ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096627

# PRIVACY PRESERVING FACE RECOGNITION WITH LENSLESS CAMERA

*Chris Henry, *M. Salman Asif, and *Zhu Li

*University of Missouri-Kansas City, USA
*University of California, Riverside, USA

## ABSTRACT

The widespread adoption of facial recognition technology is a global phenomenon. Facial recognition systems leverage upon image data containing faces. This poses serious threats to user privacy as the data is exposed to potential data breaches. In this paper, we propose a face recognition system that works without compromising user privacy. It utilizes data captured by FlatCam - a lensless camera. FlatCam captures the scene as a sensor measurement that is visually unintelligible. The proposed system preserves user privacy since it works directly on FlatCam's sensor measurements without the need of FlatCam camera parameters which are required for pixel reconstruction. We propose a frequency domain deep learning solution that computes the DCT of the sensor field at multiple resolutions and organizes it into subbands before training a classification network with attention. The multi-resolution DCT subband representation leads to huge performance gains when compared to using the sensor measurement directly for training. Our proposed system was trained and tested on a real lensless camera dataset - the FlatCam Face dataset. Privacy of user is preserved during both training and testing. Experimental results demonstrate the effectiveness of our method.

***Index Terms***— Lensless camera, FlatCam, face recognition, visual privacy, DCT

## 1. INTRODUCTION

Face recognition is a well-researched topic in the computer vision community. It deals with the identification of faces in images or videos and has attracted significant attention [1, 2, 3] due to its immense practical applicability in areas like biometrics, surveillance, etc. Facial recognition systems leverage upon huge amounts of image data containing faces. This face data is vulnerable to digital attacks that poses serious threats to user privacy.

Recently, lensless cameras have gained much attention owing to its thin form-factor, lightweight, and inexpensiveness when compared to lens-based cameras. A recent example of lensless camera is FlatCam [4]. It replaces lenses with computations and captures a scene as sensor measurements.

This work is partially supported by NSF grants 1747751 and 2148382.

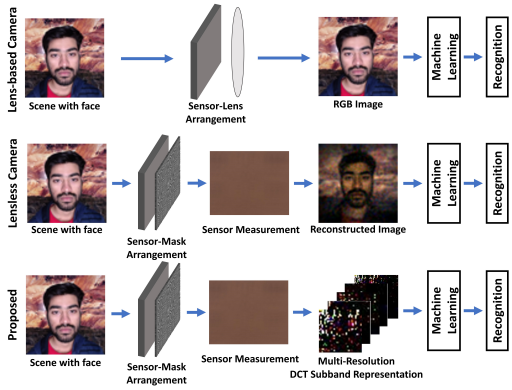


**Fig. 1**: Proposed privacy preserving face recognition system (bottom) vs. lensless camera reconstruction-based face recognition system (middle) vs. traditional lens-based face recognition system (top). Face image taken from FlatCam Face dataset [5].

Computational algorithms are used to reconstruct the scene via these sensor measurements. It is worth noting that the recorded sensor measurements are incomprehensible to humans. A sensor measurement for a face will therefore maintain the user privacy since the measurements lack spatial correlation.

In the past, a few research works [6, 7, 8] explored the use of sensor measurements from lensless cameras for performing computer vision tasks like image classification and action recognition. These approaches are reconstruction-free since they do not require reconstructing image from the sensor measurement. The study in [6] generated a local binary pattern map for optically encoded pattern and used it for image classification. The encoded pattern was obtained from a lensless camera. A transformer-based architecture was trained using optically encoded pattern from a mask-based lensless camera for image classification [7]. The work in [8] proposed a privacy preserving action recognition system that used coded aperture videos. Recently, lensless camera based privacy preserving face recognition systems [9, 10] have also been proposed. The study in [9] achieves privacy protection by initially training a network using unblurred images followed by fine-tuning it with blurred images obtained via lensless multi-

pinhole camera. On the other hand, the study in [10] reconstructs face image from FlatCam measurement, predicts the sensitive face region via facial segmentation, and separates them from the captured measurements.

