Research Article | Vol. 28, No. 2 / 20 January 2020 / Optics Express 2499
Optics EXPRESS

Learning-based attacks for detecting the vulnerability of computer-generated hologram based optical encryption

LINA ZHOU,[1] YIN XIAO,[1] AND WEN CHEN[1,2,*]

[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[2]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China
*owen.chen@polyu.edu.hk

Abstract: Optical encryption has attracted wide attention for its remarkable characteristics. Inspired by the development of double random phase encoding, many researchers have developed a number of optical encryption systems for practical applications. It has also been found that computer-generated hologram (CGH) is highly promising for optical encryption, and the CGH-based optical encryption possesses remarkable advantages of simplicity and high feasibility for practical implementations. An input image, i.e., plaintext, can be iteratively or non-iteratively encoded into one or several phase-only masks via phase retrieval algorithms. Without security keys, it is impossible for unauthorized receivers to correctly extract the input image from ciphertext. However, cryptoanalysis of CGH-based optical encryption systems has not been effectively carried out before, and it is also concerned whether CGH-based optical encryption is sufficiently secure for practical applications. In this paper, learning-based attack is proposed to demonstrate the vulnerability of CGH-based optical security system without the direct retrieval of optical encryption keys for the first time to our knowledge. Many pairs of the extracted CGH patterns and their corresponding input images (i.e., ciphertext-plaintext pairs) are used to train a designed learning model. After training, it is straightforward to directly retrieve unknown plaintexts from the given ciphertexts (i.e., phase-only masks) by using the trained learning model without subsidiary conditions. Moreover, the proposed learning-based attacks are also feasible and effective for the cryptoanalysis of CGH-based optical security systems with multiple cascaded phase-only masks. The proposed learning-based attacking method paves the way for the cryptoanalysis of CGH-based optical encryption.

1. Introduction

With the growing demands of secured communication and storage, information security has received the increasing attention and many information security systems [1–7] have been proposed. Recently, optical encryption is demonstrated to be promising, and can open up a new direction of cryptography due to its distinguished characteristics, e.g., multiple degrees of freedom and parallel processing [8–29]. Double random phase encoding (DRPE) [6] was first proposed, in which an image (i.e., plaintext) is encoded into white stationary noise (i.e., ciphertext) by using two statistically-independent random phase-only masks. The DRPE scheme has been continuously developed in different domains, e.g., Gyrator transform, fractional Fourier transform and Fresnel transform [17–23]. In addition, optical encryption systems based on imaging mechanisms [8,14,28,29], such as ghost imaging, diffractive imaging and interferometric imaging, are illustrated to be applicable for securing information. It is also found that computer-generated hologram (CGH) [30–35] can be applied for optical encryption, and its remarkable advantages, e.g., simple implementations, have been explored for practical applications. In the CGH-based optical encryption, the input image can be iteratively or non-iteratively encoded into one or

#380004 https://doi.org/10.1364/OE.380004
Journal © 2020 Received 9 Oct 2019; revised 6 Dec 2019; accepted 7 Dec 2019; published 17 Jan 2020

Optics EXPRESS

several phase-only masks (i.e., CGH) [31–35]. The CGH-based optical encryption systems can achieve high security, and convenient optical implementations become possible in practice, e.g., by using spatial light modulators for optical decryption.

In recent years, there is more and more interest on the vulnerability analysis of optical encryption systems [36–41]. Carnicer *et al*. [36] first proposed chosen-ciphertext attack to validate that the DRPE-based infrastructure is not secure when some conditions are given. Subsequently, different vulnerability analysis methods for optical encryption systems have been proposed, e.g., chosen-plaintext attack and ciphertext-only attack [37–40]. The attacking methods have been continuously improved, and play an important role in the evolution of cryptoanalysis of various optical encryption systems. However, the estimation of optical encryption keys using elaborately-designed plaintexts-ciphertexts pairs or complex phase retrieval algorithms is usually carried out. The applicability of conventional cryptoanalysis methods can be deservedly confined in practice, and a particular phase retrieval algorithm usually needs to be designed for analyzing the security of each different optical encryption system. In addition, the vulnerability of CGH-based optical encryption systems has not been effectively studied before, and it is also concerned whether the security of CGH-based optical encryption can be sufficiently high for practical applications.

In this paper, we propose learning-based attacks on CGH-based optical encryption for the first time to our knowledge. Using many pairs of the extracted CGH patterns and their corresponding input images, a designed learning model [41–44] is trained to emulate the inner representations (e.g., optical setup parameters and phase-only masks) of the training data. Then, the trained learning model is applied to recover unknown plaintexts from the given CGH patterns without the usage of optical encryption keys. The proposed learning-based attacks are demonstrated to be feasible and effective to fully analyze the vulnerability of CGH-based optical cryptosystems.

2. Principles

2.1. CGH-based optical encryption

Figure 1 shows a schematic setup for a typical optical security system based on CGH with two cascaded phase-only masks (i.e., one phase-only mask M1 to be extracted as ciphertext and one fixed phase-only mask M2 as security key). The encoding process of CGH-based optical cryptosystem is conducted by using an iterative phase retrieval algorithm to extract the phase-only mask M1 under the constraint of original input image (i.e., plaintext). In the initial iteration, phase-only mask M1 is initialized randomly in a range of $[0, 2\pi]$, and random phase-only mask M2 is pre-defined and fixed. For the encoding, the phase-only masks M1 and M2 are respectively denoted as $M_1^n(\mu, \upsilon)$ and $M_2(\xi, \eta)$, where n (i.e., integers 1, 2, 3,. . . . .) denotes the nth iteration. The wavefront $f^n(x, y)$ in the input image plane can be described by

$$f^n(x, y) = \mathrm{FrT}_{d_2,\lambda}(\{\mathrm{FrT}_{d_1,\lambda}[M_1^n(\mu, \upsilon)]\}M_2(\xi, \eta)), \tag{1}$$

where FrT denotes free-space wave propagation in the Fresnel domain [45], λ denotes wavelength of the incident wave, d_1 denotes the axial distance between phase-only mask M1 and phase-only mask M2, and d_2 denotes the axial distance between phase-only mask M2 and the input image plane. Spectrum method is adopted to describe the free-space wave propagation. To update the wavefront $f^n(x, y)$ in the input image plane, original input image $I(x, y)$ is applied as a constraint.

