Hindawi
Journal of Sensors
Volume 2021, Article ID 7129800, 7 pages
https://doi.org/10.1155/2021/7129800

Research Article

Application of Multiscale Facial Feature Manifold Learning Based on VGG-16

Huilin Ge, Zhiyu Zhu, Runbang Liu, and Xuedong Wu

School of Electronic Information, Jiangsu University of Science and Technology, Zhenjiang 212003, China

Correspondence should be addressed to Huilin Ge; ghl1989@just.edu.cn

Received 20 April 2021; Revised 21 May 2021; Accepted 18 June 2021; Published 25 August 2021

Academic Editor: Kelvin Wong

Purpose. In order to solve the problems of small face image samples, high size, low structure, no label, and difficulty in tracking and recapture in security videos, we propose a popular multiscale facial feature manifold (MSFFM) algorithm based on VGG16. *Method*. We first build the VGG16 architecture to obtain face features at different scales and construct a multiscale face feature manifold with face features at different scales as dimensions. At the same time, the recognition rate, accuracy rate, and running time are used to evaluate the performance of VGG16, LeNet-5, and DenseNet on the same database. *Results*. From the results of comparative experiments, it can be seen that the recognition rate and accuracy of VGG16 are the highest among the three networks. The recognition rate of VGG16 is 97.588%, and the accuracy is 95.889%. And the running time is only 3.5 seconds, which is 72.727% faster than LeNet-5 and 66.666% faster than DenseNet. *Conclusion*. The model proposed in this paper breaks through the key problem in the face detection and tracking problem in the public security field, predicts the position of the face target image in the time dimension manifold space, and improves the efficiency of face detection.

1. Introduction

Face recognition [1–5], as a biometric recognition technology, is one of the hot topics in the research fields of pattern recognition, image processing, machine vision, neural networks, and cognitive science in recent years. At the same time, face recognition, as a biometric identification technology with high stability, high accuracy, difficult to copy, and easy to be accepted by humans [6–9], has a wide range of application prospects in the fields of identity authentication, security monitoring, human-computer interaction, etc. With the increasing innovation of information technology, the processing of images by face recognition technology has become more and more complex. With sufficient samples, single background, and stable ambient light, most algorithms can achieve higher recognition rates.

In practical applications, how to solve the influence of environmental factors, human sentiments, and posture changes has become a difficult problem in testing various algorithms. The face recognition algorithm based on eigenface extracts face features by way of dimensionality reduction. Although the computational complexity is reduced, some effective features will be lost while reducing the dimensionality.

In order to reflect the nonlinear structure of image features, two classical manifold learning methods have been proposed by previous researchers, which are ISO metric mapping (ISOMAP) [10–12] and local linear embedding (LLE) [13–15]. By learning the mapping from environmental space to eigenspace, the structure between adjacent points after projection can be preserved. Although such manifold learning methods can model the manifold structure of the data, they require a large amount of dense data as training samples, which is not applicable to some practical applications. Therefore, this paper constructs convolutional neural network (CNN) architecture to obtain face features at different scales, so as to solve the problem of insufficient sample size.

In recent years, CNN [16–20] has become a research hotspot in the field of speech analysis and image recognition, especially in the field of face recognition. The CNN makes full use of the locality of the data itself by combining the local perception area of the face image, sharing the weight, and spatially. This feature has a certain degree of robustness to illumination changes, posture, and occlusion.

