A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 127

Multi-scale facial scanning via spatial LSTM for latent facial feature representation

Seong Tae Kim[1], Yeoreum Choi[1], Yong Man Ro[2]

Abstract: In the past few decades, automatic face recognition has been an important vision task. In this paper, we exploit the spatial relationships of facial local regions by using a novel deep network. In the proposed method, face is spatially scanned with spatial long short-term memory (LSTM) to encode the spatial correlation of facial regions. Moreover, with facial regions of various scales, the complementary information of the multi-scale facial features is encoded. Experimental results on public database showed that the proposed method outperformed the conventional methods by improving the face recognition accuracy under illumination variation.

Keywords: Face recognition, facial feature representation, spatial LSTM, deep learning

1 Introduction

In the past few decades, automatic face recognition has been an important vision task for many applications such as video surveillance and biometric identification [JRP04, CRP12, KKR16a]. For biometric identification, it is important to extract discriminative features which discriminate inter-person differences while being robust to intra-personal variations (e.g. illumination variations) [DCTD16].

As recent progress of deep learning, convolutional neural networks (CNN) have shown outstanding performance on many fields of computer vision such as image classification [KSH12, DCTD16], object detection [RHGS15], and action recognition [JXYY13]. Recently, the CNN has also been used to solve face recognition problems by learning latent and discriminative features [PVZ15, CKR16, KKR16b]. Generally, the CNN is comprised of one or more convolutional layers with a subsampling layer and followed by one or more fully-connected layers. In the convolutional layer, the filters slide over input images with convolutional operation to encode local image features. The neurons of feature maps obtained by convolution layer are connected to neurons of the fully-connected layer. In other words, spatial information extracted from local regions is simply aggregated to construct the image features. However, there are spatial relationships in facial local regions, which could not be encoded in the conventional CNN framework for face recognition [PVZ15, CKR16].

[1] School of Electrical Engineering, KAIST, Daejeon 34141, Republic of Korea, Email: {stkim4978, cyr0703}@kaist.ac.kr. Both authors are equally contributed to this manuscript.
[2] School of Electrical Engineering, KAIST, Daejeon 34141, Republic of Korea, Email: ymro@kaist.ac.kr. Corresponding author

In this paper, we propose a novel face recognition framework using deep network to solve the abovementioned limitation of the conventional CNN for face representation. To exploit the spatial relationships of facial local regions, we devise a long short-term memory (LSTM) network with which the whole face is scanned sequentially. The LSTM network originally introduced for sequence learning [HS97, GMH13, KKR16a]. It incorporates memory cells with three control gates (i.e., input, forget, output). The memory cells can store, modify, and access an internal state to learn long-term dependencies [BSF94]. In the proposed method, spatial long short-term memory has been devised to learn spatial dependencies of facial local features extracted from facial local regions. The contributions of this paper are summarized as followings: 1) A novel framework has been devised to encode latent facial features from spatial relationship of facial local regions. First, the facial local features are encoded by the CNN. Then the each facial local feature is used to construct facial latent spatial relationship-feature by scanning the whole face image. In other words, the face is scanned by the spatial LSTM network to learn relationship and dependencies of spatially sequential facial local regions. The memory cells of the spatial LSTM enable the proposed deep network to discover latent relationship of facial local regions. 2) The effectiveness of the proposed framework has been validated on the public face database. By the experiments, it is verified that the proposed method is robust to extract facial features under illumination variation. Moreover, the performance of face recognition could be further improved with multi-scale spatial long short-term memory, which combines latent facial features learned from multi-scale facial local regions.

The rest of this paper is organized as follows: The proposed latent facial feature representation using facial scanning is described in Section 2. Face recognition with multi-scale facial scanning is explained in Section 3. Section 4 presents and discusses experimental results. Finally, Section 5 provides concluding remarks.