$$f_{update}^n(x, y) = \sqrt{I(x, y)}\frac{f^n(x, y)}{|f^n(x, y)|}, \tag{2}$$

Optics EXPRESS

where | | denotes modulus operation, and $f^n_{update}(x, y)$ denotes the updated wavefront in the input image plane. Subsequently, phase-only mask M1 is further updated by

$$M_1^{n+1}(\mu, \upsilon) = \frac{\mathrm{FrT}_{-d_1,\lambda}(\{\mathrm{FrT}_{-d_2,\lambda}[f^n_{update}(x, y)]\}[M_2(\xi, \eta)]^*)}{|\mathrm{FrT}_{-d_1,\lambda}(\{\mathrm{FrT}_{-d_2,\lambda}[f^n_{update}(x, y)]\}[M_2(\xi, \eta)]^*)|}, \tag{3}$$

where asterisk denotes complex conjugate. Then, the wavefront in the input image plane can be further updated by

$$f^{n+1}(x, y) = \mathrm{FrT}_{d_2,\lambda}(\{\mathrm{FrT}_{d_1,\lambda}[M_1^{n+1}(\mu, \upsilon)]\}M_2(\xi, \eta)). \tag{4}$$

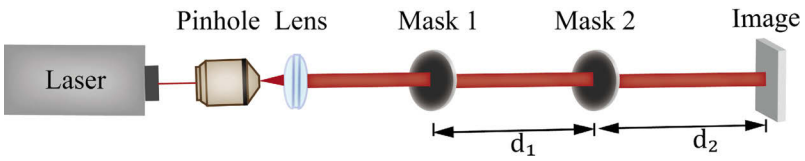

Fig. 1. A schematic for a typical CGH-based optical encryption. Mask 1 (M1) and Mask 2 (M2): phase-only masks. For optical decryption, the input image plane can be replaced by using a CCD camera.

To evaluate the difference between the estimated input image $|f^n(x, y)|$ and original input image $I(x, y)$, correlation coefficient (CC) is implemented by using a ready-made Matlab function 'corr2'. The higher value of CC means the higher similarity between the estimated input image and original input image. When a preset CC value (i.e., threshold) is achieved, the iterative process can be stopped and the updated phase-only pattern $M_1^n(\mu, \upsilon)$ is correspondingly used as the ultimate estimation of phase-only mask M1, i.e., as $M_1(\mu, \upsilon)$.It has been widely demonstrated that without optical encryption keys (e.g., wavelength, axial distances and the fixed phase-only mask M2), it is impossible to extract original input image (i.e., plaintext) from a given phase-only mask M1 [31–35] during optical decryption.

2.2. Learning-based attacks

Learning-based attacks are proposed and designed here to analyze the security of CGH-based optical encryption, which can extract senior representations from the given data [42–44,46–49]. One of the most popular learning architectures for the imaging problems is convolutional neural network (CNN), which is widely used for object classification [46] and object reconstruction [42–44]. The CNN architecture belongs to supervised learning algorithms, which are trained by using input-output pairs $D = \{\mathbf{x}, \mathbf{y}\}_{n=1}^N$, where $\mathbf{x}$ denotes an input vector and $\mathbf{y}$ denotes an output vector. Based on a loss function $L(\mathbf{y}, \hat{\mathbf{y}})$, the CNN model devotes to finding the optimal model parameters Θ with the given data pairs. After training, the trained model can be denoted as $f(\hat{\mathbf{x}}; \Theta)$, where $\hat{\mathbf{y}}$ denotes the prediction of arbitrary $\hat{\mathbf{x}}$ obtained from the function $f(\hat{\mathbf{x}}; \Theta)$. Through the end-to-end learning, the designed learning models are trained to learn the inner mapping relationships of the input data and output data without any subsidiary conditions (e.g., parameters of the environments for data acquisition). This breathtaking scheme [42–44] has been thoroughly applied in various areas, which can also make significant progress in the development of cryptoanalysis of CGH-based optical encryption systems in this study. With sufficient pairs of ciphertexts and the corresponding plaintexts fed to a designed learning model, the trained machine learning model can be used to retrieve the unknown plaintext from a given ciphertext without the usage of optical encryption keys existing in the typical CGH-based optical encryption setup shown in Fig. 1, e.g., optical setup parameters and phase-only mask M2.

Optics EXPRESS

A framework of learning-based attacks using CNN model for analyzing the vulnerability of CGH-based optical encryption is shown in Fig. 2. Many pairs of ciphertexts and the corresponding plaintexts are sent to the designed learning model. Each ciphertext is processed by n groups of cascaded convolution and pooling layers. The convolution layer is labeled as C_1, C_2, ..., C_n, and the pooling layer is identified as P_1, P_2, ..., P_n. It is worth noting that the sequence of convolution layers and pooling layers is not confined to the arrangement of a convolution layer followed by a pooling layer. The arrangement of convolution layers and pooling layer shown in Fig. 2 is an exemplification of the proposed CNN-based machine learning attacks. Assume that the ciphertext (denoted as $\mathbf{x}$) is of size $m \times m$, and then the input ciphertext is convolved with certain kernels forming the first convolution layer C_1. Size of the kernels for the first convolution layer is denoted as $L_1 \times L_1$ with the number of the kernels as p, and the kernels are initialized to be $\Theta_1 = \{\mathbf{w}^1, \mathbf{b}^1\}$, where $\mathbf{w}^1$ denotes a set of weights and $\mathbf{b}^1$ denotes a set of biases for the first convolution layer C_1. Hence, the feature map for the first convolution layer $\mathbf{x}_1$ can be described by

