LETTER

Image Pattern Similarity Index and Its Application to Task-Specific Transfer Learning

Jun WANG†, Nonmember, Guoqing WANG†(a), Student Member, and Leida LI†, Member

SUMMARY A quantized index for evaluating the pattern similarity of two different datasets is designed by calculating the number of correlated dictionary atoms. Guided by this theory, task-specific biometric recognition model transferred from state-of-the-art DNN models is realized for both face and vein recognition.

key words: face recognition, vein recognition, transfer learning, sparse representation

1. Introduction

Transfer learning based feature extraction refers to the situation that a state-of-the-art Deep Convolutional Neural Network (DCNN) model trained on a large image dataset can be adopted as a universal image feature descriptor, and doing so leads to impressive performance for a variety of image recognition tasks [1] especially for the condition that the supply of target training data is limited [2]. Traditional transfer learning models (as illustrated in Fig. 1) usually adopt the activations from a single DCNN layer (the second fully-connected layer) as final representation. The performance, however, will be degraded greatly if the pattern distribution between the source training data (domain A) and the target training data (domain B) is greatly different, thus making the idea that adopting state-of-the-art DCNN model (such as VGG [3] and AlexNet [4]) trained on large-scaled dataset for generating discriminative feature representation for small-scaled dataset unreliable, and making great similarity between domain A and domain B a necessity for generating effective and discriminative feature representation.

To solve the problem that there is not a standard criterion for evaluating the pattern similarity of images from two different domains, an quantized image pattern similarity index is formulated as calculating the number of correlated dictionary atoms from the perspective of reconstructing image based on sparse dictionary. Based on the proposed theory, simple pattern similarity evaluation experiments with different face images [5, 6], PolyU palmprint [7] and fingerprint images [8] and also the lab-made hand-dorsa vein images are conducted and results as expected are obtained.

On the other hand, driven by the assumption that better feature representation could be obtained by adopting the DCNN trained with dataset which lies in a more similar feature space with the target dataset as feature extractor, we design a hierarchical transfer learning framework where the higher similarity dataset (generated by the designed similarity index) is incorporated for re-training the original DCNN so as to enhance the ability of the model in exploiting the most discriminative feature representation of the target dataset for recognition.

Such hierarchical transfer learning procedure is evaluated with hand-dorsa vein image based gender and identity recognition experiments, and the competitive accuracy fully demonstrates the effectiveness of the proposed model. Further experiments with face recognition also verify the conclusion.

2. Quantized Evaluation on Shared Patterns

In this part, the sparse representation [9], which involves generating common dictionary with training images by optimizing a pre-defined objective function followed by reconstructing testing images with the dictionary, is introduced for figuring out the shared patterns by designing a structural sparse dictionary, where the atoms representing the similarity of input images are obtained. The atoms to dictionary is what the words to sentence, and the atoms are usually generated by clustering algorithm, and dictionary is just a codebook by grouping the learnt atoms, and the number of the shared atoms is the objective quantized index for evaluating the shared patterns between different images.

Note that the defined dictionary model is not used for image reconstruction or classification, and all the images are gathered for dictionary generation. Let \( E_i \in \mathbb{R}^{d \times M_i} \), \( i = 1, \cdots , N \), represent a group of evaluation samples of class \( i \), and \( D_i \in \mathbb{R}^{d \times A_i} \) as the corresponding dictionary, and
$d$ is the dimension of evaluated sample, $M_i$ is the number of evaluated samples belonging to $i$th class, and $A_i$ represent the number of atoms with dictionary $D_i$. Assuming that there exist shared patterns between the input, the learned dictionary could be divided into two constitute parts: (1) a group of shared atoms denoted as $D_s \in \mathbb{R}^{d \times A_s}$ and (2) remaining group of atoms denoted as $D_r \in \mathbb{R}^{d \times (A_i - A_s)}$. Given that $D_i = [D_s, D_r]$, $C_i = [C_s, C_r]$ the learning algorithm for generating specific dictionary with $N$ categories could be figured out by minimizing a pre-defined objective function as expressed in (1):

$$\min_{D_s, D_r, C_i} \sum_{i=1}^{N} \|[E_i - (D_s, D_r) [C_s, C_r]]_F^2 + \lambda \|C_i\|_1\]$$

