深層学習で分類するステガノグラフィのロバスト性評価

繁田大輝[1,a)]　森田光[1]

概要：著者らは，ステガノグラフィ構成法として，カバーメディアを必要としない直接ステゴメディアを生成する方法を提案している．候補とするステゴメディアを分類器で分類し，分類結果を埋め込み情報として扱うことでカバーメディアを必要としないのである．分類器として深層学習を用いた手法 [1], [3] とハッシュ関数を用いた手法 [2] を提案しているが，両者を比較すると，深層学習を用いた手法は速度では劣るがロバスト性に優れていると期待できる．そこで，ステゴメディアが画像の場合，どの程度の改ざんで分類結果が変わり，正しく復号きるかを課題とし，本稿では，深層学習を用いた手法におけるロバスト性を定量的に評価する．

キーワード：ステガノグラフィ，情報ハイディング

Robustness evaluation on deep learning in steganography

Taiki Shigeta[1,a)]　Hikaru Morita[1]

Abstract: The authors have proposed methods that don't need any covermedium and directly classifies stegomedium. They don't require covermedia by classifying stegomedium with a classifier and treating the result as embedded information. As a result of comparing methods using deep learning and hash function as a classifier, a method using deep learning is inferior in speed but can be expected to good in robustness. If a stegomedium including is selected image, how much it's classification result is changed. In this paper, the authors quantitatively evaluate the robustness of the method using deep learning.

Keywords: steganography, information hiding

1. はじめに

ステガノグラフィは，第三者に気づかれずに情報を共有する技術である．例えば，**図 1** で示すように，オリジナルのカバーメディアに情報を埋め込み，ステゴメディアに変換する．カバーメディアとステゴメディアのデータが異なるため，データの違いによってステゴメディアに埋め込んだ情報の復元が可能となる．見た目からは，第三者はカバーメディアとステゴメディアの違いは大きくなく，埋め込まれた情報は鍵 k 無しでは検出することが困難になる．

しかし，従来のステガノグラフィでは，カバーメディア，ステゴメディアの両データが流出すると，データを比べることで埋め込み情報を解析される恐れがあった．この問題を解決するために著者らは，カバーメディアを必要としない，直接ステゴメディアを分類するステガノグラフィ構成法を提案した (**図 2**)．ステゴメディア生成関数 s において分類器を用いてステゴメディアを決定し，復号関数 d で同じ分類器を用いることで，カバーメディアなく情報を共有することを可能としている．提案方法では，分類器として

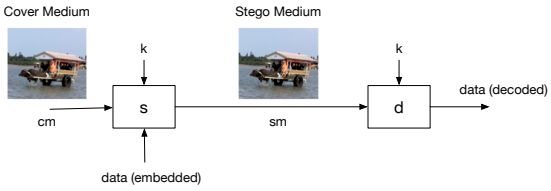

図 1　従来のステガノグラフィ

[1] 神奈川大学大学院
Graduate School of Kanagawa University
[a)] r201870120yv@jindai.jp

深層学習を用いる手法とハッシュ関数を用いた手法があり，ハッシュ関数を用いた手法は，速度において勝るが，ステゴメディアが 1 ビットでも変わると正確に復号することができず，ロバスト性に欠けることが分かる．自然発生なノイズ，または攻撃者による改ざんなどでステゴメディアに何らかの変化が加わり，ロバスト性が欠けた場合，埋め込み情報が正確に復号できない．深層学習は性質として，ロバスト性が期待できる．本稿では，ステガノグラフィーにおいて深層学習を用いた提案手法におけるロバスト性の評価をする．

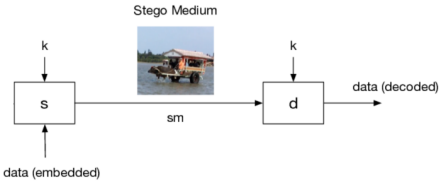

図 2　カバーメディアのないステガノグラフィ

本稿の構成は，2 章でカバーメディアを必要としないステガノグラフィ構成法について説明する．3 章で評価実験をする．4 章で考察し，5 章でまとめる．

2. カバーメディアを必要としないステガノグラフィ構成法

カバーメディアを必要としないステガノグラフィは，分類器の分類結果を埋め込み情報として扱う．そのため，メディアに直接情報が埋め込まれないため，従来のステガノグラフィにあった，カバーメディア，ステゴメディアの両データがペアで流出した際に埋め込み情報が解析される恐れがなくなる．例えば，「231」という情報を相手と共有したい場合，**図 3** に示すように，関数 s 内の分類器により，それぞれ「2」,「3」,「1」に分類されるメディアをステゴメディアが格納されたデータベースから選択し，相手に送信する．受信者も同じ分類器を用い，「231」という情報を復号することができる．