2 Proposed latent facial feature representation by spatial LSTM

Figure 1 shows the overview of the proposed latent facial feature representation. The proposed method consists of facial local feature representation and spatial LSTM network. To learn the proposed latent facial feature representation, each face image is divided into regions horizontally and vertically, as shown Fig. 1. The objective of spatial LSTM is to learn relationship and dependencies of spatially sequential facial local regions. The spatial LSTM network consists of horizontal LSTM networks and vertical LSTM network. For horizontal scan, we divide face evenly into N_h parts with overlapping between two eye centers. The two eye centers are located based on facial landmark detection method [AZCP14]. For vertical scan, N_v horizontal patch sets are evenly divided between an eye corner and a lip corner. Eye corner and lip corner are also located by the facial landmark detection method. In this way, we can acquire $N_h \times N_v$ facial local patches from a face image (as shown in Fig. 1).

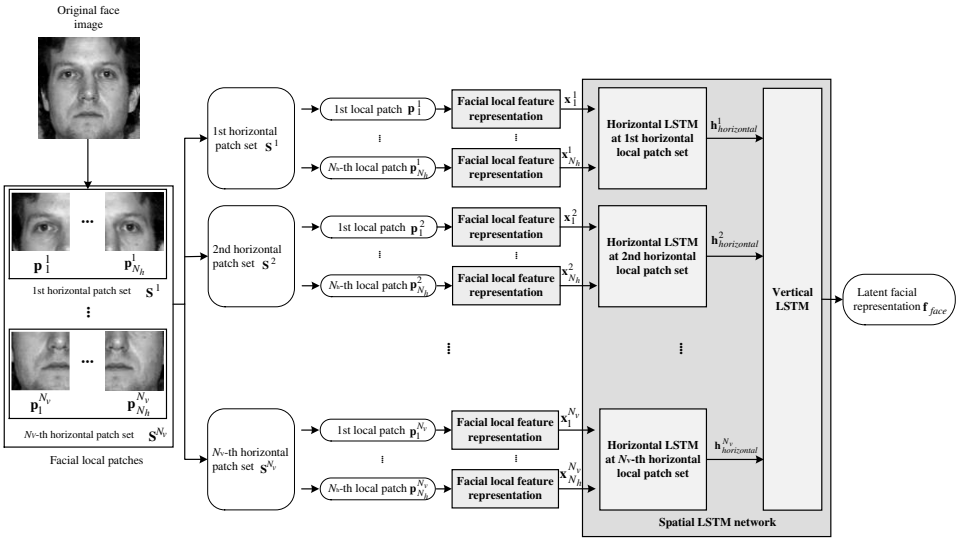

Fig. 1. Overall framework of the proposed latent facial feature representation. It consists of facial local feature representation and spatial LSTM.

The facial local features such as texture and shape are encoded by a CNN. The facial local features are used for input sequences of a spatial LSTM network. Let $\mathbf{F}^m = \left\{\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots \mathbf{x}_{N_h}^m\right\}$ denotes facial local features, which are extracted from the m-th horizontal patch set $\mathbf{S}^m = \left\{\mathbf{p}_1^m, \mathbf{p}_2^m, \ldots \mathbf{p}_{N_h}^m\right\}$ where $m = 1, 2, \ldots, N_v$. $\mathbf{p}_n^m$ denotes the n-th local patch in m-th horizontal patch set and $\mathbf{x}_n^m$ is the facial local feature encoded from $\mathbf{p}_n^m$.

We employ bidirectional LSTM to consider both directions in face scanning as:

$$\mathbf{h}_{fwd,n}^m = LSTM_{fwd}(\mathbf{x}_n^m, \mathbf{h}_{fwd,n-1}^m), \tag{1}$$

where $LSTM_{fwd}(\cdot)$ denotes a function which performs the operation of the LSTM layer in forward direction and $\mathbf{h}_{fwd,n}^m$ is the hidden state of the forward LSTM at n-th local patch in m-th horizontal patch set.

