Internship Report

Abstract. In this project, we tried to develop a plugin to obfuscate images to enhance face privacy on social media. The main approach is to perform an adversarial attack so that the resulting images are wrong classified or detected by systems while the face still looks like the original one.

Keywords: Adversarial Attack · Face privacy

1 Introduction

Facial image data is incredibly valuable and sensitive because it is key to your identity. It is used to identify you in many critical applications like smile-to-pay [1] and facial recognition shoplifter detection[2]. The fear is that detailed profiles of yourself may be created if your data gets in the wrong hands. So, your face data must be changed in a way that facial recognition systems can not be trained on it. However, the image must be still represent the content it is intended to have. Therefor, we choose to perform an adversarial attack. Moreover, our attack must be able to do well in black-box setting at the end.

2 Related Work

Many attacks have been proposed on attacking facial recognition models like adversarial attacks using gradient estimation [4] Evolutionary methods [3] generative networks [7] as well as on attacking facial detection models [5].

3 Enhancing Face Privacy

In order to fool the recognition systems, there are two approaches we can take: fooling face detection and fooling face recognition. Face detection is proposing areas in the image that contains a face and face recognition is classifying the face to the right person. At first, we try to attack face detection.

3.1 Attacking Face Detection

In order to create adversarial examples we use the Projected Gradient Descent [8]. We defined a loss function and computed the gradient of loss function w.r.t the input and update the example using the gradient direction. The initial loss function was:

$$Loss = \sum_{b \in boxes} Score_b * (sign(d_b - t)) \tag{1}$$

where d is the distance from the real face position and t is a threshold. The real face was successfully undetected by the local model but we observed that fake face detection boxes had small areas which makes it unreasonable. So we changed the loss function to:

$$Loss = \sum_{b \in boxes} Score_b * Surface_b * (sign(d_b - t)) \tag{2}$$

where surface is surface of proposed area for a face. Using this loss function we were able to have fake detection with big enough boxes. This attack is targeted since not only it lowers the probability of an are containing a face, it also generates some fake proposals with reasonable sizes. Unfortunately, this attack does not seem to be transferable.

Fig. 1. Comparing unperturbed an perturbed images for Attack on Face Detection

3.2 Attacking Face Recognition

We use FaceNet[9] to embed faces in a vector, then dot product means similarity. So, we perform adversarial attack on FaceNet so that the victim will be recognized as similar to target. The percie algorithm is:

1- All faces in the image are detected and cropped and resized to 160*160 (Embedding model input size). (Using SSD Mobilenet from Face-API.js)

2- For each face distance (dot product of embeddings) to potential victims are computed and closest victim is chosen. (Embedding computed using FaceNet)

3- Projected Gradient Descent method with momentum and input diversity is applied to the objective function which is average of dot product of embeddings.

3-1 Input is resized to rnd*rnd which rnd is a random number uniformly sampled from [135,160] and then it is padded to 160*160 and rescaled image is not in center necessarily. The new image is chosen with probability of 0.9 otherwise original image is used [10]. (Input Diversity)

3-2 The image values are normalized to [-1,1] and a random uniform noise is added. Then we compute embeddings and do L2 normalization.

3-3 We compute objective function which is average of dot product of image embeddings with victim embeddings.

3-4 We compute gradient of objective function w.r.t. input image then we add the gradient to previous gradient multiplied to 0.9 [6].

