A Novel Convolutional Neural Network for Facial Expression Recognition

Jing Li[1] , Yang Mi[1], Jiahui Yu[2], Zhaojie Ju[3,4]

[1] School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]School of Information Science & Engineering, Shenyang Ligong University, Shenyang, 110159
[3] School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, U.K.
[4] The State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China

jingli@ncu.edu.cn; 997562713@qq.com; zhaojie.ju@port.ac.uk

Abstract. Facial expression recognition is becoming a hot topic due to its wide applications in computer vision research fields. Traditional methods adopt hand-crafted features combined with classifiers to achieve the recognition goal. However, the accuracy of these methods often relies heavily on the extracted features and the classifier's parameters, and thus cannot get good result with unseen data. Recently, deep learning, which simulates the mechanism of human brain to interpret data, has shown remarkable results in visual object recognition. In this paper, we present a novel convolutional neural network which consists of local binary patterns and improved Inception-ResNet layers for automatic facial expression recognition. We apply the proposed method to three expression datasets, i.e., the Extended Cohn-kanade Dataset (CK+), the Japanese Female Expression Database (JAFFE), and the FER2013 Dataset. The experimental results demonstrate the feasibility and effectiveness of our proposed network.

Keywords: Facial expression recognition, Deep learning, LBP, Inception-ResNet layers

1 Introduction

In human daily communications, audio and visual signals are mixed to understand each other. Audio signal is the most direct way to express ourselves and visual signals help us get potential information. As a part of visual signals, facial expressions, which refer to the movements of the mimetic musculature of the face [25], provide rich information about one's real emotions and intentions. A study [23] reported that facial expressions constitute 55% of the effect of a communication message which is higher than language and voice do. Due to the important role in conveying information, automatic facial expression analysis is becoming an interesting and challenging topic in today's computer vision research area, including human-computer interaction, data-driven animation, and so on. Though much progress has made in this study, with the

variability, complexity and subtlety of facial expressions, it is still difficult to achieve a desirable recognition accuracy in practical applications.

Ekman and Friesen [26] provided six basic facial expression forms, i.e., anger, happiness, disgust, sadness, surprise and fear, which are widely accepted, and most of the current articles are aiming at automatically identifying these six prototypic expressions. Early in 1978, Suwa et.al [28] presented a preliminary investigation on automatic facial expression analysis. Picard [24] showed that affective computing could be a useful way to expand human-computing communication. Ekman and Friesen [27] developed the famous Facial Action Coding System (FACS) to detect subtle variations in facial expressions. Since then, extracting facial features and classifying different facial expressions based on computer technologies have attracted more and more attention. However, these works did not significantly progress until the last decade, Hinton et al. [29] proposed that with many hidden layers in deep learning neural network, the ability of learning characteristics could be improved, as well as the model prediction and classification accuracy.

Our aim is to explore and design a system that could perform automated facial expression analysis. Generally, three main steps are involved in tackling the problem: 1) facial detection and preprocessing; 2) facial feature extraction; and 3) facial expression classification. As we can see in Fig.1. In order to accomplish the classification tasks, traditional approaches such as Support Vector Machine (SVM) perform well on classifying posed facial expressions in a controlled environment and lab settings. However, in a spontaneous uncontrolled circumstance, these methods cannot work effectively. In 1980s, LeCun et al. [30] proposed the Convolutional Neural Network (CNN), which contains convolutional layers (C layers) and subsampling layers (S layers) as two basic layers, made the classification more robust.

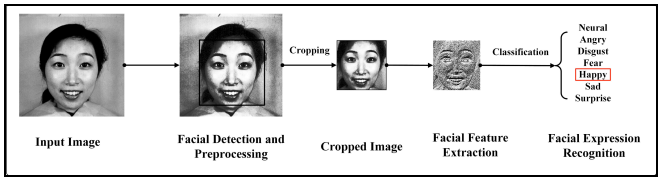

Fig. 1. An example of a facial expression recognition system.

In this paper, we propose a novel deep convolutional neural network for recognizing facial expressions. Inspired by [7], improved Inception-ResNet is used since it would allow Inception to reap the benefits of the residual approach while maintaining its computational efficiency. Furthermore, Local Binary Patterns (LBPs), which contains image texture information, is entered into the network. We conduct our proposed method on three well-known facial expression datasets, which are CK+ [14], JAFFE [16], and FER2013 [33].

The main contributions of this paper are as follows:

1) We introduce an improved double-input feature extraction block to recognize facial expressions automatically. Except for the raw images, we add LBPs to the two middle-layers of the network, because they contain texture

information of the face and can reflect subtle face changes. In this way, the performance can be significantly improved.