$$\mathbf{h}_{bwd,n}^m = LSTM_{bwd}(\mathbf{x}_n^m, \mathbf{h}_{bwd,n+1}^m), \tag{2}$$

where $LSTM_{bwd}(\cdot)$ denotes a function which performs the operation of the LSTM layer in backward direction and $\mathbf{h}^{m}_{bwd,n}$ is the hidden state of the backward LSTM at n-th local patch in m-th horizontal patch set. Then horizontal feature $\mathbf{h}^{m}_{horizontal}$ encoded at m-th horizontal patch set is represented as

$$\mathbf{h}^{m}_{horizontal} = [\mathbf{h}^{m}_{fwd,N_h}, \mathbf{h}^{m}_{bwd,1}]. \tag{3}$$

The vertical sequence acquired from horizontal LSTM network $\mathbf{h}_{horizontal} = \left\{ \mathbf{h}^{1}_{horizontal}, \mathbf{h}^{2}_{horizontal}, \ldots, \mathbf{h}^{N_v}_{horizontal} \right\}$ is used for the vertical LSTM to encode the facial feature vector $\mathbf{f}_{face}$ as followings:

$$\mathbf{f}_{fwd,m} = LSTM_{fwd}(\mathbf{h}^{m}_{horizontal}, \mathbf{f}_{fwd,m-1}), \tag{4}$$

$$\mathbf{f}_{bwd,m} = LSTM_{bwd}(\mathbf{h}^{m}_{horizontal}, \mathbf{f}_{bwd,m+1}), \tag{5}$$

$$\mathbf{f}_{face} = [\mathbf{f}_{fwd,N_v}, \mathbf{f}_{bwd,1}], \tag{6}$$

where $\mathbf{f}_{fwd,m}$ is the hidden state of the forward LSTM at m-th horizontal feature and $\mathbf{f}_{bwd,m}$ is the hidden state of the backward LSTM at m-th horizontal feature. Consequently, both horizontal LSTM networks and vertical LSTM network can learn gradual changes with respect to facial local distributions.

Fig. 2. Various scales of local patches at eye region for multi-scale facial scanning.

3 Face recognition with multi-scale facial scanning

From aforementioned spatial LSTM network which consists of horizontal LSTM networks and vertical LSTM network, we obtain a facial feature vector. By changing the

size of local region which is used to encode facial local features, the various facial feature vectors can be encoded in the spatial LSTM network. Therefore, combining these multi-scale facial features obtained from facial scanning, the complementary information could be encoded for face recognition. For this purpose, local patches are extracted with various sizes for considering multi-scale local features. In details, we acquire local patches with scale factor α which determines the ratio of the size of local region to the size of whole face image as shown in Fig. 2. Finally, the facial feature vectors extracted from various scales are combined as followings:

$$\mathbf{f}_{multiscale} = [\mathbf{f}_{face,1}, \cdots, \mathbf{f}_{face,\alpha}, \cdots, \mathbf{f}_{face,N_s}], \tag{7}$$

where $\mathbf{f}_{face,\alpha}$ denotes the facial feature vector obtained from facial scanning using local patch size of $\frac{1}{\alpha}$ and N_s denotes the number of multi-scale approach. Finally, a feature vector $\mathbf{f}_{multiscale}$ is used for face recognition. For the face recognition, 1-nearest neighborhood classifier is used based on Euclidean distance.

4 Experiments

4.1 Experimental conditions

To verify the proposed method, we performed experiments with the publicly available CMU Multi-PIE database which was collected from the face images under 20 illumination conditions (as seen in Fig. 3) [GMCKB10]. Particularly, the effectiveness of the proposed method under environment variation (i.e., varying illumination conditions) was investigated in this paper. We followed the experimental protocol in [CKR16] as followings. Among 337 subjects, we used mutually exclusive setting between the training set and the test set for evaluating the proposed method. The first 200 subjects were used for the training set and the remaining 137 subjects were used for the test set. In the case of test phase, the gallery images were set with only one frontal illumination condition and the probe images were chosen with other varying illumination conditions. In other words, the face images with 19 other illumination conditions of the database were included in the probe images. The number of gallery and probe images was 137 and 2,603, respectively.