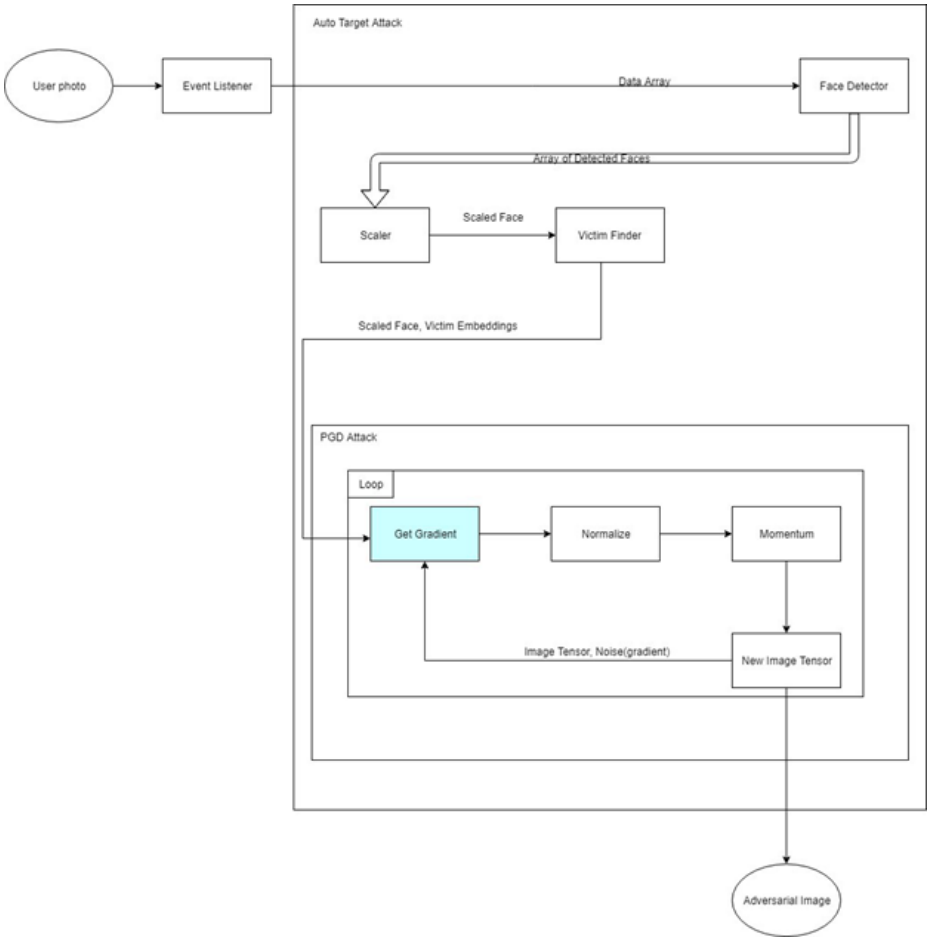
3-5 We use Gradient Sign Method to get an adversary and clip it to L-infinity bound. We repeat this process to certain number of times [8]. (Projected Gradient Descent)