Our work is most similar to [6, 7], since image reconstruction is not needed, however, these works were tested on dummy datasets like MNIST, Fashion MNIST, and/or cat-vs-dogs dataset. On the contrary, our work deals with face recognition which has real-life applicability. In case of [9], it still requires pre-training with unblurred images while [10] requires image reconstruction and facial segmentation which adds to the computational resources and time required. A pictorial comparison for the proposed approach and traditional approaches is shown in Figure 1.

In this paper, we propose a facial recognition system that maintains user privacy while being able to perform face recognition. The proposed system uses sensor measurements recorded by FlatCam for face recognition. These sensor measurements are incomprehensible to humans and cannot be used to recover the face images without the knowledge of camera parameters. Sensor measurements are transformed into frequency domain via Discrete Cosine Transform (DCT). The DCT is computed at multiple resolutions and organized into subbands to form a multi-resolution DCT subband representation before inputting it to a CNN. Particularly, we use VGG network with attention [11] (referred to as VGG-ATT in the following text) as the network for training and inference. The proposed system is camera parameter blind and hence, preserves privacy during both training and inference. Our contributions are mentioned below:

1. We propose a privacy preserving face recognition system that only requires sensor measurements from lensless camera during both training and inference time. These sensor measurements are unintelligible to humans which enables the protection of user's privacy.

2. Lensless image sensor field is heavily blurred with a point spread function (PSF) that will need extremely large receptive field to untangle the convolution. To remedy this, we developed a frequency domain learning solution that converts the sensor measurements from lensless camera into the frequency domain via computing DCT at multiple resolutions forming a multi-resolution DCT subband representation. This representation is used to train VGG with attention which results in huge accuracy gains when compared to using sensor measurements directly for training.

3. Experiments are conducted on real lensless camera dataset. The data contains sensor measurements captured by FlatCam. No simulated data was used during training or testing the face recognition system which validates the effectiveness of our proposed approach under real-world scenarios.

The remainder of the paper is organized as follows. Section 2 describes the working of our proposed method while the experimental setup and results are presented in Section 3. Conclusion and future work are covered in the last section.

## 2. PROPOSED METHOD

This section describes the proposed approach for privacy preserving face recognition. The overall approach is illustrated in Figure 2.

### 2.1. FlatCam Imaging

This subsection provides the background knowledge for understanding the working of FlatCam [4]. It mainly consists of a large bare-sensor along with a coded binary mask. The entire camera system is devoid of optical lenses which makes it a lightweight and thin device.

Imagine a light point source enters the image sensor through the aperture in the mask; the point spread function of the system would be a shadow of the mask. For a complex scene, each sensor pixel represents multiplexed light from multiple scene elements. For a separable mask, the sensor measurement $Y$ can be written as:

$$Y = \Phi_L X \Phi_R^T + E \tag{1}$$

where $X$ is the scene radiance, $\Phi_L$ and $\Phi_R$ are the system matrices computed once by calibrating the camera model, and $E$ is the sensor noise. Given $Y$, the original scene can be reconstructed by solving a $\ell_2$ regularized least-squares problem that can be represented as:

$$\hat{X} = \arg\min_X \left\| \Phi_L X \Phi_R^T - Y \right\|_2^2 + \tau \|X\|_2^2 \tag{2}$$

where $\tau > 0$ is a regularization parameter.

### 2.2. Multi-Resolution DCT Subband Representation

The first step towards building our privacy preserving face recognition system is using the sensor measurements from FlatCam for training the network. However, directly training the sensor measurements leads to mediocre performance as evident in Section 3.3. To tackle this issue, we propose to convert the sensor measurements into a multi-resolution DCT subband representation before using it for training. This step is an extension of the frequency domain learning via DCT subbands [12] proposed for aerial image classification via lensless camera.