$$\mathbf{x}_1 = \sigma[(\mathbf{w}^1 * \mathbf{x}) + \mathbf{b}^1], \tag{5}$$

where $*$ denotes the processing of convolution, and activation functions for neural networks are denoted as σ. Size of the first convolution layer is $(m\text{-}L_1+1)\times(m\text{-}L_1+1)\times p$. Followed by down-sampling, the convolution layer $C_1(\mathbf{x}_1)$ is transformed to the first pooling layer $P_1(\mathbf{x}_{1,p})$ with size of $[(m - L_1 + 1)/2] \times [(m - L_1 + 1)/2] \times p$. Similarly, the feature map for the nth convolution layer $\mathbf{x}_n$ is given by

$$\mathbf{x}_n = \sigma[(\mathbf{w}^n * \mathbf{x}_{n\text{-}1,\,p}) + \mathbf{b}^n], \tag{6}$$

where $\mathbf{x}_{n-1,p}$ denotes feature map of the $(n-1)$th pooling layer P_{n-1}, and $\mathbf{w}^n$ and $\mathbf{b}^n$ denote parameters for the nth convolution layer C_n as $\Theta_n = \{\mathbf{w}^n, \mathbf{b}^n\}$. Size of the nth convolution layer C_n is $[(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^{n-1}] \times [(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^{n-1}] \times p$, and size of the nth pooling layer P_n is $[(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^n] \times [(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^n] \times p$. Followed by a reshaping layer, the nth pooling layer P_n is converted to a one-dimensional vector $\mathbf{A}$. Then, the reshaped layer is fully connected to a one-dimensional layer of the desired plaintext by $\Theta_{FC} = \{\mathbf{w}^{FC}, \mathbf{b}^{FC}\}$, where $\mathbf{w}^{FC}$ denotes a set of weights and $\mathbf{b}^{FC}$ denotes a set of biases for the fully connected layer FC. Finally, the fully connected layer is reshaped to the size of original plaintext, and a prediction $\hat{\mathbf{y}}$ of the plaintext is generated. To evaluate the difference between the predicted plaintext $\hat{\mathbf{y}}$ and original plaintext $\mathbf{y}$, mean squared error (MSE) is used as the loss function described by

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2, \tag{7}$$

where Y_i and $\hat{Y}_i$ respectively denote the ith pixel value of original plaintext $\mathbf{y}$ and the predicted plaintext $\hat{\mathbf{y}}$ (i.e., i ranges from 1 to N), and N denotes the total pixel number. The parameters of weights and biases should be updated until the preset MSE is reached. When the calculated MSE value is larger than the preset threshold, backpropagation is implemented to further update the weights and biases. The error $\Delta\mathbf{d}$ between the prediction and original plaintext is denoted as

$$\Delta\mathbf{d} = \hat{\mathbf{y}} - \mathbf{y}. \tag{8}$$

The error $\Delta\mathbf{F}$ at the fully connected layer is given by

$$\Delta\mathbf{F} = [\mathbf{w}^{FC}]^{\mathrm{T}} * \Delta\mathbf{d}, \tag{9}$$

where $[\mathbf{w}^{FC}]^{\mathrm{T}}$ denotes transpose of $\mathbf{w}^{FC}$. Then, $\Delta\mathbf{F}$ is reshaped to the size of the nth pooling layer P_n with size of $[(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^n] \times [(m - L_1 - 2L_2 - 2^{n-1}L_n + 2^n - 1)/2^n] \times$

Optics EXPRESS

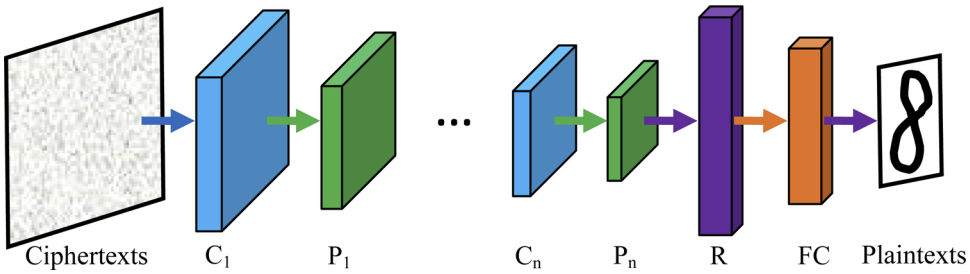

Fig. 2. A framework for the proposed learning-based attacks using a designed CNN model for analyzing the security of CGH-based optical encryption. C_1: the first convolution layer; C_n: the nth convolution layer; P_1: the first pooling layer; P_n: the nth pooling layer; R: the shaping layer; FC: fully connected layer. Many pairs of ciphertexts and plaintexts are fed to the designed CNN model. Each ciphertext is processed by n groups of convolution and pooling layers, followed by a reshaping layer to convert the three-dimensional nth pooling layer into a one-dimensional vector. Then, the reshaped vector is fully connected to a one-dimensional vector. It is worth noting that size of the one-dimensional vector coincides with the size of the transformed one-dimensional plaintext. Through the processing of reshaping, size of the one-dimensional vector is converted to that of original plaintext. After multiple layers of processing, the input ciphertext is decomposed and fully connected to original plaintext. Feeding a number of ciphertexts-plaintexts pairs to the designed learning model, the CNN model can be adequately trained to be ready for making a real-time prediction of unknown plaintexts from the given ciphertexts.

p.The reshaped error $\Delta\mathbf{P}_n$ at the nth pooling layer P_n is backpropagated to the nth convolution layer C_n by upsampling, and the error at the nth convolution layer is denoted as $\Delta\mathbf{C}_n$.Through the convolution, the error at the $(n-1)$th pooling layer P_{n-1} is denoted as $\Delta\mathbf{P}_{n-1}$. Following the aforementioned error propagation rules, the error at the first convolution layer is $\Delta\mathbf{C}_1$. Then, the gradient of $\mathbf{w}^{FC}$ can be given by

$$\mathbf{w}_g^{FC} = \Delta\mathbf{d} * \mathbf{A}, \tag{10}$$

where $\mathbf{w}_g^{FC}$ denotes the gradient of $\mathbf{w}^{FC}$. $\mathbf{b}_g^{FC}$ is given by

$$\mathbf{b}_g^{FC} = \sum \Delta\mathbf{d}. \tag{11}$$

The gradient of $\mathbf{w}^n$ at the nth convolution layer is the convolution of $\mathbf{x}_{n-1,p}$ and $\Delta\mathbf{C}_n$. Similarly, $\mathbf{b}_g^n$ is given by