2) Traditional datasets (CK+ and JAFFE) only containing frontal facial images were collected in specific environments. Some expressions are even unnatural and exaggerated which are not suitable for practical good social communications. In order to make the network generalize better for communication in daily life, we use the FER2013 dataset that contains more variations in pose, occlusion, and illumination

3) We employ data augmentation and batch normalization to avoid overfitting. For data augmentation, we randomly rotate, flip, and shift the images to expand the size of datasets. Batch normalization can reduce intra-class variation while does not distort the images. Section 3 describes this procedure in detail.

Fig. 2. Some examples of the CK+ dataset (the first row), the JAFFE dataset (the second row), and the FER2013 dataset (the third row). The CK+ dataset and the JAFFE dataset only contain frontal facial images, which were collected in conditional environments. The FER2013 dataset includes more variations in pose, occlusion, and illumination.

The rest of this paper is arranged as follows. Section 2 outlines the related work in the field of facial expression recognition. Section 3 describes the newly proposed network. Section 4 presents the experimental results and Section 5 concludes the whole paper

2 Related work

Facial expressions represent one's most powerful and immediate means for emotions and intentions communication [15]. Because of its potential applications, facial expression recognition can be used in many fields, such as human-robot interaction [16], surveillance [17], health-care [18], and so on.

The facial action coding system developed by Ekman [27] is the inspiration for many research papers. Luecy et al. [14] created the Extended Cohn-kanade dataset (CK+). In the classification stage, the dataset was evaluated using Active Appearance Models (AAMs) in combination with SVMs. The architecture achieved the final accuracy over 65% for each expression. In [21], facial expressions were classified into six basic emotions. First, the active patches, which are the most useful parts when people give an expression, were extracted from the facial region. Then, LBP used as

the feature descriptor were fed into Support Vector Machines (SVMs) for classification. The system achieved 94.39% and 92.22% for extended Cohn-Kanade (CK+) dataset and JAFFE dataset respectively.

Recently, it has been a popular way to use deep neural network to recognizing facial expression and visual objects. In [19], the earliest convolutional neural network LeNet-5 was presented for handwriting recognition. Then, many variants of this basic design are prevalent in image classification. In 2012, Hinton proposed the AlexNet [20] as the beginning of the larger and deeper convolutional neural networks. It classifies 1.2 million high-resolution images into 1,000 different classes.

Szegedy et al. [9] proposed GoogLeNet, which increases both the width and depth of the network to improve the architecture performance. This is a 22-layer deep network and the main structure is "Inception" layers, which allow the architecture to make a more complex decision. In 2015, He et al. [5] produced the deep residual learning framework called ResNet to address the degradation problem. The residual block uses shortcut connection to add outputs to the outputs of the stacked layers. When the model layer to deepen, the simple operation can solve the degradation problem.

The previous studies [3, 4] had shown remarkable improvement in recognition rates by using Inception, and also ResNet had achieved remarkable results in very deep networks [5, 6]. Inspired by these advantages, the Inception architectures were combined with residual connections in [7], which was called as Inception-ResNet layers. Wherein, it also presented clear empirical evidence that residual connections accelerate the training stage of Inception networks, meanwhile maintain the computation effectiveness.

3 A Double Input Architecture

In this paper, we utilize a double-input Inception-ResNet block (DIB) to address the facial expression recognition problem. In the proposed neural network, we incorporate LBP features during the training step. These features provide texture information and thus reflect slight changes on the face, which helps the network pay attention to the facial feature so as to improve the recognition accuracy. Beyond that, we employ batch normalization to guarantee the stable distribution of the input data at each layer. Last but not the least, we use data augmentation to prevent overfitting. We will explain each module in detail in the following sub-sections.

3.1 A Double Input Inception-ResNet Block

We propose a modified Inception-ResNet architecture, which has double-input feature blocks to automatically recognize seven expressions. As inspired by the Inception-ResNet architecture [7], we choose Inception-v4 layers, which achieve better recognition rates compared with the other models. Fig.3 shows the overall schema for the modified Inception-ResNet network. We feed the input images with the size of 299*299*1 into the "stem" module and add LBP features into the Inception-ResNet-A

module and the Inception-ResNet-B module. The LBP feature maps are resized to their corresponding filter size in the network. Batch normalization layer is added to all of the convolution layers to avoid overfitting and ReLU [10] activation function is added to all of the batch normalization layers to avoid the vanishing gradient problem.