In the experiment, N_h and N_v were set to 7 for cropping facial local regions. Each cropped facial local region was resized to 32×32 pixels. To extract feature vectors from facial local regions, the CNN structure [SKR15, CKR16] which consisted of three convolutional layers with a max-pooling layer, and two fully-connected layers was

adopted. For the case of LSTM network, each forward and backward LSTM layer has 512 memory cells, respectively. The proposed deep network was implemented by using the Keras framework with Theano backend [Ch15]. To avoid over-fitting, fully-connected layers and LSTM layers were constrained using drop out [SHKSS14].

Fig. 3. Example face images from CMU Multi-PIE under 20 different illumination conditions.

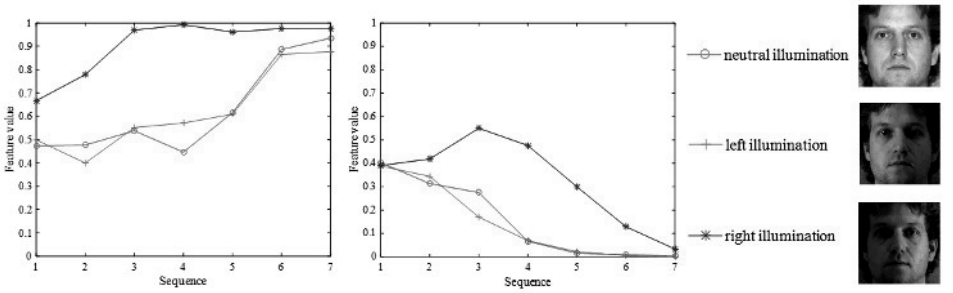

Fig. 4. Examples of feature changes according to sequential inputs of horizontal LSTM network. For visualization purpose, one of the feature value was selected from $\mathbf{h}^1_{horizontal}$ and normalized to [0, 1]. (Red (o): neutral illumination, green (+): left illumination, blue (*): right illumination)

4.2 Analysis of spatial LSTM for each illumination

Figure 4 shows the process of feature changes according to sequential inputs. The direction of input sequences was left to right. Each figure represents specific feature value of output feature vector from the LSTM network. One of the feature values obtained from the first horizontal LSTM network was used for the visualization. There were three face examples with different illumination conditions, which were neutral, left, and right illumination. As shown in each figure, the feature values of face images

showed similar changes under neutral and left illumination. On the contrary, the values of face images showed different tendency under right illumination. Nevertheless, all the values converged to feature values as bright parts of face images were put into LSTM network. These results indicated that the proposed method had the ability to store important information and forget noisy information, which resulted in encoding discriminative features under illumination variations.

Method	Accuracy
LBP [AHP06]	68.33%
GradientFace [ZTFS09]	84.75%
Weber-Face [WLYL11]	90.47%
VGG-Face [PVZ15]	85.06%
Two-step CNN [CKR16]	96.24%
Proposed method (α=2)	**96.73%**
Proposed method (multi-scale)	**98.08%**

Table 1. Accuracy of face recognition of the proposed method on CMU Multi-PIE database.

4.3 Face recognition performance under illumination variations

Table 1 shows the face recognition accuracy of the proposed method for CMU Multi-PIE database. For the comparison, local binary pattern (LBP) [AHP06], GradientFace [ZTFS09], Weber-face [WLYL11], VGG-face [PVZ15], and two-step CNN [CKR16] were used. The LBP was one of the popular approaches for local texture feature representation. The GradientFace and Weber-face were photometric normalization-based approaches for illumination variation. The VGG-face was CNN model learned from large scale celebrity face images. In this study, the pre-trained VGG-face model was fine tunned on the CMU MultiPIE database. The two-step CNN was the CNN-based approach which compensated illumination effects. As shown in the table, the proposed latent spatial facial feature representation achieved the accuracy of 96.73% at α=2. It outperformed other methods. This result indicated that encoding spatial sequential relationships between facial local regions was useful for face representation.

