ICIP 2019

MultiDrone

ADVERSARIAL FACE DE-IDENTIFICATION

Presenter: A. Tefas

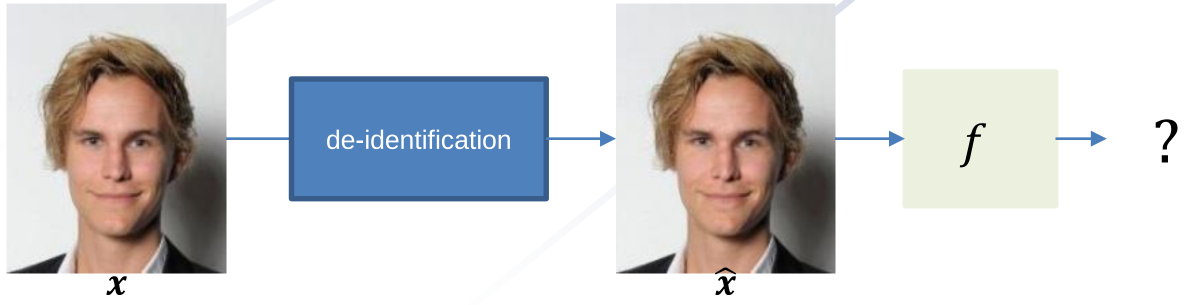
E. Chatzikyriakidis, C. Papaioannidis and I. Pitas
Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Face de-Identification problem

- Face recognition systems f take a facial image $\boldsymbol{x}$ as input and predict its corresponding identity y, $f(\boldsymbol{x}) \rightarrow y$.
- Therefore, **face de-identification** methods aim to alter the original facial image $\boldsymbol{x}$ and produce a de-identified image $\widehat{\boldsymbol{x}}$ that can no longer be identified by face recognition systems, $f(\boldsymbol{x}) \rightarrow ?$.

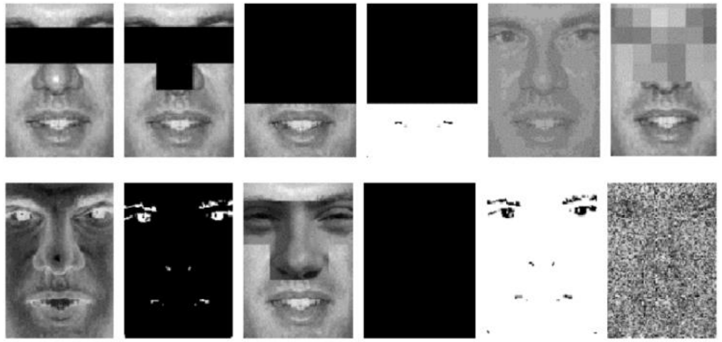


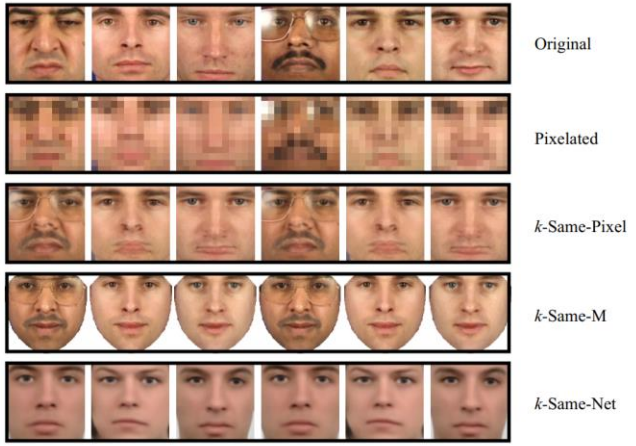
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Motivation - Drawbacks of previous methods

- Privacy protection on images and videos.
- Previous face de-identification methods strongly alter original images.
- De-identified image should retain the original facial image unique characteristics (e.g. race, gender, age, expression, pose).