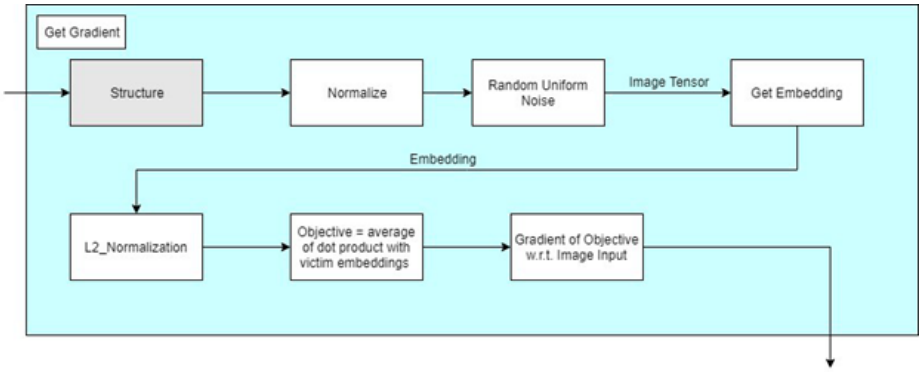
Pseudo code

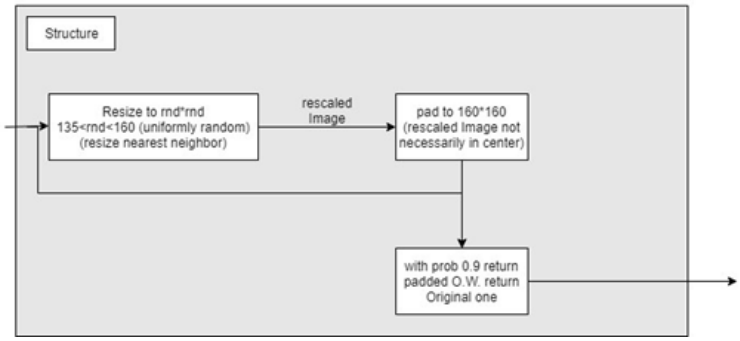
```
ObjectiveFunction(imageInput) {
   With probability 0.1 resize image to random*random
   (random uniformly from [135,160]) and
   do a random padding
   Scale imageInput to [-1.1]
   Add randomUniform noise between [ -1e-2, 1e-2] imageInput
   Get imageInput Embedding
   L2_normalize the Embedding
```

```
  Return mean of dot product of victimEmbeddings and Embeddings
  // must be maximized
}

oneStepAttack(image, grad) {
 noise = gradient of objective w.r.t. image
   // div on L2 norm
   noise = noise/ mean(|noise|)
   // momentum
   noise = grad * 0.9 + noise
   // after this, we apply sign on noise, so it will be normalized
   let adv = image + sign(noise) * 1.0
   Clip image to lower bound and upper bound
    return adv, noise
}

let input = imageBatch.toFloat();

lowerBound = image - eps , clipped to 0-255
upperBound = image + eps , clipped to 0-255

grad = zeros
for ( i = 0; i < maxIter; i++) {
           Res = oneStepAttack(input, grad);
   input = res[0];
   grad = res[1];
}
adversarial = input.toInt();
```

Diagrams


Auto Target Attack
User photo
Event Listener
Data Array
Face Detector
Array of Detected Faces
Scaler
Scaled Face
Victim Finder
Scaled Face, Victim Embeddings
PGD Attack
Loop
Get Gradient
Normalize
Momentum
Image Tensor, Noise(gradient)
New Image Tensor
Adversarial Image





References

1. https://www.theguardian.com/world/2019/sep/04/smile-to-pay-chinese-shoppers-turn-to-facial-payment-technology(CSRF)
2. https://www.cnet.com/news/with-facial-recognition-shoplifting-may-get-you-banned-in-places-youve-never-been/(CSRF)
3. Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.J., Srivastava, M.: Genattack: Practical black-box attacks with gradient-free optimization (2018)
4. Bhagoji, A.N., He, W., Li, B., Song, D.: Practical black-box attacks on deep neural networks using efficient query mechanisms. In: European Conference on Computer Vision. pp. 158–174. Springer (2018)
5. Bose, A.J., Aarabi, P.: Adversarial attacks on face detectors using neural net based constrained optimization (2018)
6. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum (2017)
7. Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition (2019)

8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
9. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298682, http://dx.doi.org/10.1109/CVPR.2015.7298682
10. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.: Improving transferability of adversarial examples with input diversity (2018)

Face Off Report

MohammadMahdi Abdollahpour, Alireza Torabian

September 2019

Notations:

H: Was able to impersonate the target person

M: Was able to change the identity to someone else

U: The original identity is not in recognizer dataset, also we couldn't impersonate the target person
. * in some cases even though we don't achieve the targets, with choosing different targets the recognizers identify different persons.

L: Couldn't fool the recognizer

1 Celebrities

1.1 Ali Daie

Original image on BetaFace: N, Clarifai: N

eps:8 | eps: 12 | eps:16

Furthest - Pete Sampras, Similarity: -0.250, BetaFace:UH(76)H(76) Clarifai:UUU

Random - Jennifer Aniston, Similarity: 0.085, BetaFace:UUU Clarifai:UUU

Random - George Robertson, Similarity: 0.083, BetaFace:UH(77)H(80) Clarifai:UUU

1.2 Chris Evans

Original image on BetaFace: Y, Clarifai: Y

eps:8 | eps: 12 | eps:16

Furthest - Mahathir Mohamad, Similarity: -0.409, BetaFace:MMM Clarifai:LMM

Random - Spencer Abraham, Similarity: -0.112, BetaFace:MMM Clarifai:LMM

Random - Michael Bloomberg, Similarity: 0.078, BetaFace:H(75)H(75)H(78) Clarifai:MMM