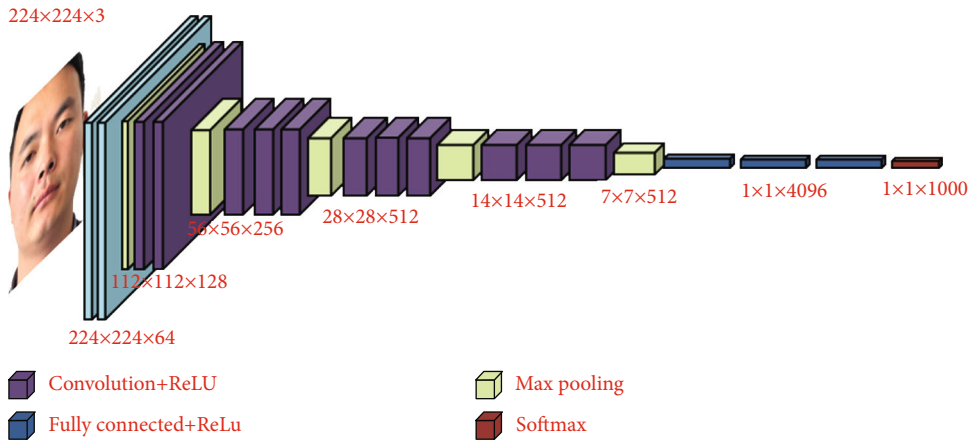

Figure 1: Network structure of VGG16.

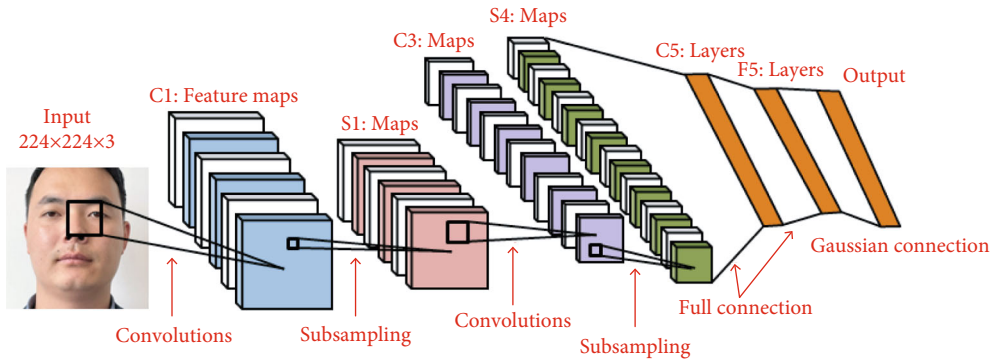

Figure 2: The network structure of LeNet-5.

Previously, there have been many studies on face recognition, such as methods based on statistical manifolds [21–24] and methods combined with deep neural networks. It can be seen that the image classification method based on the manifold will still maintain considerable research enthusiasm for a long time in the future.

In this paper, a CNN is used to construct a face recognition system. First, the VGG16 model [25–29] is used to extract facial features, and then multiscale facial feature manifold (MSFFM) [30, 31] is used for classification and tested in the actual environment.

2. Methodology

2.1. CNN Architecture

2.1.1. VGG16. VGGNet was proposed by the Oxford Visual Geometry Group of Oxford University [32]. It explored the relationship between the depth of CNN and its performance. By repeatedly stacking 33 small convolution kernels and 22 maximum pooling layers, it successfully constructed CNN with 16 to 19 layers deep. VGGNet is a modification based on AlexNet [33]; the training image size is 224×224. All images are subtracted from the mean of all training images. VGGNet contains many levels of networks, ranging in depth from 11 to 19 layers. The more commonly used ones are VGGNet-16 and VGGNet-19. VGGNet divides the network into 5 segments, and each segment connects multiple 3∗3 convolutional networks in series. Each segment of convolution is followed by a maximum pooling layer, and the last is 3 fully connected layers and a softmax layer. The network structure of VGG16 is shown in Figure 1.

2.1.2. LeNet-5. LeNet-5 [34] is a classic structure of CNN, the pioneering work of CNN, mainly used for handwritten font recognition. Although the network is simple, the structure is complete, and the convolutional layer, pooling layer, and full link layer have been used until now. The number of layers is very shallow, and the size of the kernel is single. The kernel sizes used by the three convolutional layers of C1, C3, and C5 are all 5×5. The feature map size of C5 is 1×1 because the feature map size of S4 is 5×5 and the kernel size is the same, so the result size of the convolution is 1×1. The window size used by the two pooling layers of S2 and S4 is 2×2, and there are two types of pooling here. F6 is a fully connected layer with 84 neurons. The network structure of LeNet-5 is shown in Figure 2.