For the multi-scale facial scanning, N_S was set to 5. The proposed multi-scale approach achieved 98.08% accuracy by combining the multi-scale facial features obtained from various size of facial local regions. It was mainly attributed to the fact that the multi-scale approach could exploit the complementary information of multi-scale spatial long

short-term memory.

5 Conclusions

In this paper, we proposed the multi-scale facial scanning via spatial LSTM for latent facial feature representation. By scanning the face using the spatial LSTM network, the proposed method could exploit the relationship of the facial local regions. The experimental results with CMU Multi-PIE dataset showed that sequential relationships of facial local regions encoded by spatial LSTM network were useful in face recognition under illumination varioations. It was mainly attributed to the fact that important information was stored and noisy information was deleted by considering the spatial relationships of facial local regions in the spatial LSTM network. Moreover, by combining the complementary information obtained from multi-scale approaches, the accuracy of face recognition could be further improved in the proposed method.

6 Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No. 2015R1A2A2A01005724).

References

[AHP06] Ahonen, T.; Hadid, A.; Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12): 2037-2041, 2006.

[AZCP14] Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M.: Incremental face alignment in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1859-66, 2014

[BSF94] Bengio, Y.; Simard, P.; Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks, 5(2): 157-166, 1994.

[Ch15] Chollet, F: Keras: Theano-based deep learning library. 2015.

[CRP12] Choi, J. Y.; Ro, Y. M.; Plataniotis, K. N.: Color local texture features for color face recognition. IEEE Transactions on Image Processing, 21(3): 1366-1380, 2012.

[CKR16] Choi, Y.; Kim. H.-I.; Ro, Y. M.: Two-step Learning of Deep Convolutional Neural Network for Discriminative Face Recognition under Varying Illumination. In: Electronic Imaging. 2016.

[DCTD16] Ding, C.; Choi, J.; Tao, D.; Davis, L.: Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3): 518-5531, 2016.

[GMH13] Graves, A.; Mohamed, A.-r.; Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6645-6649. 2013.

[GMCKB10] Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S.: Multi-pie. Image and Vision Computing, 28(5): 807-813, 2010.

[HS97] Hochreiter, S.; Schmidhuber, J.: Long short-term memory. Neural Computation, 9(8): 1735-1780, 1997.

[JRP04] Jain, A. K.; Ross, A.; Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(1): 4-20, 2004.

[JXYY13] Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231, 2013.

[KKR16a] Kim, S. T.; Kim, D. H.; Ro, Y. M.: Facial dynamic modelling using long short-term memory network: analysis and application to face authentication. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems. 2016.

[KKR16b] Kim, S. T.; Kim, D. H.; Ro, Y. M.: Spatio-temporal representation for face authentication by using multi-task learning with human attributes. In: IEEE International Conference on Image Processing. pp 2996-3000. 2016.

[KSH12] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp 1097-1105. 2012.

[MH08] Maaten, L.V.D.; Hinton, G: Visualizing data using t-SNE. In: ournal of Machine Learning Research. pp. 2579-2605. 2008.

[PVZ15] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. 2015.

[RHGS15] Ren, S.; He, K.; Girshick, R.; Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp 91-99. 2015.

[SHKSS14] Srivastava, N; Hinton, G; Krizhevsky, A; Sutskever, I; Salakhutdinov, R: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1): 1929-1958. 2014.

[SKR15] Seo, J.-J.; Kim, H.-I.; Ro, Y. M.: Pose-robust and discriminative feature representation by multi-task deep learning for multi-view face recognition. In: IEEE International Symposium on Multimedia. pp 166-171. 2015.

[SZ14] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. 2014.

[WLYL11] Wang, B.; Li, W.; Yang, W.; Liao, Q.: Illumination normalization based on weber's law with application to face recognition. IEEE Signal Processing Letters, 18(8): 462-465, 2011.

[ZTFS09] Zhang, T.; Tang, Y. Y.; Fang, B.; Shang, Z.; Liu, X.: Face recognition under varying illumination using gradientfaces. IEEE Transactions on Image Processing, 18(11): 2599-2606, 2009.