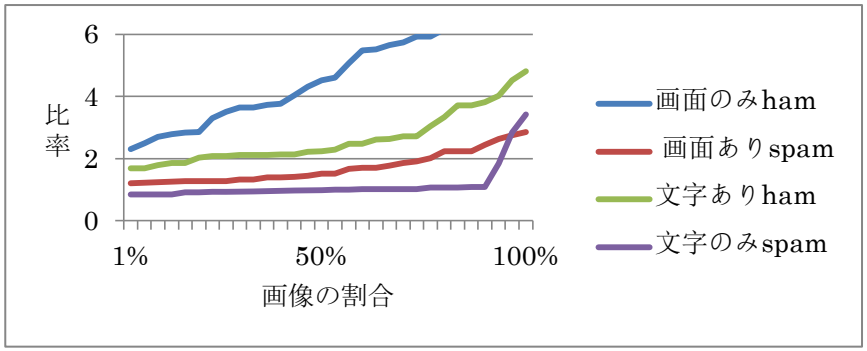

図 7　4 種類画像の実験結果

図 7 の様に, 画面のみ ham は他の種類の画像より高い比率を持っている. すなわち, 自然画像の 45 度エッジ検出のエッジ曲線の面積は水平垂直エッジ検出のエッジ曲線の面積の通常 4 倍以上である．一方，多数の文字のみ spam は，45 度エッジ検出と水平垂直エッジ検出の曲線はほとんど同じ面積であり, 通常の比率は 1 である. その他, 画面あり spam と文字あり ham は文字数の違いによって分かれている.

これらのデータにより，我々は以下の様な結論を得た.

- 自然画像は通常 4 倍以上の一番高い比率を持っているが，人工ノイズを入れた spam も高い比率結果を得られる.
- 比率結果は 1 の周りの画像は文字のみ spam のみである.
- 画面あり spam と文字あり ham は文字数の変化によって各自に不安定な比率を持っているので判別は難しいだが，フィルタの策略により閾値を調整し多重フィルタで判別することができる. すなわち我々が提案したフィルタを多重フィルタの最初の一環として文字のみ spam を検出し，残った画像を他のフィルタで処理する.文字のみ spam は全 spam の 80%以上を占めているので，その策略は効率的だと考えられるであろう.

3.6　光学文字認識

光学文字認識（Optical Character Recognition），略称 OCR とは，画像の中に書かれている文字を解析して、テキストに変換する技術である．この方法を使うことに、spammer が画像を使い spam メールを送っても、通常のテキストと同じように，内容をチェックすることができる.

OCR の仕組みは図 8 のように示す.

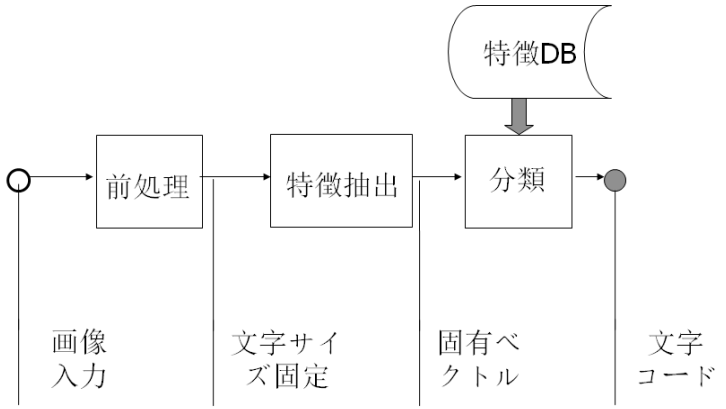

図 8　OCR の仕組み

3.7　多重フィルタ

OCR は，文字の特徴により画像にある文字を認識する技術であり，文字数は多ければ多いほど，処理時間が長くなる.

さらに，図 9 のような写真，またキーワードを含める書類の写真及びスキャンは，OCR で spam に誤判し易い.

図 9　誤判例

そこで，我々は，sobel 及び OCR を用いた多重フィルタを提案する．その仕組みは図 10 のように示す.