2.1.3. DenseNet. DenseNet [35] breaks away from ResNet's shortcomings of deepening the number of network layers [36] and Inception's shortcomings of widening network structure to improve network performance [37]. From the

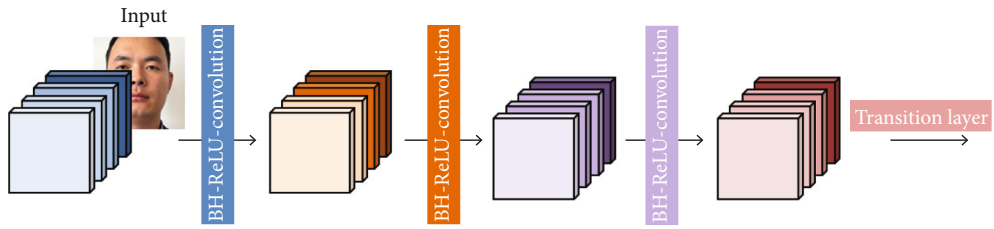

Figure 3: DenseNet's network structure.

perspective of features, through feature reuse and bypass settings, it has greatly reduced the parameter quantity of the network alleviates the emergence of gradient vanishing problem to a certain extent. Combining the assumptions of information flow and feature reuse, DenseNet deserves to be the best paper of the year at the 2017 Computer Vision Conference. DenseNet has absorbed the most essential part of ResNet and has done more innovative work on this, which further improves the network performance.

DenseNet is a CNN with dense connections. In this network, there is a direct connection between any two layers, that is, the input of each layer of the network is the unions of the outputs of all the previous layers, and the feature map learned by this layer will also be directly passed to all subsequent layers that are used as input. The network structure of DenseNet is shown in Figure 3.

2.2. Manifold Learning

2.2.1. Manifold Types Commonly Used in Image Classification. In the field of computer vision, the covariance matrix conforming to the manifold geometry of the symmetric positive definite matrix has been proven to have very good effects in image classification tasks. Because the covariance matrix can better adapt to various types of image changes, it has become a mainstream image feature representation method in image classification based on Riemannian manifolds. Commonly used manifold types for image classification are symmetric positive definite matrix manifold and Gmssmami manifold.

2.2.2. Kernel Function on Manifold. In the field of machine learning, the kernel method is a type of learning algorithm used to solve pattern recognition problems. The most classic example is the support vector machine (SVM). The main idea is to embed the data in the original space into a specific high-dimensional space through some kind of implicit nonlinear mapping, so that the linearly inseparable data in the original space becomes linearly separable after being mapped to the high-dimensional space.

Literature [18] presents a kernel function on the manifold of a symmetric positive definite matrix. According to the Frobenius norm and polarization formula, the inner product of two n-dimensional symmetric positive definite matrices X_1 and X_2 in the tangent space Tr Symn is defined as Equation (1).

$$\{\log(X_1), \log(X_2)\} = Tr\{\log(X_1)\log(X_2)\}. \quad (1)$$

It can be seen that the corresponding kernel function on Symn is defined as Equation (2).

$$k_L(X_1, X_2) = Tr\{\log X_1 \log X_2\}. \quad (2)$$

2.2.3. Face Super-Resolution Algorithm Based on Sparse Representation. The SR method adds sparse representation theory on the basis of manifold learning, uses a subset of the training sample block to linearly represent the input low-resolution image block, and uses the L1 norm to solve the optimal weight coefficient. We use the most similar face training sample block to represent the effect of the input image block.

