ステガノグラフィを用いたプライバシ保護顔認証とその安全性評価

神津 岳志[1,a)]　河合 洋弥[1,b)]　伊藤 康一[1,c)]　青木 孝文[1,d)]

概要

SNS などの利用拡大に伴ってインターネットから簡単に顔画像を収集することができる．第三者が顔画像を使って顔認証システムに対してなりすまし攻撃を行う危険性がある．顔画像を公開しつつ，なりすまし攻撃を防ぐために，プライバシ保護を備えた顔認証が求められている．本論文では，深層学習を用いたステガノグラフィにより顔画像を任意の画像に埋め込んで顔認証を行う手法を提案する．埋め込んだ画像は，顔認証に利用できるだけではなく，インターネット上に安全に公開できる．顔画像の公開データセットを用いた性能評価実験と安全性評価実験を通して提案手法の有効性を示す．

1. はじめに

パスワードや鍵に代わる認証方式として，個人の身体的・行動的特徴を用いる生体認証が注目されている [1]．生体認証の 1 つである顔認証は，一般的なカメラを使用して非接触で顔画像を撮影できるという特長があるため，スマートフォンや PC などのユーザ認証，空港の出入国管理等で実用化されている．一方で，SNS などのサービスの普及により，インターネット上から顔画像を収集することが容易になり，悪意のある第三者（攻撃者）が顔画像を収集して顔認証システムに対してなりすまし攻撃を行う恐れがある．

なりすまし攻撃を防ぐ方法としてインターネット上に公開されている顔画像をアバター画像に置き換える方法がある．本人の顔画像ではなくなるためなりすまし攻撃を防ぐことはできるが，アバター画像が顔画像の代替であるため本人との間で認証を行うことができない．セキュアな顔認証を実現するために，本人と認証することができ，プライバシが保護されているような顔画像に代わる認証媒体が求められている．著者らが知りうる限りでは，現在までに，上記のような観点からの研究が報告されていない．

[1] 東北大学 大学院情報科学研究科
[a)] kozu@aoki.ecei.tohoku.ac.jp
[b)] hiroya@aoki.ecei.tohoku.ac.jp
[c)] ito@aoki.ecei.tohoku.ac.jp
[d)] aoki@ecei.tohoku.ac.jp

本論文では，ステガノグラフィ [2] を用いて任意の画像（以下，カバー画像）に顔画像を埋め込んで顔認証を行う手法を提案する．ステガノグラフィとは，ある情報を別の情報に埋め込んで秘匿する技術であり，一般的に用いられている暗号化とは異なり，情報が秘匿されていること自体を判別困難にする．最近では，Convolutional Neural Network (CNN) を用いた手法 [3], [4] が提案され，従来使われていたステガノグラフィと比べて，多くの情報を 1 枚の画像に埋め込むことができるようになっている．Deep Steganography [4] に基づいて顔画像をカバー画像に埋め込むことで個人情報を秘匿化し，利用者のプライバシを保護する．顔画像が埋め込まれた任意の画像（以下，ステゴ画像）から顔画像を再構成することなく顔特徴量を抽出することができるように学習させる．そのため，インターネット上にステゴ画像を公開することができる．2 つの大規模な顔画像の公開データセットを用いた性能評価実験およびステゴ画像の安全性評価実験を通して提案手法の有効性を示す．

2. プライバシ保護顔認証

Deep Steganography [4] は，入力画像をカバー画像に埋め込んでステゴ画像を生成する Hiding Network (HN) とステゴ画像から入力画像を復元する Revealing Network (RN) で構成される．これに対して，提案手法では，RN ではなく，顔特徴量を抽出する Extraction Network (EN) で構成する．提案手法で用いる CNN のネットワークアーキテクチャを図 1 に示す．HN では，RGB カラーの顔画像 S と RGB カラーのカバー画像 T を結合した合計 6 チャンネルのデータを入力し，カバー画像と見た目が同じであるステゴ画像 C を出力する．HN には，U-Net [5] に基づくエンコーダデコーダネットワークを用いる．オリジナルの U-Net とは違い，カバー画像を後ろの層に伝搬するために，ResNet [6] で用いられている ResBlock を用いる．EN では，ステゴ画像 C から顔画像 S の特徴量を抽出する．EN には，ResNet18 [6] に基づくネットワークアーキテクチャを用い，Skip Connection においてチャンネルにアテンションをかけるために Squeeze-and-Excitation [7] を用

