SCIENCE CHINA
Information Sciences

• LETTER •

January 2020, Vol. 63 119103:1–119103:3
https://doi.org/10.1007/s11432-018-9862-9

The FrFT convolutional face: toward robust face recognition using the fractional Fourier transform and convolutional neural networks

Xin WU[1,2], Ran TAO[1,2*], Danfeng HONG[3,4] & Yue WANG[1,2]

[1]*The School of Information and Electronics, Beijing Institute of Technology, Beijing* 100081*, China;*
[2]*Beijing Key Laboratory of Fractional Signals and Systems, Beijing Institute of Technology, Beijing* 100081*, China;*
[3]*Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling* 82234*, Germany;*
[4]*Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich* 82234*, Germany*

Received 26 August 2018/Revised 14 January 2019/Accepted 15 March 2019/Published online 16 October 2019

Citation Wu X, Tao R, Hong D F, et al. The FrFT convolutional face: toward robust face recognition using the fractional Fourier transform and convolutional neural networks. Sci China Inf Sci, 2020, 63(1): 119103, https://doi.org/10.1007/s11432-018-9862-9

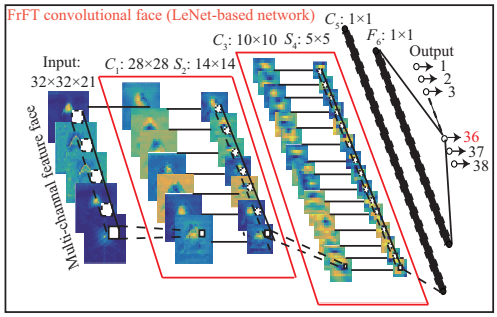
Dear editor,

Recently, advances in techniques, such as deep learning, have attracted increasing interest in the field of biometrics [1–3]. In facial recognition systems, facial images are observed to inevitably suffer from various effects, such as occlusions (facemask, sunglasses, and expressions), illumination, and pose variants, during image collection, resulting in performance degradation. Further, machine learning approaches, such as manifold regularization, have been recently proposed in [4–6] to learn the application of discriminative low-dimensional representation. Unfortunately, these approaches fail to extract the semantically meaningful information owing to the lack of adequate modeling of the spatial information.

Convolutional neural networks (CNNs) have been proven to be effective for extracting semantically meaningful information. However, the extraction of features from only the spatial domain does not improve the recognition performance. First, facial images comprise various complex components, including texture, structure, and minutiae. However, the CNN representation of multiple components is limited because of the usage of a small-scale training dataset. Second, even though CNN can easily represent various components, it is still considered to be difficult to model some features in the spatial domain because of the noise distribution, illumination condition, and posture and expression variations owing to the fact that the location information of the pixels remains unchanged while conducting spatial image transformation such as convolution or filtering. Unlike spatial transformation, frequency-based transformation, such as the two-dimensional (2D) fractional Fourier transform (FrFT) [7], not only decomposes the image into multiple components having different attributes but also adequately models the facial variations by capturing the changing location information in the frequency domain. Therefore, we obtain a discriminative and robust facial representation that is not affected by conditions such as the illumination, posture, and occlusion.

In this study, we propose the application of FrFT and CNN in a cascaded fashion, which can be referred to as the FrFT convolutional face, capable of detailed extraction of the rich facial features from both the spatial domain and the fractional Fourier domain. Using various combinations of different orders in FrFT and different components, such as the amplitude and phase information, the proposed FrFT convolutional face effec-

* Corresponding author (email: rantao@bit.edu.cn)



(a)

$\mathcal{ABIP}_p$+LeNet (%)			Network ablation study (%)		
p	Ex-YaleB	AR	Method	Ex-YaleB	AR
0	86.58	85.36	FFT	79.18	80.12
0+1.0	85.12	83.08	FrFT	83.12	82.18
0+0.1	89.87	87.12	LeNet	86.58	85.36
0:0.1:0.2	91.33	88.53	$\mathcal{AB}_p$+LeNet	87.96	90.64
0:0.1:0.4	92.25	90.65	$\mathcal{IP}_p$+LeNet	87.22	85.26
0:0.1:0.6	93.16	91.32	$\mathcal{IAB}_p$+LeNet	91.38	84.05
0:0.1:0.8	**94.06**	**92.80**	$\mathcal{I}(\mathcal{PAB})_p$+LeNet	92.42	91.68
0:0.1:1.0	93.39	91.12	$\mathcal{ABIP}_p$+LeNet	**94.06**	**92.80**

(b)

Figure 1 (Color online) (a) The proposed network architecture and (b) quantitative comparisons using different p values and network ablation study on two face datasets.

tively learns the facial variations. Figure 1(a) denotes the architectural overview of the proposed framework.