This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Motivation - Face recognition systems

- Modern face recognition systems are robust to ad-hoc de-identification methods (mask, blur, pixelization, random noise etc.).
- Wide variety of face recognition systems with different internal functionality.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Contribution

- A new face de-identification method that uses adversarial examples.
- A novel penalty term in the objective function.
- Increased misclassification rate (protection) than previous face de-identification methods.
- Minimal image distortion between original and de-identified images.
- The non-identity facial characteristics are preserved in the de-identified images.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Adversarial examples

- Adversarial examples are carefully constructed inputs that result to incorrect classification.
- Let f be a deep neural network classifier trained on a dataset and $\{\boldsymbol{x}_i, y_i\}$ is a dataset entry, with $\boldsymbol{x}_i \in \boldsymbol{X} \subseteq \boldsymbol{R}^n$ being a facial image and $y_i \in \boldsymbol{Y}$ the corresponding ground truth label.
- If $\boldsymbol{x}$ is an instance with ground truth label y, then an adversarial example $\hat{\boldsymbol{x}}$ can be crafted by adding a small perturbation to $\boldsymbol{x}$, so that $f(\hat{\boldsymbol{x}}) \neq y$.
- The added perturbation can be measured as $\boldsymbol{p} = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_p$, where $\|\cdot\|_p$ is the p-norm.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Adversarial attacks

- Fast gradient-based adversarial example generation methods generate adversarial examples by using the gradient $\nabla_x l_f$ of the loss function l_f of the classifier f w.r.t. an input $\boldsymbol{x}$.
- Iterative Fast Gradient Sign Method (I-FGVM) changes the input $\boldsymbol{x}$ in the direction of the gradient $\nabla_x l_f$.
- Iterative Fast Gradient Sign Method (I-FGSM) uses only the sign of the gradient $\nabla_x l_f$ to change the input $\boldsymbol{x}$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Adversarial attacks

- Targeted adversarial attacks: generate adversarial examples that are misclassified as a specific label $\hat{y}$, $f(\hat{\mathbf{x}}) = \hat{y}$.
- Non-targeted adversarial attacks: generate adversarial examples that are misclassified in a label different than the ground truth label y, $f(\hat{\mathbf{x}}) \neq y$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

I-FGVM

- Gradient descent update equations of the I-FGVM.

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\hat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(clip_{[\boldsymbol{x}-\varepsilon,\, \boldsymbol{x}+\varepsilon]}(\hat{\boldsymbol{x}}_i - \alpha \cdot \nabla_x l_f(\hat{\boldsymbol{x}}_i, \hat{y})))$$

- α is the step size, $\boldsymbol{x}$ is the original image, $\hat{\boldsymbol{x}}_i$ is the adversarial image at step i, $\nabla_x l_f(\hat{\boldsymbol{x}}_i, \hat{y})$ is the first-order gradient term of the adversarial loss, $\hat{y}$ is the target class label and $clip_{[a,b]}$ is a constraint that keeps pixel values in the $[a, b]$ range.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

I-FGSM

- Gradient descent update equations of the I-FGSM.

$$\widehat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\widehat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(clip_{[\boldsymbol{x}-\varepsilon,\, \boldsymbol{x}+\varepsilon]}(\widehat{\boldsymbol{x}}_i - \alpha \cdot sign\left(\nabla_x l_f(\widehat{\boldsymbol{x}}_i, \hat{y})\right)))$$

- α is the step size, $\boldsymbol{x}$ is the original image, $\widehat{\boldsymbol{x}}_i$ is the adversarial image at step i, $\nabla_x l_f(\widehat{\boldsymbol{x}}_i, \hat{y})$ is the first-order gradient term of the adversarial loss, $\hat{y}$ is the target class label, $clip_{[a,b]}$ is a constraint that keeps pixel values in the $[a, b]$ range and $sign(\cdot)$ is the sign function.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Proposed face de-identification method

- Penalized Fast Gradient Value Method (P-FGVM).
- A novel face de-identification method based on adversarial examples.
- Inspired by the adversarial attack method I-FGVM.
- Combines an adversarial loss term and a ‘realism’ loss term in the objective function.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Proposed method – P-FGVM

- Gradient descent update equations of the P-FGVM.