$$\mathbf{b}_g^n = \sum \Delta\mathbf{C}_n. \tag{12}$$

Following the aforementioned method for calculating the gradient of weights and biases, the gradient of $\mathbf{w}^1$ at the first convolution layer is the convolution of $\mathbf{x}$ and $\Delta\mathbf{C}_1$. Similarly, $\mathbf{b}_g^1$ is given by

$$\mathbf{b}_g^1 = \sum \Delta\mathbf{C}_1. \tag{13}$$

Here, stochastic gradient descent is applied to update the weights and biases [49], which is described by

$$\mathbf{w}^n = \mathbf{w}^{n-1} - \mathbf{v}_w^n, \tag{14}$$

$$\mathbf{v}_w^n = m\mathbf{v}_w^{n-1} + \alpha\mathbf{w}_g^n, \tag{15}$$

$$\mathbf{b}^n = \mathbf{b}^{n-1} - \mathbf{v}_b^n, \tag{16}$$

$$\mathbf{v}_b^n = m\mathbf{v}_b^{n-1} + \alpha\mathbf{b}_g^n, \tag{17}$$

where $\mathbf{v}_w^n$ and $\mathbf{v}_b^n$ respectively denote the velocity of weights and biases, m denotes the momentum, α denotes the learning rate, $\mathbf{w}_g^n$ and $\mathbf{b}_g^n$ respectively denote the gradients of the weight and bias at

Optics EXPRESS

the nth convolution layer, and $\mathbf{w}_g^{n-1}$ and $\mathbf{b}_g^{n-1}$ respectively denote the gradients of weight and bias at the $(n-1)$th convolution layer. Through the updating rule aforementioned, the weights and biases can be continuously updated until the MSE value approaches the preset value. When the number of ciphertext-plaintext pairs used for training is insufficient, the existing ciphertext-plaintext pairs can be reused for several epochs of iterations. Finally, the designed learning model is adequately trained to be applied for retrieving unknown plaintext from the given ciphertext without the usage of optical encryption keys existing in the CGH-based optical encryption setup shown in Fig. 1.

3. Results and discussion

A typical CGH-based encryption setup shown in Fig. 1 is applied to illustrate feasibility and effectiveness of the proposed method. In practice, laser source with wavelength of 632.8 nm can be expanded by a pinhole, and then the expanded light can be collimated by a collimating lens. Subsequently, the expanded and collimated laser beam illuminates the extracted phase-only mask M1 and further modulated by the fixed random phase-only mask M2. Axial distances d_1 and d_2 are 17.0 cm and 12.0 cm, respectively. The random phase-only masks can be embedded into spatial light modulators with a dimension of 512×512 pixels in practice. In Fig. 1, the input image plane can be replaced by using a CCD camera for optical decryption. It is worth noting that in CGH-based optical encryption, a digital approach usually needs to be used for the encoding, and either a digital or optical approach can be flexibly applied for the decoding.

The input images (i.e., plaintexts) are handwritten-digit patterns from the MNIST database [47], and also the patterns with fashion products from the fashion MNIST database [48]. All input images randomly selected from the handwritten-digit patterns and fashion images are of 8-bit and 512×512 pixels which are digitally resized from pixel size of 28×28. From each database, 5000 images are randomly selected to verify validity of the proposed method. Therefore, 10000 CGH patterns, i.e., phase-only mask M1, can be correspondingly extracted by using the iterative phase retrieval algorithm. The maximum number of iterations is set as 1000 in the iterative phase retrieval algorithm, and the CC threshold is set as 0.98. In CGH-based optical encryption, it is feasible to recover the plaintext from the ciphertext during the decoding when all optical setup parameters and phase-only mask M2, i.e., security keys, are given and correctly applied. However, until now there is no systematic study about the security of CGH-based encryption systems, and nowadays it is still concerned whether CGH-based optical encryption is sufficiently secure for practical applications. In this paper, using the proposed learning-based attacks, we demonstrate for the first time to our knowledge that the CGH-based optical encryption is vulnerable, and optical encryption keys, e.g., setup parameters and the fixed phase-only mask M2, are not requested to be used for retrieving unknown plaintexts from the given ciphertexts in the proposed learning-based attacking method.

The structure of the designed learning-based attacks for the cryptoanalysis of CGH-based optical encryption is shown in Fig. 3. The designed learning model has two convolution layers, two pooling layers, one reshaping layer and one fully connected layer. To lower the computational cost, the extracted phase-only mask M1 (i.e., ciphertext) with a dimension of 512×512 pixels is resized to 200×200 pixels. The resized ciphertext convolves with 30 kernels of size 1×1, and then activated by the sigmoid function to generate the first convolution layer of size 200×200×30. The first convolution layer is downsized to the first pooling layer with size of 100×100×30. Then, the down-sampled feature maps further convolve with 30 kernels (size of 1×1) forming the second convolution layer of size 100×100×30. Followed by the second action of down-sampling, the second convolution layer is converted to the second pooling layer with size of 50×50×30. Subsequently, the feature maps obtained after two rounds of convolution and down-sampling processing are reshaped to a column vector with size of 1×75000. The reshaped vector is processed by the fully connected layer with size of 1×784 followed by an action of reshaping, and then the output layer is a vector with size of 28×28. 4800 pairs of the extracted CGH patterns and

Optics EXPRESS

their corresponding plaintexts from each database are fed to the designed learning model in the training phase. The learning rate is set as 10^{-6}, and momentum is set as −0.00095. The training epoch is set as 5. The weights are initially set as random values between 0 and 1, and the biases are initially set as 0. It takes about 8.0 h to train a CNN model for each database. To implement the proposed method, Matlab platform is used with Nvidia Geforce GTX1080Ti GPU and RAM of 64GB. After training, the trained CNN model for each database can predict unknown plaintexts from the given ciphertexts in real time without the usage of various optical encryption keys.