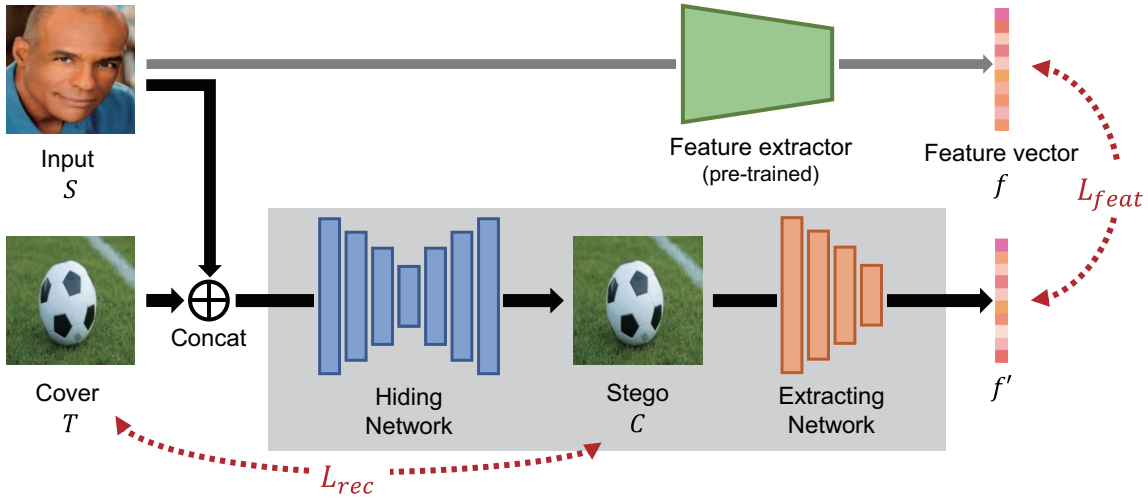

図 1 提案手法で用いる CNN のネットワークアーキテクチャの概要

いる．

EN で出力される特徴量は，顔特徴抽出器で顔画像 S から抽出される特徴と同様になるように学習される．本論文では，顔認証の標準的な手法として用いられている FaceNet [8] および ArcFace [9] を学習のための顔特徴抽出器として用いる．両手法とも特徴空間においてクラス内分散が小さくなるように，クラス間分散が大きくなるように CNN を学習させる深層距離学習を用いている．FaceNet では，Triplet Loss [10] を用いている．Triplet Loss は，基準とする特徴量に対して，同じクラスの特徴量とのユークリッド距離を近づけるとともに，異なるクラスの特徴量とのユークリッド距離を遠ざけるように学習させる損失関数である．ArcFace は，基準とする特徴量に対して，同じクラスの特徴量とのコサイン類似度が高くなるように，かつ，異なるクラスの特徴量とのコサイン類似度が低くなるように学習させる損失関数である．

HN と EN の学習では，2 つの損失関数を用いる．1 つ目は，カバー画像 T とステゴ画像 C が等しくなるように学習させるための再構成損失 L_{rec} であり，次式で定義される．

$$L_{rec} = \frac{1}{N}\sum^{N}(T - C)^2 \tag{1}$$

ここで，N はバッチサイズを示す．もう 1 つは，EN から出力される顔特徴量 f' が FaceNet または ArcFace から出力される顔特徴量 f と同じになるように学習させる特徴量損失 L_{feat} であり，次式で定義される．

$$L_{feat} = \begin{cases} \frac{1}{N}\sum^{N}(f - f')^2 & \text{(FaceNet)} \\ 1 - \cos(f, f') & \text{(ArcFace)} \end{cases} \tag{2}$$

特徴量損失 L_{feat} は，学習に使用する顔特徴抽出器によって場合分けを行う．全体の損失関数 L は，再構成損失と特徴量損失の和として次式のように定義される．

$$L = L_{rec} + \beta \cdot L_{feat} \tag{3}$$

ここで，β は，損失の重みを決定するハイパーパラメータである．

3. 性能評価実験

本実験では，CNN の学習に，CelebFaces Attributes (CelebA) Dataset [11] を用いる．CelebA[*2] は 10,177 人から撮影された 202,599 枚の顔画像からなるデータセットであり，199,599 枚を学習用に，残りの 3,000 枚を検証用に用いる．なお，本実験において，全ての顔画像は MTCNN [12] を用いて顔領域を抽出し，256 × 256 画素にリサイズさせる．学習時のバッチサイズは 32 とし，バッチを 2 分割したうちの一方を画像 S，もう一方を画像 T として使用する．最適化手法には Adam [13] を用い，検証用データに対する損失に基づいて，学習率を 10^{-5} から動的に調整しながら，150 エポック学習させる．損失関数のハイパーパラメータ β の値は 2.0 とする．学習時には，入力画像の左右をランダムに反転するデータ拡張を追加する．