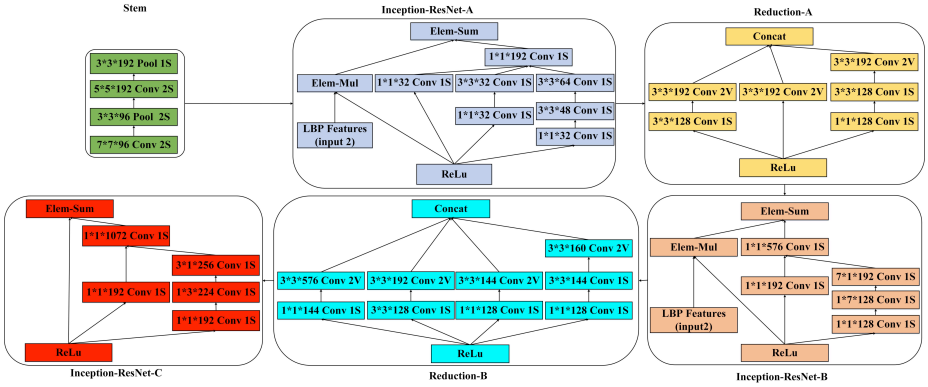

Fig. 3. The overview of the architecture.

3.2 Local Binary Pattern

Robust facial features should minimize within-class variations of expressions while maximize between-class variations. There are two types of facial feature extraction methods: Geometric feature-based methods and appearance-based methods. The geometric feature-based methods aim to use the geometric relationships between facial feature points to extract facial features, but commonly require accurate and reliable facial feature detection and tracking, which is difficult to accommodate in many situations. The appearance-based methods reflect the underlying information in a face image that can help determine the expression attributes accurately. However, this method may ignore overall information of the face to some extent. Although either of these two methods can effectively represent facial expressions, we employ appearance-based methods since it can pay attention to the changes in skin texture that is important for facial expression modelling. As a representative of the appearance-based methods, we extract local binary patterns (LBPs) [11] as the facial features. The LBPs can reflect subtle changes of the face such as wrinkles and furrows, which makes it convenient for a more detail study. The LBP operator is rotation invariant and grey invariant, and thus can solve the problem of imbalanced displacement, rotation and illumination in an image.

There are several types of LBPs. The original LBP operator introduced in [11] was defined in a 3*3 neighborhood. It thresholds each pixel with the center value, and then considers the result as a binary number. The LBP feature at this pixel can be expressed as follows:

$$LBP(x_c, y_c) = \sum_{p=0}^{7} s(i_p - i_c) 2^p \quad . \tag{1}$$

where p is the number of different pixels, (x_c, y_c) as the center pixel with intensity i_c, and i_p being the intensity of the neighborhood pixel, and s is the sign function defined as:

$$s(x)=\begin{cases}1 & if \quad x\geq 0\\ 0 & if \quad x<0\end{cases} . \tag{2}$$

In order to make it suitable for different scales of texture features and achieve gray-scale invariance, the operator extends to use different-size neighborhoods [12]. The local neighborhoods are defined as a set of sampling points evenly spaced on a circle, which are centered at the pixel to be labeled. It allows any radius and number of sampling points. Bilinearly interpolating is used when sampling points do not fall in the center of a pixel. Given a pixel at (x_c, y_c), the resulting LBP can be expressed as：

$$LBP_{P,R}(x_c, y_c)=\sum_{P=0}^{P-1} s(i_p-i_c)2^P \quad . \tag{3}$$

where U is the uniformity measure and the superscript “rius2” denotes rotation-invariant uniform pattern.

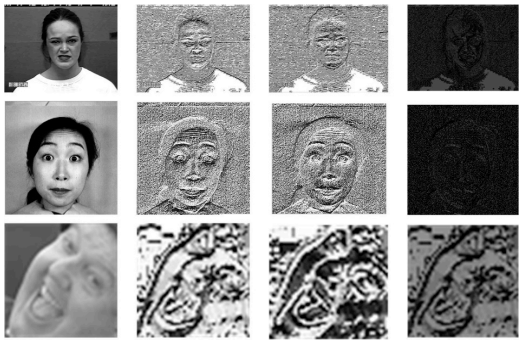

Fig. 4. Each row shows the LBP images of the CK+ dataset, the JAFFE dataset, and FER2013 dataset, respectively. From left to right is the input image, original LBP image, circle LBP image, and uniform LBP image.