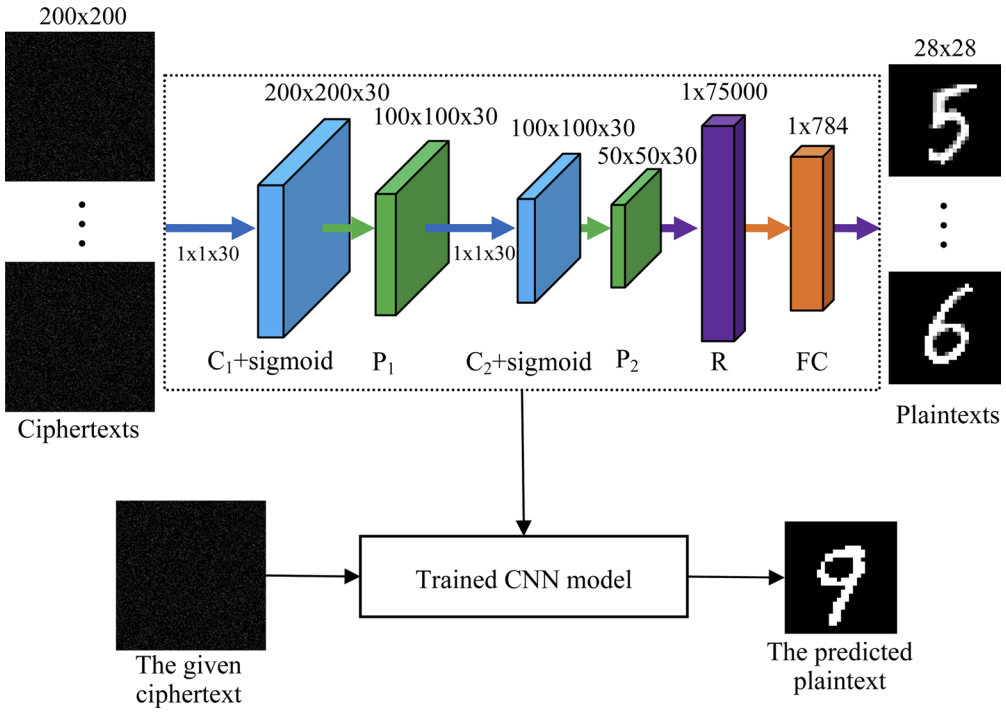

Fig. 3. A designed CNN architecture for attacking the CGH-based optical encryption shown in Fig. 1. Two convolution layers and two pooling layers are used in this study. The convolution layers are activated by the sigmoid function. The ciphertexts are resized from 512×512 pixels to 200×200 pixels to lower the computational load. With 4800 ciphertext-plaintext pairs sent to the designed learning model, the designed CNN model is well trained. Then, 200 ciphertexts without *prior* knowledge about their plaintexts are tested, and the trained CNN model is able to predict unknown plaintexts from the given ciphertexts.

Two databases are used to illustrate performance of the proposed attacks on CGH-based cryptosystem. Figure 4 shows the recovered plaintexts from the given CGH patterns by usage of the trained models. Figures 4(a), 4(g) and 4(m) in the first column, Figs. 4(c), 4(i) and 4(o) in the fourth column, and Figs. 4(e), 4(k) and 4(q) in the seventh column show the encrypted CGH patterns (i.e., ciphertexts). It can be seen from the ciphertexts in Fig. 4 that original input images are completely encrypted. Figures 4(b), 4(d) and 4(f) in the first row show the unknown plaintexts retrieved by utilizing the fashion MNIST database trained model. Figures 4(h), 4(j) and 4(l) in the second row show the unknown plaintexts retrieved by using the handwritten-digit MNIST database trained model. Figures 4(bb), 4(dd), 4(ff), 4(hh), 4(jj) and 4(ll) show the original plaintexts. To evaluate quality of the predicted plaintexts, peak signal-to-noise ratio (PSNR) and CC values are adopted here. PSNRs of the extracted plaintexts in the first and second rows are 34.96 dB, 28.34 dB, 25.63 dB, 30.02 dB, 28.90 dB and 29.81 dB, respectively. CCs of

Optics EXPRESS

the extracted plaintexts in the first and second rows are 0.96, 0.92, 0.89, 0.87, 0.83 and 0.86, respectively. In view of the PSNR and CC values, the unknown plaintexts are fully extracted from ciphertexts. It is demonstrated that the designed attacks can fully retrieve the unknown plaintexts without any requirement of setup keys. Therefore, the CGH-based encryption systems are not secure enough for practical applications under the designed attacks.