2.1 ステゴメディアの選択と復号

ここでは，ステゴメディアを生成する関数 s と復号関数 d について説明する．

図 4 で示す，関数 s では，カバーメディアに代わりステゴメディア候補を用いる．ステゴメディア候補は，ステゴメディアデータベースに格納されており，そこから分類器の分類結果と自分の埋め込みたい情報が一致するようなメディアを選択する．詳しい手順は以下の通りである．

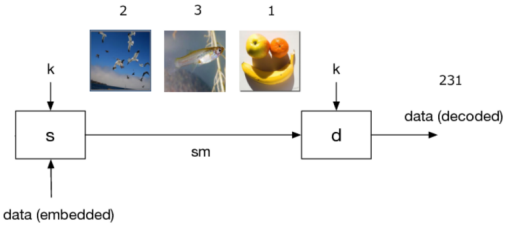

図 3　カバーメディアのないステガノグラフィにおける情報伝達の例

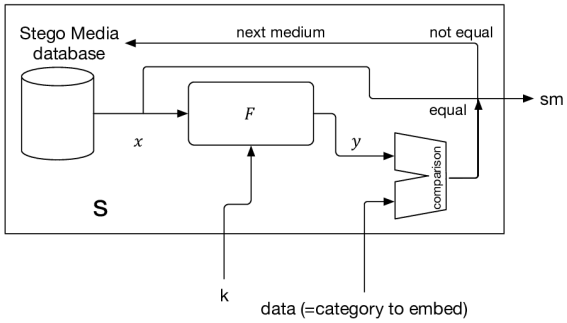

図 4　ステゴメディア生成関数 s

(1) まず，ステゴメディアデータベースから無作為にステゴメディア候補 x を選択する．
(2) 式 1 に示すように x と共通鍵 k を分類関数 F に入力し分類結果 y を求める．(n は分類数である)
(3) 求めた y と自分が埋め込みたい $data$ を比較する．
(4) $y = data$ となったら入力 x をステゴメディア sm とする．
(5) $y \neq data$ となった場合，x の選択からやり直し，sm が決定するまでこの工程を繰り返す．
(6) ステゴメディア sm が決定したら，sm を相手に送信する．
(7) 受信者は**図 5** に示すように，関数 F に受け取った sm と共通鍵 k を入力として用いることで埋め込み情報 $data$ を復元する．

$$F(x,k) = y \qquad (y \in \{1,2,\ldots,n\}) \qquad (1)$$

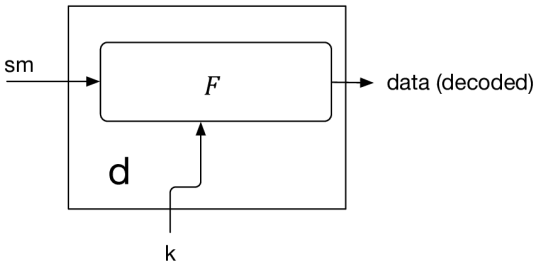

図 5　埋め込み情報復号関数 d

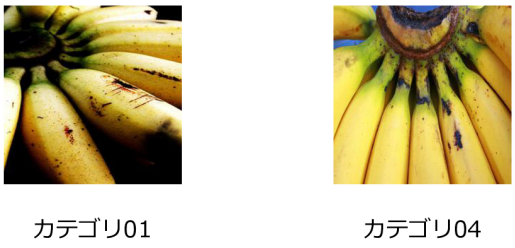

図 7　生成したステゴメディア（例）

2.2　深層学習を用いた手法

著者らは，関数 F に深層学習とハッシュ関数を用いた手法を提案している．ここでは，深層学習を用いた手法について説明する．

深層学習を用いた手法では，**図 6** に示すように，関数 F で深層学習を用い，共通鍵 k として深層学習の学習重みなどの深層学習を構成するのに必要なパラメータを用いている．受信者は共通鍵 k を用いて，関数 F で用いた深層学習と同等の深層学習を構成し，埋め込み情報 $data$ を復元する．