The proposed method. In [7], FrFT was observed to easily handle the issue of low-quality facial images in facial recognition. Further, the selection of the order $p \in (0, 1)$ in FrFT plays a vital role while designing the facial recognition systems. Theoretically, small p values result in a considerably structured data representation; conversely, the features, such as texture, can be adequately extracted using large p values. The continuous FrFT of an image can be given as

$$X_p(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(s,t)K_p(s,t,u,v)\mathrm{d}s\mathrm{d}t, \tag{1}$$

where $x(s,t)$ denotes the input image; $K_p(s,t,u,v)$ represents the kernel function. Further, we obtain

$$\begin{aligned} K_p(s,t,u,v) = {} & \frac{1}{2\pi}\sqrt{1-\mathrm{j}\cot\alpha}\sqrt{1-\mathrm{j}\cot\beta} \\ & \cdot \exp(\mathrm{j}(s^2+u^2)/2\cot\alpha - \mathrm{j}su\csc\alpha) \\ & \cdot \exp(\mathrm{j}(t^2+v^2)/2\cot\beta - \mathrm{j}tv\csc\beta), \end{aligned} \tag{2}$$

where α and β denote the rotation angles in FrFT. Based on the kernel separability, the 2D FrFT can be defined as

$$X_{\alpha,\beta}(u,v) = \mathrm{FrFT}_{\beta}^{t\to v}(\mathrm{FrFT}_{\alpha}^{s\to u}(x(s,t))). \tag{3}$$

Because the 2D FrFT is linearly reversible, the corresponding inverse FrFT (IFrFT) can be written as

$$x(s,t) = \mathrm{FrFT}_{-\beta}^{v\to t}(\mathrm{FrFT}_{-\alpha}^{u\to s}(X_{\alpha,\beta}(u,v))). \tag{4}$$

Consequently, given an input image $\mathbf{I}$, the FrFT-related features contain two aspects. One of these aspects is generated by the forward transform $\mathcal{T} = \{\mathcal{AB}_p, \mathcal{P}_p\}$, whereas the other is generated by the inverse transform $\mathcal{IT} = \{\mathcal{IAB}_p, \mathcal{IP}_p\}$. Thus, $\mathcal{AB}_p$ and $\mathcal{P}_p$ denote the amplitude and phase of p-th order FrFT, respectively, whereas $\mathcal{IAB}_p$ and $\mathcal{IP}_p$ denote the amplitude and phase, in case of IFrFT, respectively. When combined with the original RGB channels, the spatial-frequency feature set $\mathcal{C}$ can be defined as

$$\mathcal{C} := \{\underbrace{\mathbf{I}}_{\mathcal{GRAY}}, \underbrace{\Omega_p(\mathbf{I})}_{\mathcal{T}}, \underbrace{\Omega_{-p}(\mathcal{T})}_{\mathcal{IT}}\}, \tag{5}$$

where Ω denotes the FrFT operator.

Subsequently, we investigated the effects of two different spatial-frequency feature combinations, namely $\mathcal{ABIP}_p$ and $\mathcal{I}(\mathcal{PAB})_p$. Among them, $\mathcal{ABIP}_p$ is Gray(1) + $\mathcal{AB}_p(10)$ + $\mathcal{IP}_p(10)$ for a total of 21 channels, and $\mathcal{I}(\mathcal{PAB})_p$ is Gray(1) + $\mathcal{IAB}_p(10)$ + $\mathcal{IP}_p(10)$ for a total of 21 channels. Further, we identified these features as the new input that is to be fed into the network. Our method is a two-step learning framework; i.e., the FrFT-based features are initially constructed and are subsequently fed into the network. When compared with the data-augmentation-based strategy, our method does not incur any additional computation cost.