Each raw sensor measurement $Y_{raw}$ from the FlatCam Face dataset [5] has a size of $1280 \times 1024$ pixels where each $2 \times 2$ window represents the Bayer pattern. The raw sensor measurement is split into a red channel, a blue channel, and 2 green channels, each of size $620 \times 500$. The 2 green channels are averaged to generate a sensor measurement $Y$ of size

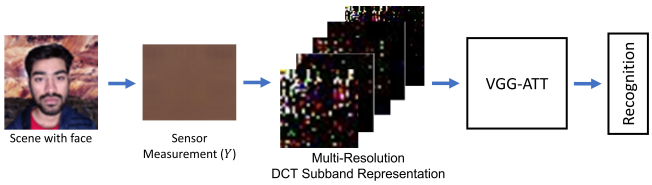


**Fig. 2**: Overall workflow of the proposed method. Face image taken from FlatCam Face dataset [5].

$620 \times 500 \times 3$. $Y$ is resized to $64 \times 64 \times 3$ and $32 \times 32 \times 3$ to generate $Y_{64}$ and $Y_{32}$, respectively.

$Y_{64}$ and $Y_{32}$ are transformed into frequency domain via computing DCT. Let $Y_{DCT} = DCT(Y)$ where $Y_{DCT}$ represent the result obtained after transforming the sensor measurement $Y$ into frequency domain using DCT. $Y_{DCT64}$ and $Y_{DCT32}$ refer to the DCT of $Y_{64}$ and $Y_{32}$, respectively. $Y_{DCT64}$ is further decomposed into $X_0$, $X_1$, $X_2$, and $X_3$ subbands, each of size $(h/2, w/2, 3)$ where $h$ and $w$ refer to the height and width of $Y_{DCT64}$. To generate the final multi-resolution DCT subband representation, we concatenate $X_0$, $X_1$, $X_2$, $X_3$, and $Y_{DCT32}$ into a multi-resolution $32 \times 32 \times 15$ input $Y_{mDCT}$. A pictorial explanation of this procedure can be visualized in Figure 3.

This multi-resolution DCT subband representation of the sensor measurement leads to significant gains in the overall accuracy of the face recognition system. This is evident from the results in Table 1. The subband organization prevents low frequency DC and AC components from dominating the filter weights and enables a large receptive field for the network.

It would be possible to recover the sensor measurement from the DCT representation. However, recovering the pixel reconstruction of the face from the sensor measurement would require the knowledge of camera parameters. It would be almost impossible to recover the face image without the knowledge of the camera parameters. This makes our system highly secure in terms of privacy during both training and inference time.

### 2.3. Frequency Domain Learning with Attention

The success of attention mechanism in language models led to its adoption in vision models [11]. Vision models with attention mechanism have shown superior performance in deep learning tasks. The attention mechanism mimics the human visual system which tends to focus on an area of interest rather than whole visual space.

Inspired by the success of attention mechanism, we use the architecture from [11] for our face recognition system. The architecture in [11] is a modified version of VGG [13] containing 15 convolutional and 2 fully connected (fc) layers. Particularly, [11] inserts attention estimators after the 7th, 10th, 13th layers and replaced the last fc layer with a new fc layer that takes input from these attention estimators. We use the network referred to as 'VGG-att3-concat-pc' in the original paper [11]. '-att3' means that the last three levels contain the attention layers, 'concat' means that the attention outputs are concatenated before inputting to the fc layer, and 'pc' means that parametrised compatibility was used to calculate the compatibility scores. For simplicity, 'VGG-att3-concat-pc' is referred as 'VGG-ATT' in this paper. More details about the architecture can be obtained from [11]. We also trained on VGG16 network [13].

Contrary to the original work [11], we train the network in the frequency domain with the multi-resolution DCT subband representation $Y_{mDCT}$ obtained from the procedure in Section 2.2. The benefit of training in the frequency domain with $Y_{mDCT}$ is evident from Section 3.3.