1.3 Cristiano Ronaldo

Original image on BetaFace: Y, Clarifai: Y

eps:8 | eps: 12 | eps:16

Furthest - Jose Maria Aznar, Similarity: -0.256, BetaFace:LMH(76) Clarifai:LLL

Random - Pervez Musharraf, Similarity: -0.242, BetaFace:MMM Clarifai:LLL

Random - John Bolton, Similarity: -0.048, BetaFace:LMM Clarifai:LMM

1.4 Jackie Chan

Original image on BetaFace: Y, Clarifai: Y

eps:8 eps: 12 eps:16

Furthest - Donald Rumsfeld, Similarity: -0.268, BetaFace:LLL Clarifai:LLL

Random - Julianne Moore, Similarity: -0.020, BetaFace:LLL Clarifai:LLL

Random - Tiger Woods, Similarity: 0.066, BetaFace:LLL Clarifai:LLL

1.5 Kim Jong un

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Trent Lott, Similarity: -0.301, BetaFace:UUU Clarifai:UUU

Random - Fidel Castro, Similarity: -0.067, BetaFace:UUU Clarifai:UUU

Random - Joschka Fischer, Similarity: 0.064, BetaFace:UUU Clarifai:UUU

1.6 Robert Downey

Original image on BetaFace: Y, Clarifai: Y

eps:8 | eps: 12 | eps:16

Furthest - Joschka Fischer, Similarity: -0.268, BetaFace:MMM Clarifai:LLM

Random - Arnold Schwarzenegger, Similarity: 0.130, BetaFace:LLH(80) Clarifai:LLH

Random - Winona Ryder, Similarity: -0.137, BetaFace:LLL Clarifai:LLL

1.7 Scarlet Johansson

Original image on BetaFace: Y, Clarifai: Y

eps:8 eps: 12 eps:16

Furthest - Jean Chretien, Similarity: -0.278, BetaFace:LMM Clarifai:MMM

Random - Trent Lott, Similarity: -0.035, BetaFace:LLM Clarifai:LLM

Random - Mohammed Al-Douri, Similarity: -0.191, BetaFace:LLL Clarifai:LMM

1.8 Taylor Swift

Original image on BetaFace: N, Clarifai: Y

eps:8 | eps: 12 | eps:16

Furthest - Jeremy Greenstock, Similarity: -0.240, BetaFace:UUU Clarifai:LLL

Random - George W Bush, Similarity: -0.128, BetaFace:UUU Clarifai:LML

Random - Paul Bremer, Similarity: -0.041, BetaFace:UUU Clarifai:LLL

1.9 Trump

Original image on BetaFace: Y, Clarifai: Y

eps:8 eps: 12 eps:16

Furthest - Hamid Karzai, Similarity: -0.316, BetaFace:LMH(78) Clarifai:L–

Random - Mahathir Mohamad, Similarity: -0.158, BetaFace:MMM Clarifai:—

Random - Hugo Chavez, Similarity: 0.014, BetaFace:LH(78)H(80) Clarifai:—

1.10 Vladimir Putin

Original image on BetaFace: Y, Clarifai: Y

eps:8 eps: 12 eps:16

Furthest - Abdullah Gul, Similarity: -0.256, BetaFace:LMM Clarifai:LLM

Random - Hugo Chavez, Similarity: 0.184, BetaFace:LLL Clarifai:LLL

Random - Hamid Karzai, Similarity: -0.114, BetaFace:LMM Clarifai:LLM

2 Famous

2.1 Ian Goodfellow

Original image on BetaFace: N, Clarifai: N

eps:8 | eps: 12 | eps:16

Furthest - Kofi Annan(BetaFace:Y), Similarity: -0.283, BetaFace:UUU Clarifai:UUU

Random - Gray Davis(BetaFace:Y), Similarity: 0.093, BetaFace:UUU Clarifai:UUU

Random - Michael Bloomberg(BetaFace:Y), Similarity: -0.095, BetaFace:UUU Clarifai:UUU