(1)

And $C_i = [c_{i1}, \ldots, c_{iM_i}] \in \mathbb{R}^{A_i \times M_i}$

(2)

Where $C_i$ is defined as the corresponding sparse coefficient matrix of sample $E_i$ over the learned dictionary $D_i$, $\lambda$ is defined as sparsity constraint coefficients with scalar value. The TwIST [10] algorithm is adopted to solve the optimization problem, and extra threshold $\tau$ value. The TwIST [10] algorithm is adopted to solve the optimization problem, and extra threshold $\tau$ is added during optimization to evaluate the similarities between iteratively optimized dictionary matrices column by column.

$$\min_{D_s, D_r} \|[E_i - D_s C_s - D_r C_r]_F^2\] \text{ s.t. } \|d_i\|_2^2 \leq 1, \forall i = 1, \ldots, A_i$$

(3)

$$\min_{D_s, D_r} \|[E_i - D_s C_s]_F^2\] \text{ s.t. } \|d_i\|_2^2 \leq 1, \forall i = 1, \ldots, A_0$$

(4)

Where $C_i^r \equiv [C_i^r, \ldots, C_i^N]\]$

$$E_a \equiv [E_1 - C_i^r D_i^r; \ldots; E_N - C_N^r D_N^r]\]$$

The specific procedure for quantizing the shared patterns is described as follows:

**Input:** Sample data matrix $\{E_i\}_{i=1}^{N}$, size of learned dictionary $A_i$, $i = 1, \ldots, N$, sparsity constraint parameter $\lambda$ and similarity threshold $\tau$

1. **Initialization of $D_i$ and $A_i$**
   (a). Initializing the element of $\{D_i\}_{i=1}^{N}$ and $\{C_i\}_{i=1}^{N}$ according to the size of input image and database
   (b). Fix $D_i$ and update $A_i$ class by class by solving
   $$\min_{D_s, D_r} \|[E_i - D_s C_s]_F^2 + \lambda \|C_i\|_1\]$$
   (c). Similarly, fix $C_i$ and update $D_i$ class by class by solving
   $$\min_{D_s, D_r} \|[E_i - D_s C_s]_F^2\]$$
   using the Lagrange dual of (1)
   Repeat step (a-c) until convergence.

2. **Initial Shared-Dictionary Generation**
   For each dictionary $\{D_i\}_{i=1}^{N}$, calculate the inner product column by column, and stack those vectors whose inner product is bigger than the predefined threshold $\tau$ to form the initial $D_s$.

3. **Optimized and Complete Shared-Dictionary Generation**
   (d). Compute the initial $D_s$ corresponding to the initial $D_s$
   (e). Forming $D_i = [D_s, D_r]$ and fix it, update $A_i$ class by class by solving
   $$\min_{D_s, D_r} \|[E_i - D_i C_i]_F^2\]$$
   using the Lagrange dual of (1)
   (f). Fix $A_i$ and update $D_i$ by solving the dual value of Eq. (3)
   (g). Similarly, solving the dual of (4) to update $D_r$
   Repeat step (d-g) until convergence to obtain the optimized and complete shared dictionary.

Based on the procedure described above, we design experiments to generate shared dictionary with all databases described in Sect. 3. Before feeding the mixture mode of images into the model, the size of each input is normalized as 256*256. Besides, we set the dictionary size of each mode to be equal for simplicity, and the parameters setup are “$\lambda = 0.2$” and “$\tau = 0.9$”. For simplicity and consistence with lab-made database in class numbers, only some representative experiments are conducted and the samples randomly selected from VGG face, LFW, PolyU NIR face and ILSVRC are 200 respectively. The number of the shared-atom for those modes could be referenced from Table 1, and it should be noted that each result is the one after convergence, and also the results are as expected. What’s more, we argue that the proposed quantized model is also applicable for other image pattern similarity analysis tasks.