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

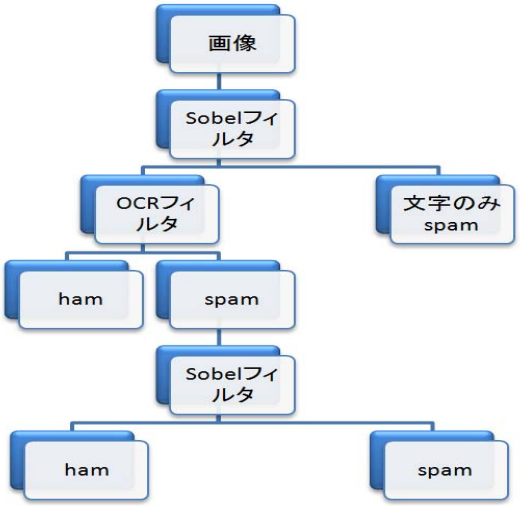

図 10　多重フィルタの仕組み

45 度 sobel オペレータを用いたフィルタは，画像にある文字数が多いほど，検出率が高くなる．また，書類のスキャン及び写真を誤認せず判別することができる．そこで，OCR フィルタで一番時間がかかる文字のみ spam を，まず 45 度 sobel フィルタで検出し，残った画像を OCR で処理する．次，OCR で spam に判定された画像を、再び sobel フィルタで校正を行う．

多くの spam フィルタは，新聞や雑誌の写真及びスキャンを spam に誤認し易い[9]. OCR など image のテキスト部分の内容によって image を分類するフィルタも弱みが存在する． spam 画像をカメラで撮影した写真を，OCR は spam だと判定されるが，実際は ham である.sobel オペレータを用いたフィルタでは，書類のスキャン及び写真を誤認せず判別することができる．OCR で spam に判定された画像を，sobel フィルタで再処理し，誤認された画像を校正する．

図 11 の様に，書類の写真やスキャンなど，自然画像として余白の部分に自然的ノイズは溢れているので，誤判はしない．すなわち，カメラで撮影した写真そしてスキャンは，sobel フィルタで ham に判定される．

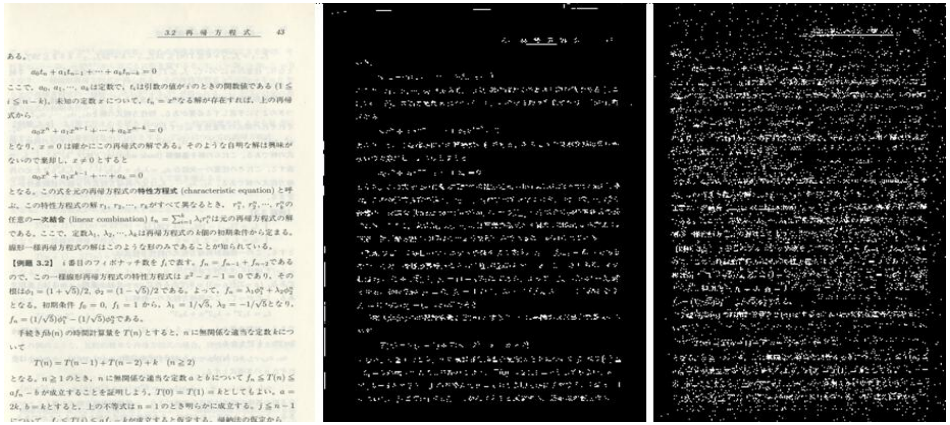

図 11　　書類のスキャン例

4. 評価

4.1 評価システム

全部で N 個のメールがあると仮定すると，spam フィルタの評価システムは以下のように示す.

表 2　評価システム

	Spam メール	Ham メール
Spam と判定される	A	B
Ham と判定される	C	D

$N=A+B+C+D=N_s+N_1$, その中$N_s=A+C$ は実際の spam の数で, $N_1=B+D$ は実際の ham の数，spam フィルタの効果は，以下の五つの基準で評価できる．

- リコール率（Recall）：$R=\frac{A}{A+C}=\frac{A}{N_s}$，spam の検出率のことである，このインジケータは，spam を検出するためのフィルタの能力を反映する．
- 正解率（Precision）：$R=\frac{A}{A+B}$，spam 検出の正解率のことである．正解率は高ければ高いほど，フィルタは ham を spam に誤判する比率が低くなる．
- 精確率（Accuracy）：$Accur=\frac{A+D}{N}$，全てのメール（spam と ham を含めて）に対

Vol.2012-DPS-150 No.18
Vol.2012-CSEC-56 No.18
2012/2/29

しての正解率のこと.

- 誤判率（Error rate）：$Err = \frac{B+C}{N} = 1 - Accur$， 全てのメール（spam と ham を含めて）に対しての誤判率のこと
- 誤認率（miss rate）：$Mis = \frac{B}{B+D}$，フィルタは ham を spam に誤認する比率のことである.