For each image block $X(i, j)$ of the input low-resolution image, all training sample blocks at the same position in the low-resolution sample space are sparsely learned to reconstruct the representation coefficients, and the objective function is expressed as Equation (3).

$$argmin\|w(i,j)\| \, s.t. \left\| X(i,j) - \sum_{m=1}^{M} Y^m(i,j) w_m(i,j) \right\|_2^2. \quad (3)$$

Then, the objective function is transformed into Equation (4) for solving.

$$argmin \left\| X(i,j) - \sum_{m=1}^{M} Y^m(i,j) w_m(i,j) \right\|_2^2 + \rho \|w(i,j)\|_1. \quad (4)$$

We linearly weigh the representation coefficient obtained by Equation (4) and the corresponding high-resolution training sample block to obtain the high-resolution prediction image block. This algorithm solves the problem that the solution is not unique in the location-based image block algorithm.

2.3. Data Set

2.3.1. Data Sources. In practical applications, the biggest challenge of the face recognition system is that the recognition effect is not ideal when there are pose changes and occlusions. Therefore, this paper collects many types of face images in practical applications as experimental samples. The normal face sample is basically aligned or inclined at a small angle, without occlusion, and single expression, as shown in Figure 4(a). The face samples with posture changes have a variety of expressions and side faces as shown in Figure 4(b). A sample of an occluded face is shown in Figure 4(c). In the system test, 20 face images of each of 36

(a) (b) (c)

Figure 4: Face samples in various situations based on (a) the normal face sample, (b) the face samples with posture changes, and (c) the face samples with an occluded face.

humans were collected, and a total of 720 images were collected. The classification and pixels of the data set are shown in Table 1.

2.3.2. Experimental Environment. The software platform, processor, and operating system of the experimental environment are shown in Table 2. In order to ensure the fairness of the experimental results, all methods are implemented using the source code provided by the original author as much as possible, and the main parameters are adjusted and set according to the instructions in the original document.

2.4. Evaluation Criteria. This paper uses recognition rate, accuracy rate, and running time to evaluate the performance of VGG16, LeNet-5, and DenseNet.

First of all, the recognition rate refers to the ratio of the number of all recognized images to the total images, and the calculation method is as shown in Equation (5).

$$\text{Recognition rate} = \frac{n_R}{N}. \tag{5}$$

In Equation (5), n_R refers to the number of all recognized images, and N knows that it is the total number of images.

Second, accuracy refers to the ratio of the number of correctly identified images to the total number. The calculation method is as shown in Equation (6).

$$\text{Accuracy} = \frac{n_A}{N}. \tag{6}$$

In Equation (6), n_A refers to the number of correctly identified images, and N refers to it is the total number.

3. Experimental Results

3.1. Recognition Rate. Figure 5 shows the comparison curve of the recognition rate of VGG16, LeNet-5, and DenseNet.

It can be seen from Figure 5 that the recognition rate is proportional to the number of samples, and the recognition rate of MSFFM based on VGG16 is the highest. When the number of samples is 240, the recognition rate is 97.588%.

3.2. Accuracy. Figure 6 shows the accuracy comparison curve of VGG16, LeNet-5, and DenseNet. It can be seen from the comparison result that the accuracy rate of the face recognition algorithm based on VGG16 is 95.889%.

Table 1: Classification and pixels of the data set.

Normal	Posture change	Covered	Pixel
240	240	240	224×224

Table 2: Software platform, processor, and operating system of the experimental environment.

Software platform	Processor	Operating system
Matlab 2020a	Intel E7-2684 v5	IOS

3.3. Computational Time. By comparing the data in Table 3, it can be seen that on the database, the time required for the three network pairs to complete an operation is about 3.5 seconds, 9.1 seconds, and 11.6 seconds, respectively, and the model we proposed can shorten the calculation time by nearly half. In contrast, our proposed VGG16 has achieved excellent results in reducing computational complexity. In general, VGG16 can effectively reduce the computational complexity, so it has better feasibility in practical applications.

4. Discussion

This paper introduces a CNN that integrates manifold learning. It uses the spatial manifold information of the image as an additional feature and integrates it into the improved CNN model, so as to improve the pertinence of the model to the data and improve the accuracy. The model proposed in this paper makes up for the lack of generalization ability.

Facial super-resolution is a specific scene application of image super-resolution technology, and facial super-resolution has attracted widespread attention from scholars. In actual scenes, the acquired face images are usually blurry and low quality, which is caused by a variety of reasons.

First, the location of the surveillance camera is high, the shooting range is large, and the target face image is small; second, the surveillance equipment is limited by storage space, and the video image is highly compressed, so the image loses

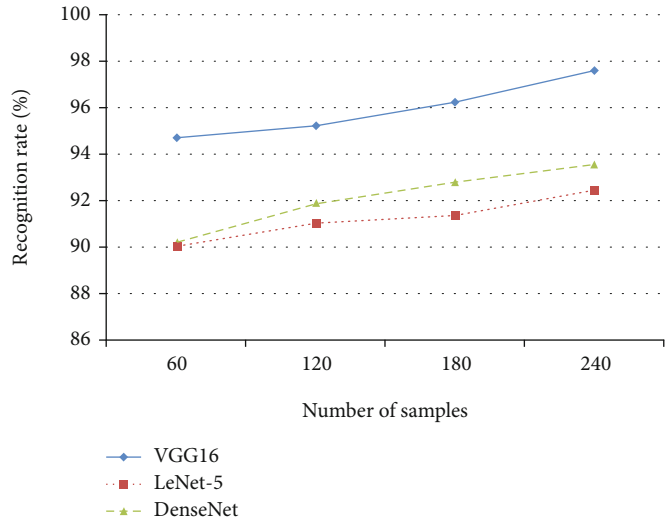

Figure 5: The recognition rate of VGG16, LeNet-5, and DenseNet.

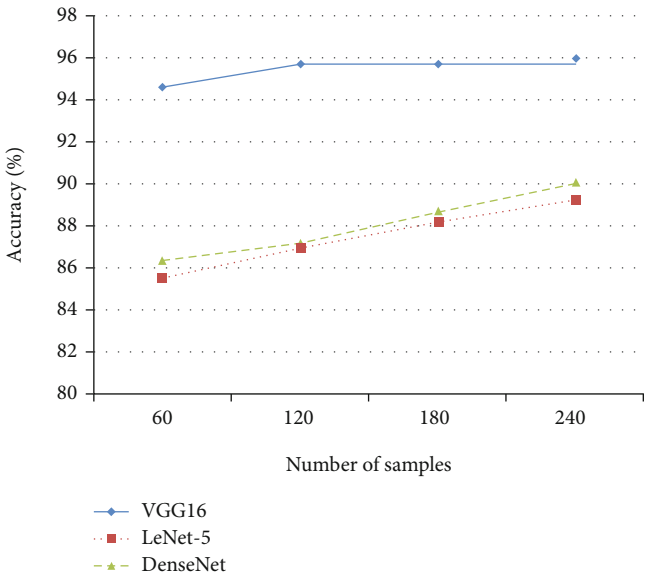

Figure 6: The accuracy comparison curve of VGG16, LeNet-5, and DenseNet.

Table 3: The time required for the three types of networks to complete an operation.

VGG16 (s)	LeNet-5 (s)	DenseNet (s)
3.5	9.1	11.6

detailed information; third, the external environment is rainy weather and low light at night will further reduce the quality of the captured images.

The existing image feature representation methods focus on matrix-type manifolds. How to fuse multiple types of Riemannian manifolds, such as linear subspaces, probability distributions to form a multimodel representation method, and perform different manifolds the effective unification of the characteristic information will be a problem worthy of attention.

With the continuous deepening of research on CNN, vector-based convolution and pooling processing have been fully studied. In fact, in a network, we can apply Riemannian manifold geometry to the data in the middle layer for processing. This pooling and iterative process in the form of a matrix can have a positive effect on the final output of the network. For future implementation, we will continue to carry out related research in this direction.

5. Conclusion

We researched and adopted a popular multiscale facial feature algorithm based on VGG16 and designed and implemented face recognition on this basis. The system first intercepts each frame of image in the video stream for face detection and then recognizes the detected faces. The actual test results show that the system has a high recognition rate for face pose, expression, and occlusion changes when the training samples are sufficient. There is still a lot of research space for the algorithms based on manifold learning in face recognition and the application of these algorithms in face recognition systems.

Data Availability

The [face image] data used to support the findings of this study have not been made available because [we collected face images on our own which includes faces of members from our project. They are not willing to give their information to the public].

Conflicts of Interest

The authors declare that there is no conflict of interests.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62006102).

References

[1] J. H. Lai, P. C. Yuen, and G. C. Feng, "Face recognition using holistic Fourier invariant features," *Pattern Recognition*, vol. 34, no. 1, pp. 95–109, 2001.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 55, no. 1, pp. 307–316, 2009.

[3] P. J. Phillips, Hyeonjoon Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[4] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Asahimas Flat Glass*, vol. 4778, no. 6, pp. 1635–1650, 2007.

[5] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[6] T. Grüter, M. Grüter, and C. C. Carbon, "Neural and genetic foundations of face recognition and prosopagnosia," *Journal of Neuropsychology*, vol. 2, no. 1, pp. 79–97, 2008.

[7] D. Masip, G. Lapedriza, and J. Vitria, "Boosted online learning for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 530–538, 2009.

[8] N. Pinto and D. Cox, "An evaluation of the invariance properties of a biologically-inspired system for unconstrained face recognition," in *ICST Conference on Biologically Inspired Network, Information, and Computing Systems*, pp. 505–518, Berlin, Heidelberg, 2010.

[9] B. Tistarelli, E. Grosso, and Y. Grosso, "Identity management in face recognition systems," *Lecture Notes in Computer Science*, vol. 5372, pp. 67–81, 2008.

[10] S. Weng, C. Zhang, and Z. Lin, "Exploring the structure of supervised data by Discriminant Isometric Mapping," *Pattern Recognition*, vol. 38, no. 4, pp. 599–601, 2005.

[11] X. H. Fu, "The isometric extension of the into mapping from the unit sphere $S_1(E)$ to $S_1l^{\infty}(\Gamma)$," *Acta Mathematica Sinica*, vol. 24, no. 3, pp. 87–92, 2008.

[12] N. Zhang and X. M. Tian, "Nonlinear dynamic fault detection method based on isometric mapping," *Journal of Shanghai Jiaotong University*, vol. 45, no. 8, pp. 1202–1206, 2011.

[13] G. Wen and L. Jiang, "Globalizing local neighborhood for locally linear embedding," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, pp. 3491–3496, Taipei, Taiwan, 2006.

[14] X. Teng, B. Wu, W. Yu, and C. Liu, "A hand gesture recognition system based on local linear embedding," *Journal of Visual Languages & Computing*, vol. 16, no. 5, pp. 442–454, 2005.

[15] S. F. Daniel, A. Connolly, J. Schneider, J. Vanderplas, and L. Xiong, "Classification of stellar spectra with local linear embedding," *The Astronomical Journal*, vol. 142, no. 6, p. 203, 2011.

[16] L. Shang, Q. Yang, J. Wang, S. Li, and W. Lei, "Detection of rail surface defects based on CNN image recognition and classification," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 45–51, Chuncheon, Korea (South), 2018.

[17] G. Zheng, M. Tan, J. Yu, Q. Wu, and J. Fan, "Fine-grained image recognition via weakly supervised click data guided bilinear CNN model," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 661–666, Hong Kong, China, 2017.