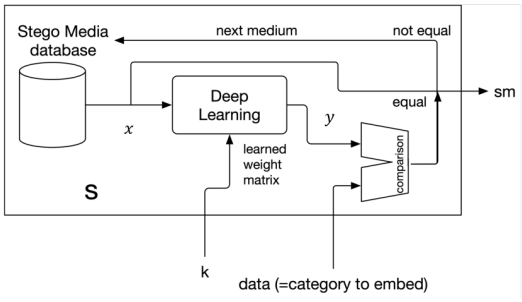

図 6　深層学習を用いた手法

図 7 は実際に分類されたステゴメディア sm である．ここでは，深層学習の学習データとステゴメディアデータベースに格納されているデータは関係のないデータを用いるため，図 7 のように人の判断では，何を基準に分類しているか判別できないため，共通鍵 k の秘密が保たれている限り，埋め込み情報は解読されない．

3.　実験

ハッシュ関数を用いた手法では，ロバスト性がないことが分かっている．しかし，深層学習は性質としてロバスト性を持つ．そこで，本提案方法に対してもロバスト性があるかを確かめる評価実験を行う．本実験で用いる深層学習は人の顔の画像を学習データとし，ステゴメディアとして，人の顔とは関係のない画像を用いている．また，分類数は 01〜04 までの 4 分類である．本実験の環境は以下の通りである．

表 1　提案の実験環境

OS	macOS High Sierra
深層学習ライブラリ	Chainer 4.1.0
CPU	Intel Core i5-7500 3.40GHz

3.1　ガウシアンノイズによる実験

ここではノイズに対する評価の実験を行う．ガウシアンノイズは，確率変数 z に対して，平均 μ，標準偏差 σ からなる式 2 のガウス関数で与えられる確率分布に従ってピクセルの要素ごとにノイズを発生させる．σ に基づく乱数を生成し，原画と加算合成することでノイズを付加した画像を生成する．

$$f(z) = \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{(z-\mu)^2}{2\sigma^2}} \tag{2}$$

本実験では，Python を用いて，ノイズ付き画像に得られたノイズを元のステゴメディアの画像に付加した．RGB 画像を用いるため，1 ピクセルの RGB それぞれ 256 階調を持ち，各 1 バイトを持つ．ピクセルごとの RGB それぞれの要素に独立で，平均 μ，標準偏差 σ，でガウシアンノイズを確率変数 z に発生させる．ここでは，平均 $\mu = 0$ とし，標準偏差 σ の値を増加させ，ノイズ強度を上げる．ノイズの強度は，**図 8** に示すように，$\sigma = 20, 50, 100, 150, 300$ のように増加させる．

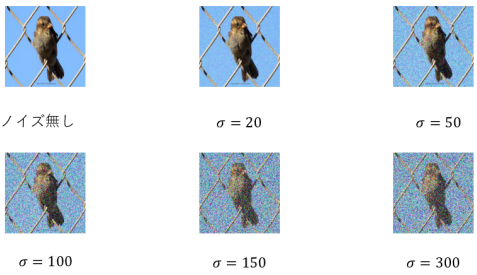

図 8　ノイズ画像

σ =20〜300 のノイズ強度に対して，ステゴメディア 4 カテゴリに，それぞれ 100 枚ずつ深層学習で分類する．分類結果を**表 2** に示す．

表 2　ノイズの強度別の正解数

	カテゴリ			
σ	01	02	03	04
20	99/100	96/100	99/100	100/100
50	78/100	81/100	83/100	99/100
100	63/100	33/100	36/100	100/100
150	66/100	11/100	6/100	100/100
300	3/100	0/100	0/100	100/100

3.2　回転による実験

ここでは，画像の回転に対する耐性実験をする．**図 9** のように，画像の角度を，15 度，30 度，45 度，60 度，90 度，135 度，180 度の順に回転させ，それぞれの角度，カテゴリに対して 100 枚ずつ分類する．その結果を**表 3** に示す．

図 9　回転画像

表 3　回転の角度別の正解数

	カテゴリ			
angle	01	02	03	04
15	82/100	60/100	68/100	90/100
30	57/100	28/100	65/100	86/100
45	43/100	13/100	66/100	80/100
60	31/100	12/100	52/100	85/100
90	63/100	18/100	39/100	60/100
135	41/100	10/100	49/100	67/100
180	65/100	15/100	31/100	52/100

4.　考察

4.1　ノイズに対する耐性実験

ノイズに対する耐性実験では $\sigma = 20$ では 394/400 枚で 98.5% の高精度で復号が可能であることが分かった．また，$\sigma = 50$ では 341/400 で 85.3% の精度で復号が可能であった．しかし，$\sigma = 100$ では，01 カテゴリでは 6 割の精度を保っているが，02，03 カテゴリでは 4 割を切る精度まで低下している．$\sigma = 150$ では，02，03 カテゴリは 1 割かそれ以下まで低下し，$\sigma = 300$ では，01〜03 カテゴリで正解が 1% という結果となった．カテゴリ 04 において，**表 4**，**表 5**，**表 6** は $\sigma = 20, 150, 300$ の 100 枚のカテゴリ内訳である．また，**表 7** は，通常 01 に分類されるメディアにノイズ強度ごとの深層学習が出したカテゴリ出現確率を示している．ノイズが増えるにつれて，どのカテゴリもカテゴリ 04 に分類される割合が多くなり，分類の正解率も 04 カテゴリの値が増加している．