The proposed FrFT convolutional face is based on the LeNet [8] network architecture. The FrFT convolutional face begins with a 5×5 convolutional layer and a 2×2 pooling layer; further, it contains another 5×5 convolutional layer and 2×2 pooling layer; finally, it contains a 5×5 convolutional layer and one completely connected layer with a softmax activation function.

Experimental results. We used the following two common face datasets: Extended-YaleB (Ex-YaleB) and AR [9] to quantitatively evaluate the performances of the proposed method. The Ex-YaleB dataset comprises the frontal face images of 38 persons, and each image is captured under 64 illumination conditions without any occlusion, whereas AR comprises the frontal face images of 100 persons captured using various combinations of illumination, partial occlusion, and facial ex-

pressions. The order p of FrFT is an important parameter in our method, which considerably affects the recognition performance. Figure 1(b) presents the precision obtained using different values of p ranging from 0 to 1 at an interval of 0.1. Note that we have determined the network parameters by performing ten-fold cross-validation using the training and validation sets.

(1) The proposed FrFT convolutional face (FrConf) exhibits a higher recognition performance than that exhibited by FrFT, FFT, and LeNet, as denoted in Figure 1(b), because of the robust FrFT-based input and the powerful learning ability of LeNet. (2) Further, the recognition precision of $\mathcal{ABIP}$ in both the datasets is higher than that of the other methods mentioned in this study. The FrFT-based input with more than one order (p value) tends to yield a better result, as depicted in Figure 1(b). Furthermore, the image texture information and illumination changes were well captured by the $\mathcal{IP}$ component, and $\mathcal{AB}$ is sensitive to the image change in the local frequency domain, which has been used to detect the partial facial occlusion. (3) $\mathcal{ABIP}$ + FrConf with cumulative p values of 0–0.8 achieved the highest recognition precision in both the datasets. The introduction of different frequency information by increasing the order p facilitates network training in cases of illumination, occlusion, and facial expressions. Once the value of p exceeds the saturation point (0.8), the classification performance begins to degrade. (4) We evaluated the performance of the proposed FrConf using different FrFT-based feature combinations as the networks input. Figure 1(b) depicts that LeNet with an input of $\mathcal{ABIP}_p$ outperforms the other four inputs, demonstrating the effectiveness and superiority of the proposed framework.

Conclusion. In this study, we reviewed the LeNet deep learning architecture that encounters various challenges such as facial deformations. Further, we proposed the FrFT convolutional face by introducing a multi-channel spatial-to-frequency domain transform. The proposed framework is considered to be sufficiently robust to mitigate the effect of illumination, posture, and occlusion. We conducted quantitative experiments using the Ex-YaleB and AR face datasets. Our results demonstrated that the proposed FrFT convolutional face improved the recognition performance. Future studies should investigate the potential of the proposed method by applying a more advanced network structure and designing a novel end-to-end learning strategy.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61421001, U1833203).

References

1 Hong D F, Liu W Q, Su J, et al. A novel hierarchical approach for multispectral palmprint recognition. Neurocomputing, 2015, 151: 511–521
2 Li X L, Cui G S, Dong Y S. Discriminative and orthogonal subspace constraints-based nonnegative matrix factorization. ACM Trans Intell Syst Technol, 2018, 9: 1–24
3 Hong D F, Liu W Q, Wu X, et al. Robust palmprint recognition based on the fast variation Vese-Osher model. Neurocomputing, 2016, 174: 999–1012
4 Li X L, Cui G S, Dong Y S. Graph regularized nonnegative low-rank matrix factorization for image clustering. IEEE Trans Cybern, 2017, 47: 3840–3853
5 Li X L, Lu Q M, Dong Y S, et al. SCE: a manifold regularized set-covering method for data partitioning. IEEE Trans Neural Netw Learn Syst, 2018, 29: 1760–1773
6 Li X L, Cui G S, Dong Y S. Refined-graph regularization-based nonnegative matrix factorization. ACM Trans Intell Syst Technol, 2017, 9: 1–21
7 Wang X, Qi L, Tie Y, et al. Face recognition based on the band fusion of generalized phase spectrum of 2D-FrFT. In: Proceedings of the International Conference on Graphics and Image Processing (ICGIP), 2016. 685: 137–145
8 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
9 Hong D F, Yokoya N, Xu J, et al. Joint & progressive learning from high-dimensional data for multi-label classification. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 469–484