性能評価に Labeled Faces in the Wild (LFW) [14] と CASIA-WebFace Dataset (CASIA) [15] の 2 つの大規模な顔画像データセットを用いる．LFW は，インターネット上で収集された 5,749 人分の計 13,233 枚の顔画像からなるデータセットである．LFW で推奨されている実験プロトコルに従って，本人ペアと他人ペアの各 3,000 ペアを用いる．CASIA は，10,575 人の 494,414 枚の顔画像からなるデータセットである．全顔画像うち，MTCNN によって顔検出ができた 490,740 枚を用いる．各人からランダムに 2 ペア抽出して作成された 21,150 ペアを本人ペアとし，各人に対して異なる人をランダムに 2 回選択して作成された 21,150 ペアを他人ペアとして用いる．性能評価では，各評価用データセットの顔画像を図 2 に示すカバー画像 T に HN を用いて埋め込み，ステゴ画像 C を生成する．そ

[*2] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

図 2 性能評価実験で使用するカバー画像

の後，EN を用いてステゴ画像 C から抽出された顔特徴量 f' と顔画像 S を FaceNet に入力し得られた特徴量 f とのユークリッド距離を認証精度の評価に用いる．また，顔画像 S を ArcFace に入力した場合は，得られた f と f' とのコサイン類似度を認証精度の評価に用いる．認証精度のベースラインとして，カバー画像 T への埋め込みを行わず，顔画像 S から FaceNet または ArcFace を用いて抽出された特徴量 f を使用した場合の認証精度を求める．また，Deep Steganography を用いて生成したステゴ画像から抽出した顔画像 S を FaceNet または ArcFace に入力した場合の認証精度も求める．

生成されたステゴ画像の例を図 3 に示す．インターネット上から著作権フリーで公開されている画像[*3] をカバー画像として用いた．提案手法を用いることでカバー画像と見分けが付かないステゴ画像が生成できていることが確認できる．表 1 に特徴抽出器と評価用データセットごとの各手法の認証精度を示す．性能評価指標として，全体のペアに対して本人ペアあるいは他人ペアを正しく推定できた割合を示す Accuracy を用いる．さらに，False Accept Rate (FAR) が 0.01 の時の False Reject Rate (FRR) と Equal Error Rate (EER) を用いる．FAR は他人ペアを本人ペアと誤って受け入れた割合を示し，FRR は本人ペアを他人ペアとして排除した割合を示す．EER は，FAR と FRR の値が一致するときのエラー率である．提案手法は，プライバシ保護を備えない従来の顔認証手法や顔画像を再構成する Deep Stegnaography の認証精度と遜色ない結果となっていることが確認できる．

4. 安全性評価実験

提案手法を用いてプライバシ保護を適用したステゴ画像の安全性を評価する．まず，提案手法で生成されたステゴ画像から埋め込んだ顔画像が抽出できるかどうかを確認する．インターネットなどから簡単に取得できる画像をカバー画像に使用した場合，生成されたステゴ画像とカバー画像との差分から埋め込まれた顔画像を抽出できる可能性がある．ここでは，カバー画像として顔画像と黒一色の画像を用いてステゴ画像との差分を確認する．図 4 に顔画像，カバー画像，ステゴ画像とカバー画像の差分を 20 倍と 40 倍に増幅させた結果を示す．どちらのカバー画像を用いたときも埋め込まれた顔画像は確認できなかった．

[*3] https://www.photo-ac.com

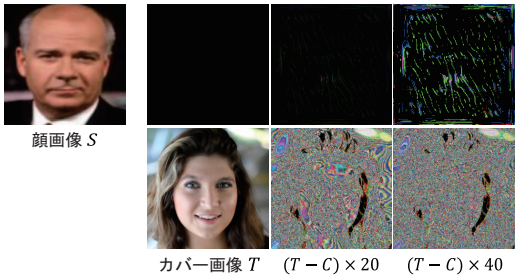

図 4 カバー画像とステゴ画像の差分

表 2 異なる学習データセットを用いた時の認証精度

HN	EN	Accuracy	FRR@FAR=0.01	EER
IMDb	IMDb	0.9513	0.1243	0.04767
CelebA	IMDb	0.4998	0.9990	0.5969
IMDb	CelebA	0.5000	0.9963	0.5960