Judging from the similarity evaluation results in Table 1, the different number with different modes fully demonstrates the effectiveness of the proposed sparse representation based image similarity index, and we tend to attribute such results to the fact that the more factor they share, the higher index would be obtained. For example, the palmprint/fingerprint share the texture feature with hand-dorsa vein, the VGG Face shares the attribution of face, and the NIR Face shares near-infrared imaging condition with the

| Database Setup | Number of Shared-atoms | Shared Ratio |
|----------------|------------------------|-------------|
| Palmprint + Vein | 125 | 48% |
| Fingerprint + Vein | 132 | 56% |
| Palmprint + ILSVRC | 32 | 7% |
| Vein + ILSVRC | 16 | 4% |
| VGG Face + LFW | 168 | 38% |
| VGG Face + ILSVRC | 39 | 12% |
| NIR face + Vein | 95 | 32% |
| NIR face + VGG Face | 65 | 25% |
| NIR face + LFW | 52 | 20% |

**Fig. 2** Proposed framework with AlexNet and VGG-16 net
3. Database for Feature Extractor Generation

To take advantage of the transferability of existing model for semantic hand vein feature extraction, a more similar distribution with target database (the lab-made hand-dorsa vein image) but relatively large-scaled database is necessary for model re-training, and the PolyU palmprint [7] with 7K images and the fingerprint database [8] with 3K images are adopted to re-train new high-level vein feature extractor based on the original DCNN. The additional face recognition experiments are also designed both with the 2.6M VGG face database [5] and the 1.3K LFW databases [6]. Besides, another one is the lab-made hand-dorsa vein database containing 98 females and 102 females whose ages vary from 19 to 62. For each sample, 10 hand-dorsal vein images were acquired in two specifically set sessions separated by a time interval of more than 10 days, and at each time, five images were acquired from each subject at the wavelength of 850nm, Fig. 3 shows some image samples of male and female in the dataset. To the fullest of the dorsal vein information, we set the size of the images as 460*680 with extremely high-quality.

Apart from keeping gender as 1:1, diversified samples differ in ages; hand thickness as well as capturing session is included in the database. What’s more, it should be noted that the original models are also adopted directly as feature extractor with the vein images and face images for results comparison.

4. Recognition Experiments with Generated Feature Extractor Re-Trained under Hierarchical Transfer Learning

4.1 Model Selection and Re-Training

VGG [3] and AlexNet [4] models are selected for direct feature extraction and re-training due to their great performance on ILSVRC recognition challenge, and the second fully connected layer (FC7) is used for feature extraction, and both models could learn the high-level 4096-dimensional feature vector. Apart from re-training the model with the selected databases, the parameters configuration of the two models are both assigned with a weight decay of 0.0005, a momentum of 0.9, a $\gamma$ of 0.1 with the initial learning rate of 0.001. Besides, the re-training iterations for VGG and AlexNet are 30000 and 50000 respectively.

4.2 Experimental Results and Analysis

After obtaining discriminative feature representation with both the original and re-trained DCNN models, the simple but effective linear-SVM is adopted as a basis for robust classifier design, and the entire framework of the proposed strategy is as illustrated in Fig. 2. It should be noted that Data I in Fig. 2 represents the one selected by the proposed similarity index for re-training feature extractor (FC7*) and Data II is the target one for recognition experiments.

In the gender classification experiment with hand-dorsa vein information, the linear-SVM is adopted for vein and face recognition by being directly used as bi-class classifier and multi-class classifier with combination of bi-class linear-SVM training within grouped samples. The distribution of the dataset for different task share the same ratio (train : validation : test = 0.67 : 0.08 : 0.25). The parameters of the classifier during training is obtained in grid-search manner with five-fold cross-validation, and the corresponding cost values for linear kernel is set as $C = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, and the corresponding classification results with specific experimental setup could be referenced from Table 2.

Judging from the experimental results in Table 2, it could be concluded that whatever the biometric recognition task (vein based gender classification or identity recognition, face recognition), all the average classification results share similar trends: the model re-trained with the biometrical database holding higher similarity index performs far better than those with the large-scaled ILSVRC database, which seems against the common theory that larger database would generate better model with greater representation ability. By analyzing the results, reasonable conclusion could be drawn that the feature generated from the re-trained DCNN model with the database lies in a more similar feature space with the target database is more discriminative than the one generated directly from the DCNN trained with a less-similar database despite the scale of the database for
Table 2  Average recognition accuracy with different models.