表 4　$\sigma = 20$ におけるカテゴリ別の分類先内訳

	分類先カテゴリ			
正解カテゴリ	01	02	03	04
01	99	0	0	1
02	1	96	0	3
03	0	0	99	1
04	0	0	0	100

表 5　$\sigma = 150$ におけるカテゴリ別の分類先内訳

	分類先カテゴリ			
正解カテゴリ	01	02	03	04
01	34	0	0	66
02	5	11	2	82
03	28	1	6	65
04	0	0	0	100

表 6　$\sigma = 300$ におけるカテゴリ別の分類先内訳

	分類先カテゴリ			
正解カテゴリ	01	02	03	04
01	3	0	0	97
02	1	0	0	99
03	2	0	0	98
04	0	0	0	100

表 7 01 画像のノイズ強度別のカテゴリ出現確率例

	カテゴリ			
σ	01	02	03	04
0	0.99	5.0×10^{-5}	1.0×10^{-4}	5.7×10^{-4}
20	0.99	6.4×10^{-5}	9.0×10^{-5}	1.0×10^{-3}
50	0.99	5.7×10^{-5}	8.4×10^{-5}	2.4×10^{-3}
100	0.85	7.4×10^{-4}	1.9×10^{-3}	0.15
150	0.02	7.2×10^{-5}	4.7×10^{-5}	0.98
300	5.7×10^{-3}	6.5×10^{-5}	2.8×10^{-5}	0.99

4.2 画像の回転に対する耐性実験

画像の回転に対する耐性実験では，カテゴリごとに耐性が大きく異なることが分かった．カテゴリ 02 では，15 度までは正解率 6 割まで保っているが，30 度からは 3 割を切り，それ以降は 1 割から 2 割の正解率となっている．また，カテゴリ 01 では，15 度で 8 割を超え，30 度，90 度，180 度では 6 割程度の正解率を保っているが，それ以降は 4 割かそれ以下の正解率となっている．03 カテゴリは 45 度まで 6 割の正解率を保ち，それ以降は 5 割かそれ以下の正解率となっている．04 カテゴリでは，180 度以外で 6 割以上の正解率を示し，180 度でも 5 割を超える結果となった．全てのカテゴリで 6 割以上の正解率があるものは角度 15 度だけとなっているため，本実験では回転への耐性は 15 度までの耐性しかないといえる．今回の実験では．画像を回転させた際に，画像に黒い部分が残っている．傾けた状態で黒い部分が残らないように処理をすると，深層学習の分類が不可能だったため，このような処理をするに至った．そのため，本実験だけでは，ロバスト性を判断するには不十分だと言える．

5. まとめ

本稿では，カバーメディアを必要としないステガノグラフィの深層学習を用いた手法に対し，ロバスト性の評価を行った．その結果，ガウシアンノイズを与えた画像に対しては $\sigma = 50$ のノイズまで正解率の平均が 8 割程度あり，人の目に違和感がない画像まではロバスト性を維持できることが分かった．回転を加えた画像に対しては，実験データが不十分ながら，角度 15 度までの回転には，ある程度のロバスト性があることが分かった．以上から，深層学習を用いた手法には，人の目で判断できないノイズに対して十分に有効なロバスト性を持つと言える．そのため，分類の変わるような大きなノイズは人が判断可能であり，自然的なノイズのような，人の目で判断できないノイズに対しては分類が変わらない程度のロバスト性を持つため有効な手法である．

参考文献

[1] 繁田大輝, 森田 光, “深層学習による画像のステガノグラフィ”, ISEC, 信学技報, vol. 118, no. 478, March. 2019.

[2] 繁田大輝, 森田 光, “メディア選択にハッシュ関数を用いるステガノグラフィ”, SCIS2019, 4C2-5, Jan. 2019.

[3] 増井孝之，繁田大輝，森田光，“深層学習による文字列のステガノグラフィ”，ISEC，信学技報, vol. 118, no. 478.，March. 2019.

[4] 坂井麻守, 森田光, “深層学習の分類による情報ハイディング埋め込み方法”，SCIS2018, 3D1-2, Jan. 2018.