## 3. EXPERIMENTAL SETUP AND RESULTS

### 3.1. Dataset Details

The proposed method was trained and tested on FlatCam Face data [5] which was collected by Rice University. The dataset contains $23,838$ samples for $87$ subjects with $274$ samples per subject under different operating conditions. It is a real lensless camera dataset that was captured using FlatCam under different lighting conditions with expression and angle variations. In addition to the sensor measurements, the dataset also provides reconstructed and webcam-captured images. In our experiments, we take every 10th (1st, 11th, 21th, and so on) sample from each subject for testing and all the remaining samples for training. The test samples include $28$ different variations. No simulated data was used during our experiments to ensure real-life applicability of our system.

### 3.2. Experimental Setup

The system was implemented on a desktop computer with Intel Core i5-8400 CPU and 40 gigabytes of RAM. A single NVIDIA GTX 1080Ti GPU was used for training and testing the network. PyTorch framework was used for implementing the networks used in this paper. The implementation for VGG-ATT was taken from [14]. VGG16 was trained for $200$ epochs while VGG-ATT was trained for $100$ epochs. Both networks were trained using using stochastic gradient descent with a batch size of $256$, momentum of $0.9$, and weight decay of $0.0005$. The learning rate for both the networks was initially set to $0.001$. The learning rate was divided by $5$ every

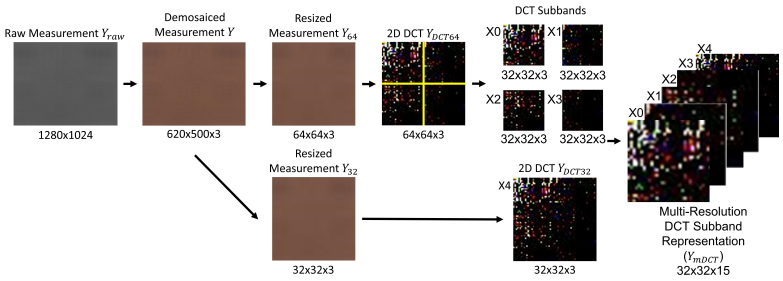


**Fig. 3**: Overall workflow for obtaining multi-resolution DCT subband representation.

60th epoch for VGG16. In case of VGG-ATT, the learning rate was decreased by a multiplicative factor of 0.1 for every 50th epoch.

### 3.3. Experimental Results

In this subsection, we discuss the experimental results for our proposed privacy preserving face recognition system on the FlatCam Face dataset [5]. Table 1 presents results on two different types of inputs on VGG16 and VGG-ATT networks. The networks were trained on sensor measurement $Y_{64}$ $(64 \times 64 \times 3)$ and the proposed multi-resolution DCT subband representation $Y_{mDCT}$ $(32 \times 32 \times 15)$ (obtained with procedure mentioned in Section 2.2)

Learning directly using $Y_{64}$ results in a mediocre accuracy of $46.59\%$ and $42.81\%$ for VGG and VGG-ATT, respectively. When learned with the proposed multi-resolution DCT subband representation $Y_{mDCT}$, the performance boosts by about $27\%$ from $46.59\%$ to $73.60\%$ for VGG. This performance gain is even higher (about $45\%$) in case of VGG-ATT with a jump from $42.81\%$ to $87.71\%$. This validates the effectiveness of our multi-resolution DCT subband representation. It is worth mentioning that no synthetic data was used during training and testing of our system. In addition, in all our experiments, we train the networks from scratch and did not use any pre-trained weights. The performance might be further increased by using a pre-trained network.

We also conducted experiments with reconstructed image from sensor measurement $Y$ $(512 \times 620 \times 3)$, and images captured by standard webcam. Both, reconstructed and webcam-captured images were resized to $64 \times 64 \times 3$ before training/testing. VGG-ATT performed well with an accuracy of $98.07\%$ and $99.71\%$ for reconstructed images and webcam-captured images, respectively. On the other hand, VGG16 achieved an accuracy of $93.27\%$ for reconstructed images and $98.52\%$ for webcam-captured images. This high accuracy, at the loss of user privacy, is not surprising given that the camera parameters were known to obtain a good reconstruction.