| Task                  | Model Re-training Database | Experimental Modes | Accuracy |
|-----------------------|-----------------------------|--------------------|----------|
| Gender Classification | ILSVRC                      | VGG(I-VG)          | 26.8%    |
|                       | PolyU palmprint             | VGG(P-VG)          | 93.5%    |
|                       | PolyU fingerprint           | VGG(F-VG)          | 89.6%    |
|                       | ILSVRC                      | Alex(I-VG)         | 19.3%    |
|                       | PolyU palmprint             | Alex (P-VG)        | 92.4%    |
|                       | PolyU fingerprint           | Alex (F-VG)        | 90.3%    |
| Identification        | ILSVRC                      | VGG(I-V1)          | 28.3%    |
|                       | PolyU palmprint             | VGG(P-V1)          | 95.4%    |
|                       | PolyU fingerprint           | VGG(F-V1)          | 86.3%    |
|                       | ILSVRC                      | Alex(I-V1)         | 31.5%    |
|                       | PolyU palmprint             | Alex (P-V1)        | 92.4%    |
|                       | PolyU fingerprint           | Alex (F-V1)        | 90.8%    |
| Face Recognition      | ILSVRC                      | VGG(I-LFR)         | 40.7%    |
|                       | LFW database                | VGG(L-LFR)         | 92.3%    |
|                       | ILSVRC                      | Alex (I-LFR)       | 48.3%    |
|                       | VGG database                | Alex (V-LFR)       | 93.6%    |

5. Conclusions

Impressive image recognition results have been obtained by adopting state-of-the-art DCNN, which is usually trained on a large image dataset, as a universal image feature descriptor. However, the results would degrade a lot when the pattern distribution between the original training dataset and the target one is greatly different, and there is not a standard index for evaluating the pattern similarity between two different datasets. To solve this problem, a quantized pattern similarity index is proposed by calculating the shared-atoms of common dictionary, and then a hierarchical transfer learning strategy guided by the similarity index evaluation is proposed to re-train task-specific feature extractor from the original DCNN model. Experiments with hand-vein based gender and identity recognition task fully demonstrate the effectiveness of the proposed model, and further experiment with face recognition also verifies the proposal.

What’s more, this is the first model for realizing gender recognition with hand-dorsa vein image, and also the first model with DCNN as feature extractor for realizing vein recognition. We also argue that the proposed quantized similarity index is also applicable for other image pattern similarity evaluation task.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61379143.

References

[1] Z. Liu, J. Yang, H. Liu, and J. Liu, “Learning from multiple sources via multiple domain relationship,” IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1941–1944, July 2016.
[2] Y. Cheng, X. Wang, and G. Cao, “Multi-source tri-training transfer learning,” IEICE Trans. Inf. & Syst., vol.E97-D, no.6, pp.1668–1672, June 2014.
[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp.1–9, 2015.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” Proc. 25th International Conference on Neural Information Processing Systems, pp.1097–1105, 2012.
[5] O.M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” Proc. British Machine Vision Conference, pp.1–12, 2015.
[6] T. Danisman, I.M. Bilasco, and C. Djeraba, “Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings,” Proc. 22nd International Conference on Pattern Recognition, pp.3144–3149, 2014.
[7] D. Zhang, W.-K. Kong, J. You, and M. Wong, “Online palmprint identification,” IEEE Trans. Pattern Anal. Mach. Intell., vol.25, no.9, pp.1041–1050, Sept. 2003.
[8] Q. Zhao, D. Zhang, L. Zhang, and N. Luo, “High resolution fragmentary fingerprint alignment using pore-valley descriptors,” Pattern Recogn, vol.43, no.3, pp.1050–1061, March 2010.
[9] L. Huo, X. Feng, C. Huo, and C. Pan, “Learning deep dictionary for hyperspectral image denoising,” IEICE Trans. Inf. & Syst., vol.E98-D, no.7, pp.1401–1404, July 2015.
[10] J.M. Bioucas-Dias and M.A.T. Figueiredo, “A new twist: Two-step iterative shrinkage/thresholding algorithm for image restoration,” IEEE Trans. Image Process., vol.54, no.11, pp.4311–4322, Dec. 2007.
[11] G. Ozbulak, Y. Aytar, and H.K. Ekenel, “How Transferable Are CNN-Based Features for Age and Gender Classification?” Proc. 2016 International Conference of the Biometrics Special Interest Group, pp.1–6, 2016.