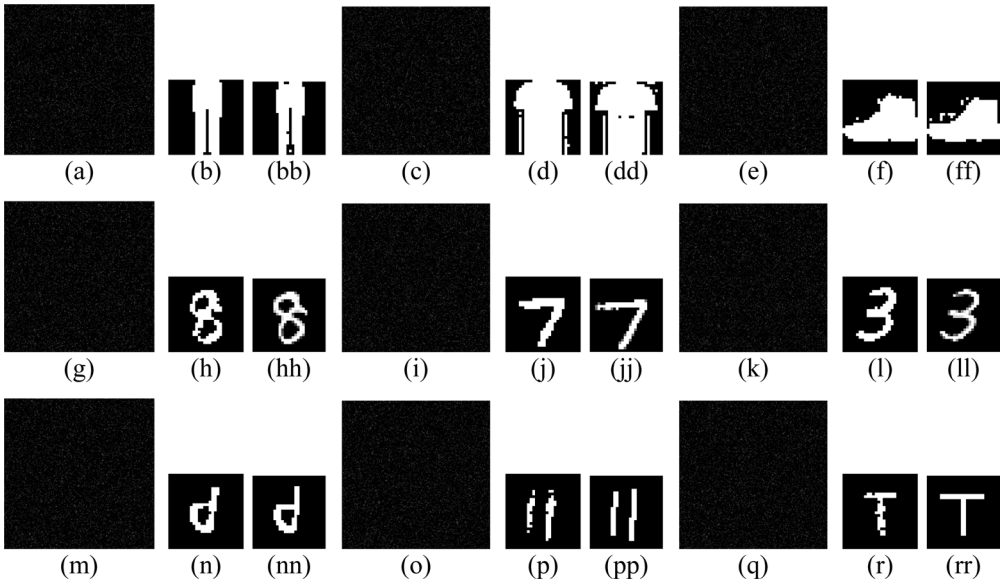

Fig. 4. The attacks on the encryption based on CGH. The first column ((a), (g), (m)), the fourth column ((c), (i), (o)) and the seventh column ((e), (k), (q)) the ciphertexts by the optical cryptosystem based on CGH. The first row (b), (d) and (f) the retrieved plaintexts by the fashion MNIST database trained model corresponding to (a), (c) and (e), respectively. The second row (h), (j) and (l) the retrieved plaintexts by the MNIST database trained model corresponding to (g), (i) and (k), respectively. The third row (n), (p) and (r) the plaintexts (from different databases) retrieved by the MNIST database trained model corresponding to (m), (o) and (q), respectively. The third column ((bb), (hh), (nn)), the sixth column ((dd), (jj), (pp)) and the ninth column ((ff), (ll), (rr)) original plaintexts respectively corresponding to the ciphertexts in the first column ((a), (g), (m)), the fourth column ((c), (i), (o)) and the seventh column ((e), (k), (q)).

To illustrate the robustness and universality of the attacks on CGH-based encoding, the trained model is further applied to extract the plaintexts from different databases, which are comprised of distinct patterns (e.g., lowercase, double digits and uppercase letters). It is worth mentioning that these distinct patterns have never been used in the training stage. Figures 4(m), 4(o) and 4(q) in the third row present the ciphertexts of these input images encrypted by the CGH-based encryption setup in Fig. 1. Figures 4(n), 4(p) and 4(r) show the unknown plaintexts retrieved by utilizing the CNN model which is trained by making use of the handwritten-digit MNIST database. The original plaintexts corresponding to Figs. 4(m), 4(o) and 4(q) are shown in Figs. 4(nn), 4(pp) and 4(rr), respectively. PSNRs of the recovered plaintexts are 33.28 dB, 29.93 dB, 26.47 dB, respectively. CCs of the extracted plaintexts are 0.88, 0.81 and 0.77, respectively. It is demonstrated that the trained CNN model is applicable to correctly predict or extract the data which is from different databases. According to learning concept, the nonlinear mapping relationship between the input and output data can be learned. Hence, the trained learning model for CGH-based encryption is also available for the data which is not used in the training phase. Although the plaintexts are from different databases, the ciphertexts obtained by the CGH-based

Optics EXPRESS

cryptosystem have the similarity in some aspects. Hence, the trained model can be applied to predict the data from different databases, and the vulnerability of CGH-based encryption is effectively detected.

It has been illustrated in the literature [31–35] that when more random phase-only masks are used in CGH-based encryption setup, the higher security for CGH-based encryption system can be achieved. In this study, the proposed attacks on CGH-based encryption with multiple random phase-only masks are also investigated. The CGH-based setup with multiple random phase-only masks (i.e., phase-only mask M1 to be extracted and fixed phase-only masks M2 and M3 as principal security keys) is schematically shown in Fig. 5. The axial distances d_1, d_2 and d_3 are 17.0 cm, 12.0 cm and 15.0 cm, respectively. 5000 images are randomly selected from each database as the plaintexts, and the CGH patterns, i.e., M1, are sequentially retrieved as ciphertexts. 4800 ciphertext-plaintext pairs generated by using each database are used to train the model. Another 200 extracted CGH patterns are used for the testing. The time taken to train each database is about 8.0 h. Figures 6(a), 6(g) and 6(m) in the first column, Figs. 6(c), 6(i) and 6(o) in the fourth column, and Figs. 6(e), 6(k) and 6(q) in the seventh column show the ciphertexts by usage of the CGH-based encoding setup shown in Fig. 5. The ciphertexts shown in the first row are further processed by usage of the fashion MNIST database trained model. The ciphertexts shown in the second row are further processed by usage of the handwritten-digit MNIST database trained model. The predicted plaintexts are also shown in Fig. 6 just following the corresponding ciphertexts. The ciphertexts given in the third row are obtained by using the input images selected from different databases. It is also worth mentioning that these images have never been used to train the model in the training phase. By using the model trained by using the handwritten-digit MNIST database, the ciphertetxs in the third row can be successfully processed, and their correspondingly predicted plaintexts are shown in Figs. 6(n), 6(p) and 6(r). The original plaintexts are shown in Figs. 6(bb), 6(dd), 6(ff), 6(hh), 6(jj), 6(ll), 6(nn), 6(pp) and 6(rr), respectively. PSNRs of the retrieved plaintexts in Fig. 6 are 18.71 dB, 15.09 dB, 32.31 dB, 29.15 dB, 27.46 dB, 28.70 dB, 28.94 dB, 28.34 dB, 29.26 dB, respectively. CCs for the retrieved plaintexts in Fig. 6 are 0.74, 0.62, 0.95, 0.83, 0.82, 0.84, 0.82, 0.82, 0.76, respectively. It is demonstrated that the method presented here is feasible and effective for detecting the vulnerability of CGH-based cryptosystem with multiple cascaded random phase-only masks.

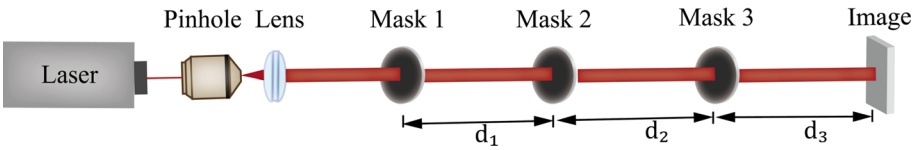

Fig. 5. A schematic for CGH-based encryption with multiple cascaded random phase-only masks. Mask 1 (M1), Mask 2 (M2) and Mask 3 (M3): phase-only masks. For optical decoding, the input image plane can be replaced by using a CCD.