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{x},$$
$$\hat{\boldsymbol{x}}_{i+1} = clip_{[0,1]}(\hat{\boldsymbol{x}}_i - \alpha \cdot \left(\nabla_x l_f(\hat{\boldsymbol{x}}_i, \hat{y}) + \lambda \cdot (\hat{\boldsymbol{x}}_i - \boldsymbol{x})\right))$$

- α is the step size, $\boldsymbol{x}$ is the original image, $\hat{\boldsymbol{x}}_i$ is the adversarial image at step i, $\nabla_x l_f(\hat{\boldsymbol{x}}_i, \hat{y})$ is the first-order gradient term of the adversarial loss, $\hat{y}$ is the target class label, λ is a weight coefficient, $clip_{[a,b]}$ is a constraint that keeps pixel values in the $[a, b]$ range and $sign(\cdot)$ is the sign function.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Proposed method – P-FGVM advantages

- De-identified images are imperceptibly different from original images.
- Can be used to attack any deep neural network classifier.
- The novel objective function leads to higher misclassification rate compared to simple adversarial attack methods (I-FGVM, I-FGSM).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Experiments

- Experimental evaluation of the proposed P-FGVM method and the baseline adversarial attack methods I-FGVM and I-FGSM.
- Two deep convolutional neural networks were used as target models.
- Both models were pre-trained on a subset of the CelebA dataset.
- The CelebA subset contains 900 random, aligned, cropped RGB facial images of 30 different persons.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Experiments - Models

- Model A has a simple architecture consisting of two convolution layers and two fully connected layers.
- Model B is the state-of-the-art VGG-Face convolutional neural network, which utilizes the VGG-16 architecture.

Model A
Conv(32, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3))
BatchNormalization+Relu
MaxPooling(PoolSize(2, 2), Strides(2, 2))
Conv(64, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3))
BatchNormalization+Relu
MaxPooling(PoolSize(2, 2), Strides(2, 2))
FC(512, L2Regularizer(1e-3))
BatchNormalization+Relu
Dropout(0.9)
FC(30)+Softmax

Model B
VGG-Face CNN descriptor (VGG-16)
FC(256, L2Regularizer(1e-3))
BatchNormalization+Relu
FC(30)+Softmax

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Evaluation metrics

- L_2-norm of the adversarial perturbation, $L2 = ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_2$.
- Mean Structural Similarity Index (MSSIM) between the original and the de-identified facial image.
- Misclassification Rate (MR) of the pre-trained models when tested with the de-identified facial images.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Results

- Comparison between the proposed P-FGVM method and the baseline I-FGVM, I-FGSM methods using the evaluation metrics L2, MSSIM, MR.