最後に，提案手法を別のデータセットで学習させたときの EN を用いてステゴ画像から顔特徴量が抽出できるかどうかを確認する．第三者によって生成された EN を用いて顔特徴量を抽出できる場合，抽出された顔特徴量がなりすまし攻撃に利用される恐れがある．検証のため，IMDb dataset [16] を用いて前節と同じ条件で提案手法を学習させる．なお，顔特徴抽出器は FaceNet のみを用いる．IMDb は 20,284 人の 460,723 枚の顔画像からなるデータセットであり，MTCNN によって顔領域の取得ができた 427,326 枚を学習用に，5,000 枚を検証用に用いる．表 2 に IMDb を用いた時の提案手法の認証精度と HN と EN で異なる学習データセットを用いた時の認証精度を示す．表 2 より，HN と EN で異なるデータセットを用いた場合，正しく顔特徴量が抽出できず，認証が行えないことがわかる．

5. まとめ

本論文では，ステガノグラフィによるプライバシ保護を備えた顔認証手法を提案した．大規模な顔画像の公開データセットを用いた性能評価実験を通して，提案手法の有効性を実証した．また，安全性評価実験を通してプライバシ保護を適用したステゴ画像の安全性について実証した．今後は，提案手法の認証精度の改善を行うとともに，学習データセットと認証精度の関係性について評価を行う予定である．

参考文献

[1] A.K. Jain, P. Flynn, and A.A. Ross, Handbook of Biometrics, Springer, 2008.

[2] G.J. Simmons, “The prisones’ problem and the subliminal channel,” Advances in Cryptology (Proc. CRYPTO ’83), pp.51–67, June 1983.

[3] D. Volkhonskiy, I. Nazarov, B. Borisenko, and E. Burnaev, “Steganographic generative adversarial networks,” Proc. NIPS 2016 Workshop on Adversarial Training, pp.1–8, March 2016.

図 3 提案手法で生成したステゴ画像例

表 1 各データセットを用いた場合の認証精度

Facial Feature Extractor	Dataset	Method	Accuracy	FRR@FAR=0.01	EER
FaceNet	LFW	FaceNet	0.9737	0.04733	0.02600
		Deep Steganography	0.9597	0.09500	0.03975
		Proposed	0.9513	0.1403	0.04700
	CASIA	FaceNet	0.8486	0.4437	0.1561
		Deep Steganography	0.8462	0.4856	0.1590
		Proposed	0.8343	0.5178	0.1697
ArcFace	LFW	ArcFace	0.9707	0.06933	0.03033
		Deep Steganography	0.9422	0.1856	0.05833
		Proposed	0.9340	0.2470	0.06707
	CASIA	ArcFace	0.9007	0.3041	0.1035
		Deep Steganography	0.8587	0.4574	0.1457
		Proposed	0.8538	0.4938	0.1495

[4] S. Baluja, “Hiding images in plain sight: Deep steganography,” Proc. Advances in Neural Information Processing Systems, vol.30, pp.2069–2079, Dec. 2017.
[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention, pp.234–241, Oct. 2015.
[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.770–778, June 2016.
[7] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.7132–7141, June 2018.
[8] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.815–823, June 2015.
[9] J. Deng, J. Guo, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.4685–4694, June 2019.
[10] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1386–1393, June 2014.
[11] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” Proc. Int’l Conf. Computer Vision, pp.3730–3738, Dec. 2015.
[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multi-task cascaded convolutional networks,” IEEE Signal Processing Letters, vol.23, no.10, pp.1499–1503, April 2016.
[13] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proc. Int’l Conf. Learning Representations, vol.abs/1412.6980, pp.1–15, May 2015.
[14] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Technical Report 07–49, University of Massachusetts, Amherst, Oct. 2007.
[15] D. Yi, Z. Lei, S. Liao, and S.Z. Li, “Learning face representation from scratch,” CoRR, vol.abs/1411.7923, pp.1–9, Nov. 2014.
[16] R. Rothe, R. Timofte, and L.V. Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” International Journal of Computer Vision, vol.126, no.2-4, pp.144–157, Aug. 2018.