4.2 シミュレーション実験

3 章で提案した 45 度 sobel オペレータと水平垂直 sobel オペレータによるエッジ検出の結果対比よる spam 判別手法また sobel と OCR を用いた多重フィルタをシミュレーション実験で評価する. 評価環境は matlab7.0 とし,インターネットから収集した 3299 枚画像 spam と 2027 枚 ham を対象として実験を行った. 評価インジケータはリコール率と誤認率とする.

表 3　sobel フィルタ閾値の選択によるリコール率と誤認率の変化

閾値	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6
リコール率	45.30%	78.20%	81.90%	85.25%	85.75%	87.64%	88.64%	89.54%	90.12%
誤認率	0.01%	0.01%	0.01%	0.01%	0.01%	0.44%	1.30%	4.86%	8.23%

sobel フィルタ，OCR 及び多重フィルタ各自のリコール率及び誤認率の関係は図 12 に示す，sobel フィルタの閾値は 1.8 とする.

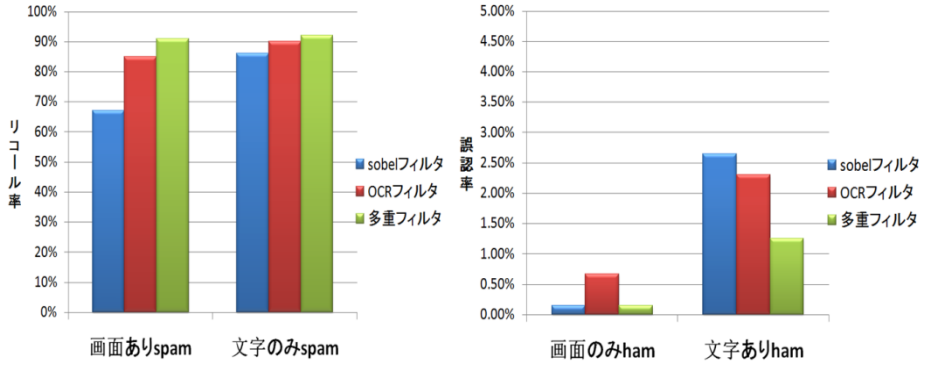

図 12　　三種類フィルタのリコール率と誤認率

これらのデータにより，我々は以下の様な結論を得た.

- sobel フィルタでは閾値は高ければ高ほどリコール率と誤認率が高くなる．閾値は 1.8 の回りとすると，リコール率と誤認率両方からも良好な効果が出る
- 多重フィルタで，リコール率は 90%を超え，効果的に誤判率を低くした.
- 全部の三種類のフィルタは人工ノイズには弱い.

5.　おわりに

本稿では，エッジ検出において，水平垂直 sobel オペレータと 45 度 sobel オペレータによる処理の結果曲線の面積を比較し，その比率によって，spam を判別する手法を提案し，それを基づいて，OCR を用いた多重フィルタを提案した.

今後の課題としては，人工ノイズに対する改進及びフィルタの学習機能の添加を行っていく予定である.

参考文献

1)　中国インターネット協会 spam 対策センター，http://www.anti-spam.cn/
2)　Guy Di Mattina：Spam and Open Relay Blocking System，A thesis submitted to the School of Information Technology and Electrical Engineering The University of Queensland（2003）
3)　Mori Tatsuya：PrBL: Probabilistic BlackList for E-mail Spammers, IEICE technical report 108（457），pp.15-20（2009）
4)　Mori Tatsuya：On the use and misuse of E-mail sender authentication mechanisms，IEICE technical report 110（115），pp.101-106（2010）
5)　Sugii Manabu，Matsuno Hiroshi：Decision Tree Representation of Spam Mail Features by Machine Learning，IPSJ SIG Notes 2007（16），pp.183-188（2007）
6)　Wang Zhan，Hori Yoshiaki，Sakurai Kouichi：A Design of Image-based Spam Filtering Based On Textual and Visual Information，IPSJ SIG Notes 2008（21），pp.279-284（2008）
7)　Bhaskar Mehta，Saurabh Nangia，Manish Gupta：Detecting Image Spam using Visual Features and Near Duplicate Detection，WWW 2008 / Refereed Track: Security and Privacy – Misc（2008）
8)　Wang Zhan，Hori Yoshiaki，Sakurai Kouichi：A Design of Image-based Spam Filtering Based On Textual and Visual Information，IPSJ SIG Notes 2008（21），pp.279-284（2008）
9)　Yang Gao, Ming Yang, Xiaonan Zhao：Image Spam Hunter, ICASSP 2008, pp.1765-1768（2008）