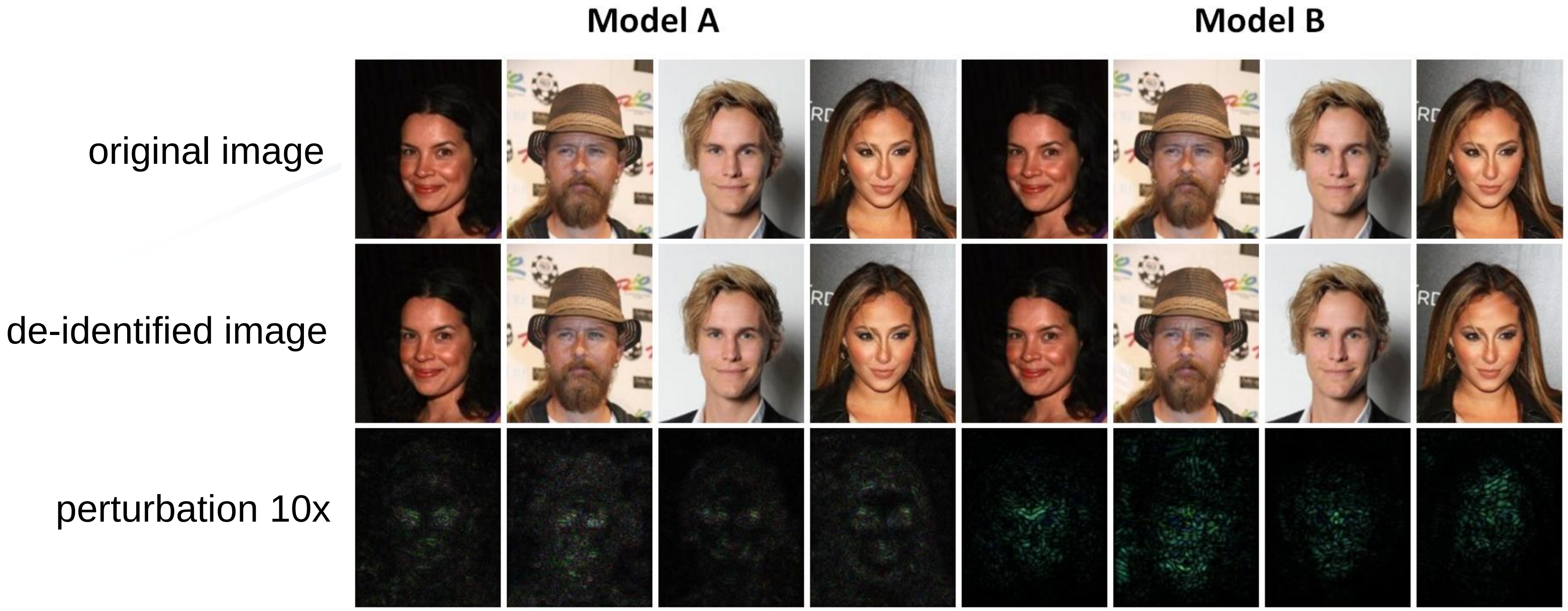
Model A			Model B		
L2	SI	MR	L2	SI	MR
Experimental Results					
P-FGVM					
3.38	0.986	99.6%	2.11	0.995	96.0%
I-FGVM					
5.31	0.963	99.4%	2.67	0.993	93.2%
I-FGSM					
5.68	0.962	98.9%	5.74	0.968	94.4%
Percentage Improvement					
I-FGVM					
36.3%	2.3%	0.2%	20.9%	0.2%	3.0%
I-FGSM					
40.4%	2.4%	0.7%	63.2%	2.7%	1.7%

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Results

- Examples of de-identified images generated using the proposed P-FGVM method and the adversarial perturbation.

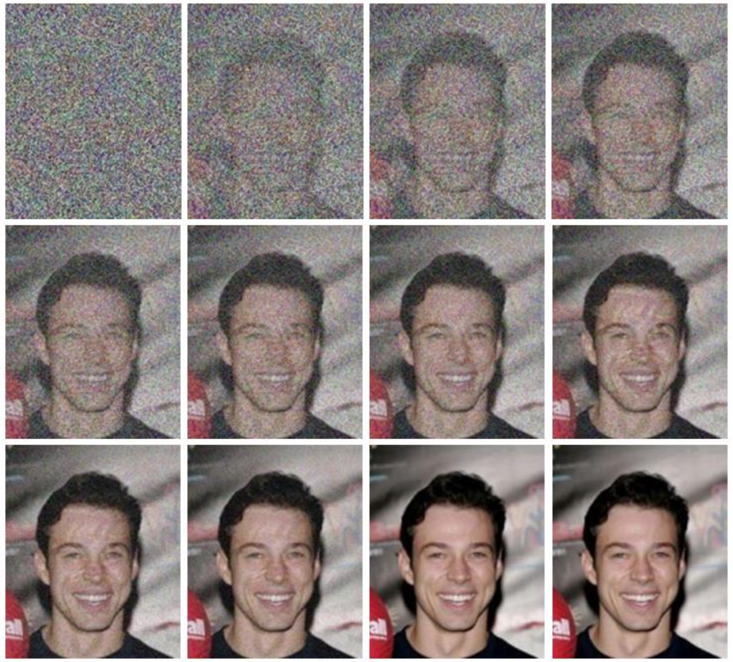


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Results

- Evolution of an example de-identified facial image generated by the proposed de-identification method P-FGVM using as input Gaussian random noise.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

Conclusions

- P-FGVM is a novel adversarial attack method for face de-identification.
- The proposed P-FGVM method generates realistic, visually imperceptible de-identified images.
- Higher misclassification rate compared to previous methods.
- Successfully fool various deep convolutional neural network face classifiers.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)

Adversarial Face De-Identification

MultiDrone

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 MULTIDRONE.

Q & A

Thank you very much for your attention!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE)