PURSUING FACE IDENTITY FROM VIEW-SPECIFIC REPRESENTATION TO VIEW-INVARIANT REPRESENTATION

Ting Zhang[1], *Qiulei Dong*[1,2,3*], *Zhanyi Hu*[1,2,3]

1. National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, P. R. China
2. Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Beijing 100190, P. R. China
3. University of Chinese Academy of Sciences, Beijing 100049, P. R. China

ABSTRACT

How to learn view-invariant facial representations is an important task for view-invariant face recognition. The recent work [1] discovered that the brain of the macaque monkey has a face-processing network, where some neurons are view-specific. Motivated by this discovery, this paper proposes a deep convolutional learning model for face recognition, which explicitly enforces this view-specific mechanism for learning view-invariant facial representations. The proposed model consists of two concatenated modules: the first one is a convolutional neural network (CNN) for learning the corresponding viewing pose to the input face image; the second one consists of multiple CNNs, each of which learns the corresponding frontal image of an image under a specific viewing pose. This method is of low computational cost, and it can be well trained with a relatively small number of samples. The experimental results on the MultiPIE dataset demonstrate the effectiveness of our proposed convolutional model in contrast to three state-of-the-art works.

Index Terms— Face recognition, View-invariant representations, Convolutional neural network

1. INTRODUCTION

Over the last twenty years, there has been huge progress in face recognition. Deep learning has achieved great success on face recognition [2, 3, 4, 5, 6] and significantly outperformed the existing systems using low-level features [7, 8, 9, 10, 11, 12, 13, 14, 15]. The performances of face recognition systems depend heavily on face representation, which is naturally coupled with many types of face variations, such as viewing pose, illumination, expression, occlusion and so on. Among them view variation is a particular challenge because intra-person variance caused by view variation sometimes exceeds inter-person variance. Addressing this problem, a lot of methods have been proposed, which can be roughly divided into two categories: 2D techniques [2, 16, 17, 18] and 3D techniques [19, 20, 21].

The existing 3D methods usually estimate 3D models from 2D input or capture 3D facial data, and then match them to 2D probe face images. Asthana et al. [20] projected a non-frontal face image onto an aligned 3D face model, and then rotated it to render a frontal face image. Li et al. [21] used a set of 3D displacement fields sampled from a 3D face database to generate a virtual view for the probe image and compared the synthesized faces with each of the gallery faces.

Different from these 3D methods, the existing 2D methods attempt to handle view variations by 2D image matching or by encoding a test image with some exemplars. Li et al. [16] proposed an elastic matching method based on Gaussian Mixture Model (GMM) to align the patches and match the face images at different poses. Zhu et al. [2] proposed a deep neural network that can transform a face image with an arbitrary view and illumination to a frontal face image with neutral illumination. Yim et al. [17] presented a new deep architecture which can rotate a face image with an arbitrary view and illumination to a target-view face image.

The recent work [1] found that a macaque monkey has a face-processing system consisting of six interconnected face-selective regions, where neurons in some of these regions were view-specific. Motivated by this intriguing function of the macaque monkey's brain, we propose a novel deep convolutional learning model, called VS2VI, which learns view-invariant facial representations by explicitly enforcing this view-specific mechanism. There are two concatenated modules in our proposed model. The first one is to learn the viewing pose of each input image, while the second one contains multiple CNNs, each of which learns the corresponding frontal image of the input image according to the learnt viewing pose.

Our contributions are as two-fold: **1.** We propose a new deep model for face recognition, which learns view-invariant representations by explicitly enforcing the view-specific mechanism, inspired by the face-processing network

[*] Corresponding author: qldong@nlpr.ia.ac.cn

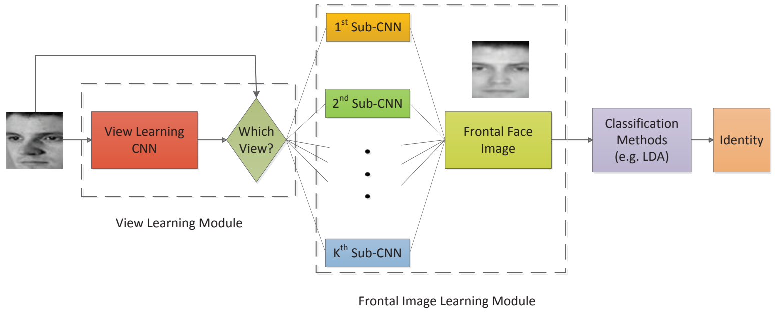


Fig. 1. Architecture of the proposed VS2VI network.

of the macaque monkey's brain. **2.** The proposed model can be trained with a small number of samples since deconvolutional layers instead of fully connected layers are used to reconstruct the frontal face images.

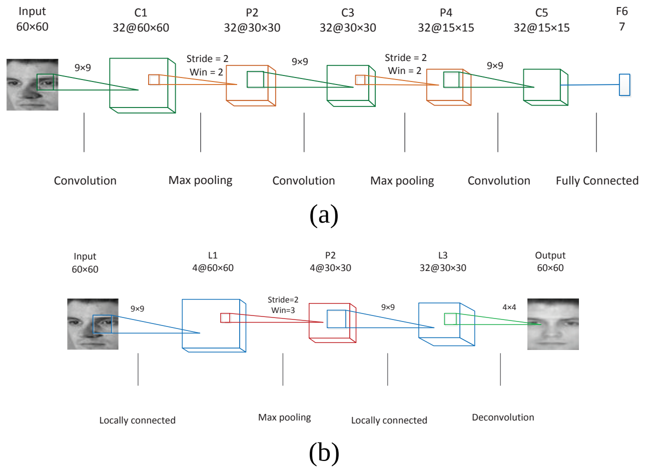


Fig. 2. (a) The architecture of the view learning module. (b) The architecture of the frontal image learning module.

2. MODEL DESCRIPTION

In this section, we propose the VS2VI model, consisting of two concatenated modules (view learning module and frontal image learning module). The architecture of the VS2VI model is show in Fig. 1. As is seen, given a face image of an arbitrary viewing pose, the view learning module classifies it into a specific viewing pose first. Then the face image is dealt with by the sub-network of the corresponding viewing pose in the frontal image learning module. Finally, a classification method (here, we simply use LDA (linear discriminant analysis)) is used to classify the output image of the frontal image learning module into a certain identity. LDA is not used in the view learning module. In this paper, the used images are grayscale with size 60×60. In the following subsections, the designed two modules are described in details.

2.1. View Learning Module

The view learning module learns the corresponding viewing pose to the input face image and its architecture is shown in Fig. 2(a). This module is composed by three convolutional layers, two max-pooling layers and a fully connected layer. The input of the network are face images and the output of the network is the probability of each possible viewing pose.

The whole set of parameters in the network is expressed as Input(60×60)-C$(32, 9)$-P$(2, 2)$-C$(32, 9)$-P$(2, 2)$-C$(32, 9)$-F(7). 'C', 'P', and 'F' respectively denote the convolutional layer, pooling layer, and fully connected layer. For the convolutional layer, the first number in the bracket indicates the number of the filters and the second indicates the filter size. Convolutional layers use ReLU [22] as the activation function. For the pooling layer, the first number in the bracket indicates the filter size and the second indicates the stride. The fully connected layer uses the softmax loss as cost function.

2.2. Frontal Image Learning Module

The frontal image learning module consists of several subnetworks determined by the view learning module, each of which learns the corresponding frontal image of an image under a specific viewing pose. The architecture of each single sub-network is shown in Fig. 2(b). The sub-network for each specific viewing pose is composed of two locally connected layers, a max pooling layer and a deconvolutional layer. The first three layers are proposed to extract features and the last layer is proposed to recover the frontal face images. The input and the output of the sub-network are both face images.

We design the frontal image learning module based on the following observations: 1. Since the features of different regions of face images are quite different, we use locally connected layers without weight sharing rather than standard convolutional layers which share weights, commonly used by many previous works [23, 22]. 2. Considering that the amount of samples in many real applications is not sufficient enough to learn millions of parameters, we use a deconvolutional layer instead of a fully connected layer as the last layer of the proposed network since fewer parameters are needed in a deconvolutional layer than in a fully connected layer. Only 0.46M parameters are needed for a deconvolutional layer, while 103M parameters are needed for a fully connected layer.

The whole set of parameters in sub-networks is expressed as Input(60×60)-L($4,9$)-P($3,2$)-L($32,9$)-D($1,4$). 'L', 'P' and 'D' respectively denote the locally connected layer, pooling layer and deconvolutional layer. For the locally connected layer, the first number in the bracket indicates the number of the filters and the second indicates the filter size. The whole locally connected layer has stride 1. A PReLU [24] is applied as activation function in the first locally connected layer, where the slopes of negative parts are learned from data rather than preset constants. D($1,4$) means that this layer applies 1 filter with weight sharing with size 4 and stride 2 to implement upsampling. Upsampling is implemented by bilinear interpolation in the deconvolutional layer. The deconvolutional layer uses ℓ_2-loss as cost function.

3. TRAINING

Our method is implemented with Caffe [25], which is one of the most popular deep learning frameworks.

The basic backpropagation (BP) is used to train the view learning module. To facilitate the optimization, a new layer-wisely training strategy is used to initialize the weight parameters in the frontal image learning module in two steps. In the first step, the first locally connected layer generates 4 feature maps from the input images. Then the average of feature maps is used to compare with the target frontal face image, and the simple network is trained with the ℓ_2-loss function. In the second step, the second locally connected layer generates 32 feature maps from the previous layer. The succeeding layer is pre-trained with the sample method while keeping the weight parameters in the pre-layer fixed.

We train our models using stochastic gradient descent with batch size of 64 examples, momentum of 0.9, and weight decay of 0.004. The weights of locally connected layer are initialized from a zero-mean Gaussian distribution with standard deviation 0.05. The learning rate is initialized at 0.001 and we divide it by 10 when the test loss stops decreasing.

4. EXPERIMENTS

The proposed method VS2VI is evaluated on the popular MultiPIE dataset [26], which contains images of 337 people of various views, illuminations and expressions. In addition, Zhu et al. [2] and Yim et al. [17] are tested on the same dataset for comparison. Similar to [2, 17], a subset of MultiPIE is used to evaluate all the referred method, which is same as Setting 1 in [17]: only images in session one which contains 249 subjects with neutral expression under all the 7 poses and 20 illuminations are adopted for training and test. The images of the first 100 identities are for training (14000 images in total), and the images of the remaining 149 identities for test. One frontal image of each identity in the test set is selected to the gallery. Since 7 images are produced for all the 7 poses for each gallery image in [17], to compare fairly, we produce 7 groups of gallery images by the sub-network of the corresponding viewing pose and select one group each time according to the pose. We denote the above method as VS2VI†. The used face images are aligned according to the position of eyes and mouths, and then converted to grayscale images. Each image is subtracted by the mean value over the training set.

	-45°	-30°	-15°	+15°	+30°	+45°	Avg
Li [27]	63.5	69.3	79.7	75.6	71.6	54.6	69.3
Z.Zhu [2]	**67.1**	74.6	86.1	83.3	75.3	**61.8**	74.4
VS2VI	62.3	**84.3**	**92.3**	**91.1**	**80.5**	58.4	**78.1**
CPI [17]	66.6	78.0	87.3	85.5	75.8	62.3	75.9
CPF [17]	73.0	81.7	89.4	89.5	80.4	70.3	80.7
VS2VI†	**95.6**	**94.8**	**97.8**	**96.1**	**94.0**	**92.3**	**95.1**

Table 1. Recognition rates (%) on different poses.

	00	01	02	03	04	05	06	08	09	10	11	12	13	14	15	16	17	18	19	Avg
Li [27]	51.5	49.2	55.7	62.7	79.5	**88.3**	**97.5**	**97.7**	**91.0**	79.0	64.8	54.3	47.7	67.3	67.7	75.5	69.5	67.3	50.8	69.3
Z.Zhu [2]	**72.8**	**75.8**	75.8	75.7	75.7	75.7	75.7	75.7	75.7	75.7	75.7	**75.7**	**75.7**	73.4	73.4	73.4	73.4	72.9	**72.9**	74.7
VS2VI	68.2	72.3	**76.3**	**76.8**	**79.8**	82.9	84.8	85.6	83.4	**80.8**	**78.2**	75.6	72.9	**79.8**	**78.5**	**81.9**	**80.2**	**79.0**	67.7	**78.1**
CPI [17]	66.0	62.6	69.6	73.0	79.1	84.5	86.6	86.5	84.2	80.2	76.0	70.8	65.7	76.1	78.2	80.7	79.4	77.3	65.4	75.9
CPF [17]	59.7	70.6	76.3	79.1	85.1	89.4	91.3	92.3	90.6	86.5	81.2	77.5	72.8	82.3	84.2	86.5	85.9	82.9	59.2	80.7
VS2VI†	**87.5**	**90.8**	**95.1**	**95.6**	**97.8**	**98.0**	**98.0**	**98.1**	**98.0**	**97.1**	**95.2**	**92.3**	**89.9**	**97.1**	**97.8**	**97.7**	**97.2**	**97.0**	**87.0**	**95.1**

Table 2. Recognition rates (%) on different illuminations.

4.1. Face Recognition

In this subsection, we compare our recognition results and the number of parameters with the state-of-the-art methods, and show the features of the output layer in the frontal face learning module.

In the test stage, we extract features from the output layer. LDA is used to classify the output frontal images from all the referred methods. The recognition rates of those methods for different poses are listed in Table 1 and for different illumi-

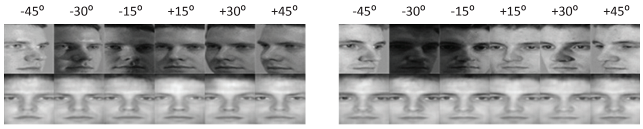


Fig. 3. Reconstructed examples: (top) Images of six poses and arbitrary illuminations for each identity. (bottom) Reconstructed frontal face images under neutral illumination.

Network	Z.Zhu [2]	Yim [17]	**VS2VI**
Number of Parameters	250.97	153.33	**65.81**

Table 3. Comparison of involved numbers of parameters in the referred methods. The least number of parameters is written in bold. The listed number of parameters for [17] only corresponds to the used number of parameters in the rotation task.

nations are in Table 2. Best results are written in bold. Experimental results are compared separately according to the number of gallery images.

As seen from Table 1, VS2VI outperforms the methods [27, 2] on four poses($+15°, +30°, -15°$ and $-30°$) and average recognition rate. In addition, VS2VI† outperforms the state-of-the-art for all the poses. The reason why recognition rates of $+45°, -45°$ are not good for VS2VI is that the number of training examples for $+45°$ and $-45°$ is really small. Only about 2000 images are provided for $+45°$ and $-45°$ in the training stage respectively. In our methods, face images of each viewing pose is trained individually. However, in the other works, all the images are trained together through one network and the training dataset is relatively large compared with our methods. If more data for $+45°$ and $-45°$ is provided, our method can also perform very well.

As seen from Table 2, VS2VI outperforms the methods [27, 2] for 10 out of 19 illuminations and VS2VI† outperforms the state-of-the-art for all the illuminations. The frontal illumination (ID 07) is not included in the test stage.

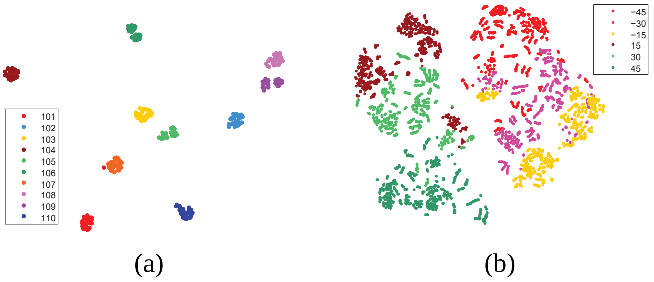
Fig. 3 shows several reconstructed face images, verifying that VS2VI can effectively remove the variation of viewing poses, meanwhile retain the intrinsic features of each identity.

The numbers of parameters of previous works [2, 17] and VS2VI are reported in Table 3. It shows that the number of parameters in VS2VI is less than those in [2, 17]. However, VS2VI still achieves comparable results on MultiPIE dataset.

Moreover, in order to verify the reconstructed frontal images by VS2VI are view-invariant, the t-SNE [28] method is used to transform the reconstructed frontal images from high-dimensional space into two-dimensional space. Fig. 4(a) shows the reconstructed frontal images of same identity are spatially close, while the reconstructed images corresponding to different identities are spatially discriminative.

view	-45°	-30°	-15°	+15°	+30°	+45°	Avg
accuracy	96.1	95.2	95.3	96.7	89.4	93.2	94.3

Table 4. Classification rates (%) on viewing poses.



(a)

(b)

Fig. 4. Results of the visualization. (a) Reconstructed frontal images of 10 identities (ID 101 to 110). The dot of the same color represents the same identity. (b) Distribution of the learnt features from the last convolutional layer in the frontal image learning module. The dot of the same color represents the feature of input images under the same viewing condition.

4.2. View Classification

In this subsection, we evaluate the VS2VI's ability for learning the viewing poses of the input images. The classification rate of viewing pose by VS2VI on each subset of images with a specific viewing pose ($-45°, -30°, -15°, 15°, 30°, 45°$) is reported in Table 4. As is seen, the classification rates of most viewing poses by the proposed VS2VI model are above 90%.

In addition, the features from the last convolutional layer are visualized using the t-SNE [28] method to demonstrate the effectiveness of the view learning module. Fig. 4(b) shows that the feature is discriminative among different viewing poses.

5. CONCLUSION

Inspired by the face-processing network of the macaque monkey's brain, we propose a deep model of two concatenated modules for face recognition. The first module learns the corresponding viewing pose to the input face image. The second module consists of several CNNs, each of which learns the corresponding frontal image of an image under a specific viewing pose. Extensive experimental results demonstrate the effectiveness of the proposed model. In the future, we will combine VS2VI with some classic face recognition approaches to further improve the performance.

6. ACKNOWLEDGEMENT

This work is supported by the Strategic Priority Research Program of the CAS (XDB02070002) and National Natural Science Foundation of China (61333015, 61273280, 61375042).

7. REFERENCES

[1] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science*, vol. 330, no. 6005, pp. 845–851, 2010.

[2] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *ICCV*, 2013.

[3] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[4] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.

[5] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014, pp. 1988–1996.

[6] Y. Taigman, M. Yang, M.A. Ranzato, and L. Wolf, "Web-scale training for face identification," *arXiv preprint arXiv:1406.5266*, 2014.

[7] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009.

[8] Y. Taigman, L. Wolf, T. Hassner, et al., "Multiple one-shots for utilizing class label information.," in *BMVC*, 2009.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009.

[10] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *CVPR*, 2011.

[11] C. Huang, S. Zhu, and K. Yu, "Large scale strongly supervised ensemble metric learning, with applications to face verification and retrieval," *arXiv preprint arXiv:1212.6094*, 2012.

[12] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *ECCV*, 2012.

[13] T. Berg and P. N. Belhumeur, "Tom-vs-pete classifiers and identity-preserving alignment for face verification," in *BMVC*, 2012.

[14] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013.

[15] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *ICCV*, 2013.

[16] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *CVPR*, 2013.

[17] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim, "Rotating your face using multi-task deep neural network," in *CVPR*, 2015.

[18] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NIPS*, 2014, pp. 217–225.

[19] X. Zhang and Y. Gao, "Face recognition across pose: A review," *PR*, vol. 42, no. 11, pp. 2876–2896, 2009.

[20] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *ICCV*, 2011.

[21] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *ECCV*, 2012.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, et al., "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[27] A. Li, S. Shan, and W. Gao, "Coupled bias–variance tradeoff for cross-pose face recognition," *TIP*, vol. 21, no. 1, pp. 305–315, 2012.

[28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.