A comparison with the original FlatCam Face dataset paper [5] would be unfair since [5] used a much larger dataset (VGG Face dataset [15]) to train the network and tested it on FlatCam Face dataset. [5] prepared a display-captured lensless version of the VGG Face dataset [15] which, to the best of our knowledge, is unavailable publicly. Hence, we were unable to train our network with this larger display-captured VGG Face dataset. Although, our results with $Y_{mDCT}$ do not match the performance of reconstructed or webcam-captured images, however, it must be noted that our approach maintains user privacy with a reasonable amount of accuracy. We strongly believe that our research is a step forward towards inference with lensless camera's sensor measurements while maintaining user privacy.

**Table 1**: Face recognition results on FlatCam Face dataset [5] for VGG16 [13] and VGG-ATT [11] with sensor measurement $Y_{64}$ and proposed multi-resolution DCT subband representation $Y_{mDCT}$.

| Model | Input Data | Accuracy |
|---|---|---|
| VGG16 | Sensor Measurement ($Y_{64}$) | 46.59 |
| | DCT Representation ($Y_{mDCT}$) | **73.60** |
| VGG-ATT | Sensor Measurement ($Y_{64}$) | 42.81 |
| | DCT Representation ($Y_{mDCT}$) | **87.71** |

## 4. CONCLUSION

In this work, we proposed a strong privacy preserving face recognition system that uses sensor measurements from FlatCam without the need to reconstruct face image in the pixel domain. We proposed to convert the sensor measurement into a multi-resolution DCT subband representation before using it for training VGG16 or VGG-ATT (VGG with attention). Training with this DCT representation boosts the performance (when compared to using sensor measurement directly for training) while preserving user privacy. The system is highly secure during both training and testing time since reconstructing the face image from the sensor measurement would require the knowledge of camera parameters. The face recognition system was trained and tested on a real lensless dataset - the FlatCam Face dataset. Experimental results show that our privacy preserving system performs almost similar to the traditional face recognition systems.

## 5. REFERENCES


[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823. 1

[2] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015. 1

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708. 1

[4] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk, "Flatcam: Thin, lensless cameras using coded aperture and computation," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 384–397, 2016. 1, 2

[5] Jasper Tan, Li Niu, Jesse K Adams, Vivek Boominathan, Jacob T Robinson, Richard G Baraniuk, and Ashok Veeraraghavan, "Face detection and verification using lensless cameras," *IEEE Transactions on Computational Imaging*, vol. 5, no. 2, pp. 180–194, 2018. 1, 2, 3, 4

[6] Xiuxi Pan, Tomoya Nakamura, Xiao Chen, and Masahiro Yamaguchi, "Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation," *Optics Express*, vol. 29, no. 7, pp. 9758–9771, 2021. 1, 2

[7] Xiuxi Pan, Xiao Chen, Tomoya Nakamura, and Masahiro Yamaguchi, "Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer," *Optics Express*, vol. 29, no. 23, pp. 37962–37978, 2021. 1, 2

[8] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang, "Privacy-preserving action recognition using coded aperture videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 1

[9] Yasunori Ishii, Satoshi Sato, and Takayoshi Yamashita, "Privacy-aware face recognition with lensless multi-pinhole camera," in *European Conference on Computer Vision*. Springer, 2020, pp. 476–493. 1, 2

[10] Thuong Nguyen Canh and Hajime Nagahara, "Deep compressive sensing for visual privacy protection in flatcam imaging," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3978–3986. 1, 2

[11] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018. 2, 3, 4

[12] Chris Henry, Birendra Kathariya, M. Salman Asif, Zhu Li, and George York, "Aerial image classification through thin lensless camera," in *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2022, pp. 27–30. 2

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3, 4

[14] Huang Lianghua, "Pytorch implementation of the iclr 2018 paper learning to pay attention," https://github.com/huanglianghua/pay-attention-pytorch, Accessed: 2022-Feb-12. 3

[15] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," 2015. 4