The results and analyses aforementioned have systematically demonstrated that the designed attacks are feasible and effective for vetting the security of CGH-based cryptosystems. The trained CNN model can recover original images from the given CGH patterns in real time. Moreover, robustness of the proposed attacks is also illustrated that the trained CNN model is applicable to attack different databases. Although multiple random phase-only masks can be used in practice, the trained model still performs well to retrieve original images. The method presented here provides a powerful tool for analyzing the vulnerability of CGH-based cryptosystems, which has never been studied before. It is believed that the method presented here could push the further developments of securer CGH-based encryption systems.

Optics EXPRESS

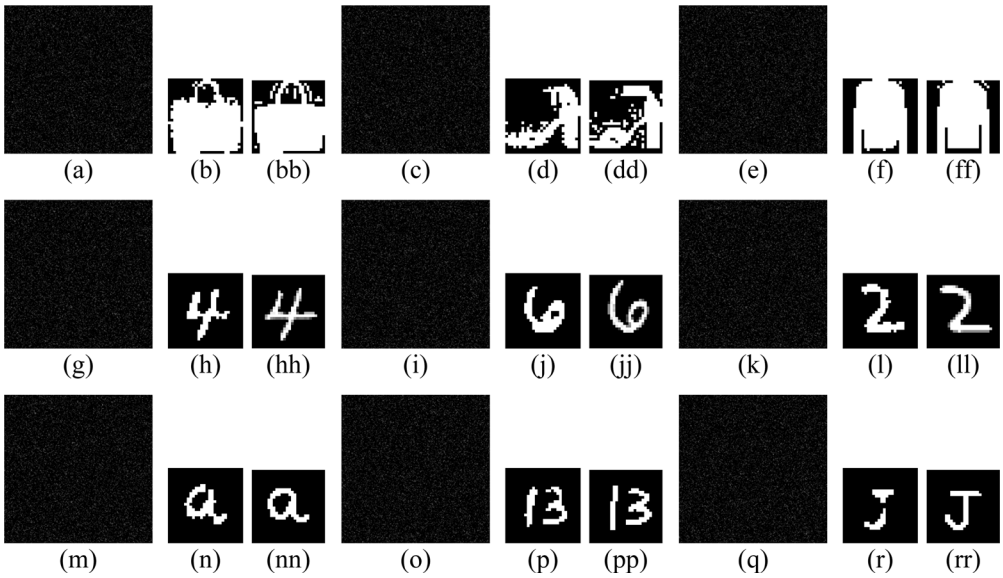

Fig. 6. The attacks on the encryption based on CGH. The first column ((a), (g), (m)), the fourth column ((c), (i), (o)) and the seventh column ((e), (k), (q)) the ciphertexts by the optical cryptosystem based on CGH. The first row (b), (d) and (f) the retrieved plaintexts by the fashion MNIST database trained model corresponding to (a), (c) and (e), respectively. The second row (h), (j) and (l) the retrieved plaintexts by the MNIST database trained model corresponding to (g), (i) and (k), respectively. The third row (n), (p) and (r) the plaintexts (from different databases) retrieved by the MNIST database trained learning model corresponding to (m), (o) and (q), respectively. The third column ((bb), (hh), (nn)), the sixth column ((dd), (jj), (pp)) and the ninth column ((ff), (ll), (rr)) original plaintexts corresponding to the ciphertexts in the first column ((a), (g), (m)), the fourth column ((c), (i), (o)) and the seventh column ((e), (k), (q)), respectively.

4. Conclusions

We have presented the learning-based attacks on CGH-based encryption, and vulnerability of CGH-based encryption has been detected. The feasibility and effectiveness of the designed model are fully verified, and input images from a different database are also tested and successfully predicted. Furthermore, the method presented here can also effectively analyze the CGH-based encryption with multiple cascaded random phase-only masks. The method presented here paves the way for the development of cryptoanalysis of CGH-based encryption, and can eventually promote the development of CGH-based encryption.

Funding

National Natural Science Foundation of China (61605165); Hong Kong Research Grants Council, University Grants Committee (25201416); Shenzhen Science and Technology Innovation Commission (JCYJ20160531184426473); Hong Kong Polytechnic University (4-BCDY, G-YBVU).

Disclosures

The authors declare that there are no conflicts of interest.