[18] Y. Chen and R. Qu, "Study on infringement identification of art works based on CNN image recognition technology," *Journal of Physics: Conference Series*, vol. 1802, no. 3, article 032084, 2021.

[19] R. Kumar, S. Joshi, and A. Dwivedi, "CNN-SSPSO: a hybrid and optimized CNN approach for peripheral blood cell image recognition and classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 47, no. 1, pp. 202–209, 2020.

[20] W. Hao, R. Bie, J. Guo, X. Meng, and S. Wang, "Optimized CNN based image recognition through target region selection," *Optik (Stuttg)*, vol. 156, pp. 772–777, 2018.

[21] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, 2011.

[22] P. K. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision," in *2008 IEEE computer society conference on computer vision and pattern recognition (CVPR 2008)*, Anchorage, Alaska, June 2008.

[23] C. Cafaro, "Information-geometric indicators of chaos in Gaussian models on statistical manifolds of negative Ricci curvature," *International Journal of Theoretical Physics*, vol. 47, no. 11, pp. 2924–2933, 2008.

[24] J. Verbeek, "Learning nonlinear image manifolds by global alignment of local linear models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 25, no. 1, p. 335, 2006.

[25] M. Valan, K. Makonyi, A. Maki, D. Vondráček, and F. Ronquist, "Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks," *Systematic Biology*, vol. 68, no. 6, pp. 876–895, 2019.

[26] N. V. Hieu and N. Hien, "Automatic plant image identification of Vietnamese species using deep learning models," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 68, no. 4, pp. 25–31, 2020.

[27] J. Qiu, X. Lu, X. Wang, and X. Hu, "Research on rice disease identification model based on migration learning in VGG network," *IOP Conference Series: Earth and Environmental Science*, vol. 680, no. 1, article 012087, 2021(10pp).

[28] U. B. Patayon and R. V. Crisostomo, "Automatic identification of abaca bunchy top disease using deep learning models," *Procedia Computer Science*, vol. 179, no. 1, pp. 321–329, 2021.

[29] J. Hagenah, M. P. Heinrich, and F. Ernst, "Deep transfer learning for aortic root dilation identification in 3D ultrasound images," *Current Directions in Biomedical Engineering*, vol. 4, no. 1, pp. 71–74, 2018.

[30] M. Cadoni, E. Grosso, A. Lagorio, and M. Tistarelli, "Interpreting 3D faces for augmented human-computer interaction," in *International Conference on Universal Access in Human-computer Interaction: Users Diversity*, pp. 535–544, Berlin, Heidelberg, 2011.

[31] B. Mandal, X. D. Jiang, and A. Kot, "Multi-scale feature extraction for face recognition," in *2006 1ST IEEE Conference on Industrial Electronics and Applications*, Singapore, 2006.

[32] X. Zhen, J. Chen, Z. Zhong et al., "Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study," *Physics in Medicine & Biology*, vol. 62, no. 21, p. 8246, 2017.

[33] A. V. Vedalankar, S. S. Gupta, and R. R. Manthalkar, "Addressing architectural distortion in mammogram using AlexNet and support vector machine," *Informatics in Medicine Unlocked*, vol. 4, article 100551, 2021.

[34] G. Wei, G. Li, J. Zhao, and A. He, "Development of a LeNet-5 gas identification CNN structure for electronic noses," *Sensors*, vol. 87, no. 2, pp. 147–156, 2019.

[35] E. Y. Huan and G. H. Wen, "Transfer learning with deep convolutional neural network for constitution classification with face image," *Multimedia Tools and Applications*, vol. 79, no. 4, pp. 110–119, 2020.

[36] W. E. Lawson, "Multi-attribute residual network (MAResNet) for soft-biometrics recognition in surveillance scenarios," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, USA, 2017.

[37] A. Satapathy and L. Livingston, "A lite convolutional neural network built on permuted Xceptio-inception and Xceptio-reduction modules for texture based facial liveness recognition," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 10441–10472, 2020.