2.2 Max Amini

Original image on BetaFace: N, Clarifai: N

eps:8 | eps: 12 | eps:16

Furthest - Vicente Fox(BetaFace:Y), Similarity: -0.322, BetaFace:UUU Clarifai:UUU

Random - Jiang Zemin(BetaFace:Y), Similarity: -0.070, BetaFace:UUU Clarifai:UUU

Random - Renee Zellweger(BetaFace:Y), Similarity: -0.210, BetaFace:UUU Clarifai:UUU

2.3 Amy tan

Original image on BetaFace: Y, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Britney Spears, Similarity: -0.331, BetaFace:LLL Clarifai:UUU

Random - Jack Straw, Similarity: -0.067, BetaFace:LLL Clarifai:UUU

Random - Ricardo Lagos, Similarity: -0.026, BetaFace:LLM Clarifai:UUU

2.4 Hamad bin Khalifa Al Thani,

Original image on BetaFace: Y, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Lance Armstrong, Similarity: -0.326, BetaFace:LMM Clarifai:UUU

Random - Halle Berry, Similarity: 0.114, BetaFace:LLM Clarifai:UUU

Random - Naomi Watts, Similarity: -0.161, BetaFace:LLL Clarifai:UUU

2.5 Ko Un

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Bill Clinton(BetaFace:Y), Similarity: -0.235, BetaFace:UUU Clarifai:UUU

Random - Julianne Moore(BetaFace:Y), Similarity: 0.018, BetaFace:UUU Clarifai:UUU

Random - George W Bush(BetaFace:Y), Similarity: 0.200, BetaFace:UUU Clarifai:UUU

2.6 Philip roth

Original image on BetaFace: Y, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Norah Jones, Similarity: -0.299, BetaFace:LLL Clarifai:UUU

Random - David Beckham, Similarity: 0.046, BetaFace:LLL Clarifai:UUH

Random - Paul Bremer, Similarity: -0.028, BetaFace:LLL Clarifai:UUU

2.7 Reza Ghoochan Nejhad

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Tom Daschle(BetaFace:Y), Similarity: -0.331, BetaFace:UUU Clarifai:UUU

Random - Rudolph Giuliani(BetaFace:Y), Similarity: -0.097, BetaFace:UUU Clarifai:UUU

Random - Nestor Kirchner(BetaFace:Y), Similarity: 0.055, BetaFace:UUU Clarifai:UUU

2.8 Simin daneshvar

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16



Furthest - Winona Ryder(BetaFace:Y), Similarity: -0.188, BetaFace:UUU Clarifai:UUU



Random - Junichiro Koizumi(BetaFace:Y), Similarity: 0.105, BetaFace:UUU Clarifai:UUU



Random - Kofi Annan(BetaFace:Y), Similarity: 0.138, BetaFace:UUU Clarifai:UUU

2.9 Steve Toltz

Original image on BetaFace: N, Clarifai: N

eps:8 | eps: 12 | eps:16

Furthest - John Snow(BetaFace:N), Similarity: -0.341, BetaFace:UUU Clarifai:UUU

Random - Ariel Sharon(BetaFace:Y), Similarity: 0.113, BetaFace:UUU Clarifai:UUU

Random - Joschka Fischer(BetaFace:Y), Similarity: -0.144, BetaFace:UUU Clarifai:UUU

2.10 Yoshua Bengio

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Tiger Woods, Similarity: -0.254, BetaFace:UUU Clarifai:UUU

Random - Julie Gerberding(BetaFace:N), Similarity: -0.003, BetaFace:UUU Clarifai:UUU

Random - Roh Moo-hyun(BetaFace:Y), Similarity: -0.035, BetaFace:UUUClarifai:UUU

3 Not Famous

3.1 Joxan Jaffar

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Richard Myers(BetaFace:Y), Similarity: -0.209, BetaFace:UUU Clarifai:UUU