Optics EXPRESS

References

1. B. Javidi, "Securing information with optical technologies," Phys. Today **50**(3), 27–32 (1997).
2. F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding-a survey," Proc. IEEE **87**(7), 1062–1078 (1999).
3. R. C. Merkle, "Secure communications over insecure channels," Commun. ACM **21**(4), 294–299 (1978).
4. K. Morita, H. Yoshimura, M. Nishiyama, and Y. Iwai, "Protecting personal information using homomorphic encryption for person re-identification," In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)* 166–167 (2018).
5. A. Surekha, P. R. Anand, and I. Indu, "E-payment transactions using encrypted QR codes," Int. J. Appl. Eng. Res. **10**(77), 461 (2015).
6. P. Refregier and B. Javidi, "Optical image encryption based on input plane and Fourier plane random encoding," Opt. Lett. **20**(7), 767–769 (1995).
7. O. Matoba, T. Nomura, E. Perez-Cabre, M. S. Millan, and B. Javidi, "Optical techniques for information security," Proc. IEEE **97**(6), 1128–1148 (2009).
8. W. Chen, B. Javidi, and X. Chen, "Advances in optical security systems," Adv. Opt. Photonics **6**(2), 120–155 (2014).
9. O. Matoba and B. Javidi, "Encrypted optical memory system using three-dimensional keys in the Fresnel domain," Opt. Lett. **24**(11), 762–764 (1999).
10. W. Chen and X. Chen, "Space-based optical image encryption," Opt. Express **18**(26), 27095–27104 (2010).
11. L. Sui, Y. Cheng, Z. Wang, A. Tian, and A. K. Asundi, "Single-pixel correlated imaging with high-quality reconstruction suing iterative phase retrieval algorithm," Opt. Lasers Eng. **111**, 108–113 (2018).
12. A. Alfalou and C. Brosseau, "Optical image compression and encryption methods," Adv. Opt. Photonics **1**(3), 589–636 (2009).
13. W. Chen, X. Chen, and C. J. R. Sheppard, "Optical image encryption based on diffractive imaging," Opt. Lett. **35**(22), 3817–3819 (2010).
14. Y. Shi, T. Li, Y. Wang, Q. Gao, S. Zhang, and H. Li, "Optical image encryption via ptychography," Opt. Lett. **38**(9), 1425–1427 (2013).
15. L. Sui, C. Du, X. Zhang, A. Tian, and A. Asundi, "Double-image encryption based on interference and logistic map under the framework of double random phase encoding," Opt. Lasers Eng. **122**, 113–122 (2019).
16. J. F. Barrera, A. Mira, and R. Torroba, "Optical encryption and QR codes: secure and noise-free information retrieval," Opt. Express **21**(5), 5373–5378 (2013).
17. N. Singh and A. Sinha, "Gyrator transform-based optical image encryption using chaos," Opt. Lasers Eng. **47**(5), 539–546 (2009).
18. O. Matoba and B. Javidi, "Encrypted optical storage with wavelength-key and random phase codes," Appl. Opt. **38**(32), 6785–6790 (1999).
19. G. Unnikrishnan, J. Joseph, and K. Singh, "Optical encryption by double-random phase encoding in the fractional Fourier domain," Opt. Lett. **25**(12), 887–889 (2000).
20. G. Situ and J. Zhang, "Double random-phase encoding in the Fresnel domain," Opt. Lett. **29**(14), 1584–1586 (2004).
21. G. Unnikrishnan and K. Singh, "Double random fractional Fourier domain encoding for optical security," Opt. Eng. **39**(11), 2853–2859 (2000).
22. R. Tao, Y. Xin, and Y. Wang, "Double image encryption based on random phase encoding in the fractional Fourier domain," Opt. Express **15**(24), 16067–16079 (2007).
23. Z. Liu and S. Liu, "Double image encryption based on iterative fractional Fourier transform," Opt. Commun. **275**(2), 324–329 (2007).
24. L. Chen and D. Zhao, "Optical image encryption with Hartley transforms," Opt. Lett. **31**(23), 3438–3440 (2006).
25. Z. Liu, Q. Li, J. Dai, X. Sun, S. Liu, and M. A. Ahmad, "A new kind of double image encryption by using a cutting spectrum in the 1-D fractional Fourier transform domains," Opt. Commun. **282**(8), 1536–1540 (2009).
26. M. R. Abuturab, "Color image security system based on discrete Hartley transform in gyrator transform domain," Opt. Lasers Eng. **51**(3), 317–324 (2013).
27. N. Singh and A. Sinha, "Chaos based multiple image encryption using multiple canonical transforms," Opt. Laser Technol. **42**(5), 724–731 (2010).
28. P. Clemente, V. Durán, V. Torres-Company, E. Tajahuerce, and J. Lancis, "Optical encryption based on computational ghost imaging," Opt. Lett. **35**(14), 2391–2393 (2010).
29. E. Tajahuerce, O. Matoba, S. C. Verrall, and B. Javidi, "Optoelectronic information encryption with phase-shifting interferometry," Appl. Opt. **39**(14), 2313–2320 (2000).
30. A. W. Lohmann and D. P. Paris, "Binary Fraunhofer holograms, generated by computer," Appl. Opt. **6**(10), 1739–1748 (1967).
31. R. K. Wang, I. A. Watson, and C. Chatwin, "Random phase encoding for optical security," Opt. Eng. **35**(9), 2464–2469 (1996).
32. Y. Zhang and B. Wang, "Optical image encryption based on interference," Opt. Lett. **33**(21), 2443–2445 (2008).
33. S. Xi, X. Wang, L. Song, Z. Zhu, B. Zhu, S. Huang, and H. Wang, "Experimental study on optical image encryption with asymmetric double random phase and computer-generated hologram," Opt. Express **25**(7), 8212–8222 (2017).
34. E. G. Johnson and J. D. Brasher, "Phase encryption of biometrics in diffractive optical elements," Opt. Lett. **21**(16), 1271–1273 (1996).

Optics EXPRESS

35. H. E. Hwang, H. T. Chang, and W. N. Lie, "Multiple-image encryption and multiplexing using a modified Gerchberg-Saxton algorithm and phase modulation in Fresnel-transform domain," Opt. Lett. **34**(24), 3917–3919 (2009).
36. A. Carnicer, M. Montes-Usategui, S. Arcos, and I. Juvells, "Vulnerability to chosen-cyphertext attacks of optical encryption schemes based on double random phase keys," Opt. Lett. **30**(13), 1644–1646 (2005).
37. X. Peng, H. Wei, and P. Zhang, "Chosen-plaintext attack on lensless double-random phase encoding in the Fresnel domain," Opt. Lett. **31**(22), 3261–3263 (2006).
38. X. Peng, P. Zhang, H. Wei, and B. Yu, "Known-plaintext attack on optical encryption based on double random phase keys," Opt. Lett. **31**(8), 1044–1046 (2006).
39. C. Guo, S. Liu, and J. T. Sheridan, "Iterative phase retrieval algorithms. Part II: Attacking optical encryption systems," Appl. Opt. **54**(15), 4709–4719 (2015).
40. X. Liu, J. Wu, W. He, M. Liao, C. Zhang, and X. Peng, "Vulnerability to ciphertext-only attack of optical encryption scheme based on double random phase encoding," Opt. Express **23**(15), 18955–18968 (2015).
41. L. Zhou, Y. Xiao, and W. Chen, "Machine-learning attacks on interference-based optical encryption: experimental demonstration," Opt. Express **27**(18), 26143–26154 (2019).
42. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
43. K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3929–3938 (2017).
44. H. Hai, S. Pan, M. Liao, D. Lu, W. He, and X. Peng, "Cryptanalysis of random-phase-encoding-based optical cryptosystem via deep learning," Opt. Express **27**(15), 21204–21213 (2019).
45. J. W. Goodman, *Introduction to Fourier optics*, 2nd ed. (McGraw-Hill, 1996).
46. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* 1, 1097–1105 (2012).
47. L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," IEEE Signal Process. Mag. **29**(6), 141–142 (2012).
48. H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708. 07747 (2017).
49. I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning," Proceedings of the 30th International Conference on Machine Learning, PMLR **28**(3), 1139–1147 (2013).