Random - Guillermo Coria(BetaFAce:Y), Similarity: 0.028, BetaFace:UUH(75) Clarifai:UUU

Random - Jiang Zemin(BetaFace:Y), Similarity: 0.418, BetaFace:H(88)H(87)H(88) Clarifai:UUU

3.2 Mohammad Reza Meybodi

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Carlos Moya(BetaFace:Y), Similarity: -0.237, BetaFace:UUU Clarifai:UUU

Random - Kofi Annan, Similarity: 0.266, BetaFace:UH(76)H(79) Clarifai:UUU

Random - Pervez Musharraf(BetaFace:Y), Similarity: 0.400, BetaFace:H(80)H(82)H(85) Clarifai:UUU

3.3 Ulrike Grossner

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Jack Straw(BetaFace:Y), Similarity: -0.309, BetaFace:UUU Clarifai:UUU

Random - Hans Blix(BetaFace:Y), Similarity: -0.052, BetaFace:UUU Clarifai:UUU

Random - John Snow(BetaFace:N), Similarity: -0.138, BetaFace:— Clarifai:UUU

3.4 Mehmet Fatih Yanik

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Bill Simon(BetaFace:N), Similarity: -0.288, BetaFace:— Clarifai:UUU

Random - Joschka Fischer(BetaFace:Y), Similarity: 0.107, BetaFace:H(81)H(85)H(86) Clarifai:UUU

Random - Guillermo Coria(BetaFace:Y), Similarity: -0.070, BetaFace:UH(79)H(77) Clarifai:UUU

3.5 Brian Lim

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - David Beckham, Similarity: -0.294, BetaFace:UH(75)H(78) Clarifai:UUU

Random - Jiang Zemin(BetaFace:Y), Similarity: 0.191, BetaFace:H(81)H(87)H(86) Clarifai:UUU

Random - John Snow(BetaFace:N), Similarity: -0.047, BetaFace:— Clarifai:UUU

3.6 Reza Shokri

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Gerhard Schroeder(BetaFace:Y), Similarity: -0.352, BetaFace:UUU Clarifai:UUU

Random - Paul Bremer, Similarity: -0.083, BetaFace:UH(76)H(76) Clarifai:UUU

Random - Vicente Fox(BetaFace:Y), Similarity: -0.352, BetaFace:UUU Clarifai:UUU

3.7 Tan Eng Chye

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Paul Bremer(BetaFace:Y), Similarity: -0.251, BetaFace:UUU Clarifai:UUU

Random - Tiger Woods, Similarity: 0.213, BetaFace:UUU Clarifai:UUU

Random - Jeremy Greenstock(BetaFace:N), Similarity: 0.012, BetaFace:UUU Clarifai:UUU

3.8 Vanessa Wood

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - Tom Daschle, Similarity: -0.272, BetaFace:UUU Clarifai:UUU

Random - John Negroponte(BetaFace:Y), Similarity: -0.101, BetaFace:UUU Clarifai:UUU

Random - Lance Armstrong, Similarity: 0.096, BetaFace:H(79)H(78)H(81) Clarifai:UUU

3.9 Helmut Bolcskei

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - James Blake(BetaFace:N), Similarity: -0.276, BetaFace:— Clarifai:UUU

Random - Tiger Woods, Similarity: 0.115, BetaFace:UUU Clarifai:UUU

Random - Jose Maria Aznar, Similarity: 0.242, BetaFace:H(83)H(85)H(86) Clarifai:UUU

3.10 Klaas Prussmann

Original image on BetaFace: N, Clarifai: N

eps:8 eps: 12 eps:16

Furthest - David Beckham, Similarity: -0.272, BetaFace:UUU Clarifai:UUU

Random - Bill Clinton, Similarity: -0.091, BetaFace:UUU Clarifai:UUU

Random - Joschka Fischer, Similarity: 0.047, BetaFace:H(77)H(76)H(79) Clarifai:UUU