3.3 Batch Normalization

In previous tasks, such as face detection [25] and face identification [34], one class only contains one person. While in facial expression recognition, the same expression class may contain numerous individuals, which leads to intra-class variations. In order to solve this problem, traditional methods always apply face alignment. However, it may distort the images. Current deep learning networks always apply batch normalization [1], which makes the network focus on images that are more real by distorting them less. Batch normalization adds standardized processing to the input data of each layer during the training process in deep neural networks. Thus, it ensures that each layer of the input data distribution is stable. Batch normalization

also alleviates the covariate shift phenomenon [2] caused by the input distribution changes. It allows the networks to decrease the need for dropout and local response normalization. In our network, we add batch normalization layers after all of the convolutional layers.

3.4 Data Augmentation

It is important to consider overfitting in our deep convolutional neural network. Overfitting makes the model perform too well on the training data, but leads to poor performance in validation data and test data. Data augmentation and dropout are two primary ways to prevent overfitting. For data augmentation, the easiest and most common way is to artificially enlarge the dataset using label-preserving transformation. In this paper, considering the images in the three datasets are grayscale, we do not manipulate the contrast, brightness or color. First, we take a random rotation on the images to capture different-angle invariance. Second, we flip the images horizontally to capture the reflection invariance. Finally, we randomly shift the images to capture the translation invariance.

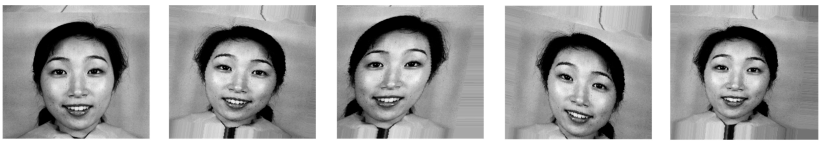

Fig. 5. An example for data augmentation: (a) is the original facial image, (b), (c), (d), (e) are the images after random data augmentation

4 Experiment Result

4.1 Face Datasets

In this work, we design a double-input Inception-ResNet block to automatic classifying static images. Except for the famous facial expression datasets, such Cohn-Kandade [14] and JAFFE [16], we also evaluate our proposed method on FER2013 [50] that is more challenging with large variations in pose, illumination and occlusion.

CK+: The Extended Cohn-Kanade database [14] contains 593 video sequences recorded from 123 subjects. Participants were from 18 to 50 years old. There are seven basic expressions, which are anger, disgust, fear, happy, sad, and surprise. Each sequence began in a neural face and ended in a peak expression. All the images are

640*490 or 640*480 pixel arrays with 8-bit gray-scale or 24-bit color values. In this paper, we select 5,287 images as the dataset.

JAFFE: The Japanese Female Facial Expression (JAFFE) database [16] contains 213 images of seven facial expressions (neutral, angry, disgust, fear, happy, sad and surprise) from 10 Japanese female models, each of which has 2-4 samples for one expression. Each image is 256*256.

FER2013: The Facial Expression Recognition Challenge 2013 (FER2013) [33] is a dataset for spontaneous facial expression. It contains 28,709 training images and 3,589 testing images of seven expressions (six basic facial expressions and neutral), which are 48*48. Unlike CK+ and JAFFE, FER2013 contains more variations in pose, illumination and occlusion.

4.2 Results

We evaluate our architecture on three datasets mentioned above, i.e. the CK+ dataset, the JAFFE dataset, and the FER2013 dataset. The model is trained with an initialized learning rate 0.001, and the batch size is around 64~128. We add batch normalization after each convolutional layer. The model obtains the best classification accuracy after 300 epochs. Our network is trained on one GTX1080 TI GPU.

We compare the performance of our proposed method with the state-of-the-art facial expression recognition methods on the three dataset. The methods include Inception-ResNet with Conditional Random Field (Inception-ResNet+CRF) [8], Salient Facial Patches (SFP) [21], VGGNet and Long Short Term Memory (VGGNet+LSTM) [22], Face Parsing and Stacked Autoencoder (FP+SAE) [31], Inception layer and Network in Network theory (Inception+NIN) [3], and Linear Support Vector Machine (LSVM) [32]. In order to demonstrate the merit of incorporating LBP features, we also provide the result of our network without LBP features, which are replaced with a simple shortcut between the input and output of the residual unit in the Inception-ResNet A and Inception-ResNet B modules. We call this Single-Input Inception-ResNet Block (SIB).

Table 1. Recognition rates (%) on the CK+ dataset, the JAFFE dataset and FER2013 dataset.

Methods \ Datasets	CK+	JAFFE	FER2013
DIB(ours)	**97.34**	**98.92**	**67.71**
SIB(ours)	94.38	97.78	58.21
Inception-ResNet+CRF [8]	93.04	-	-
SFP [21]	94.39	92.22	-
VGGNet+LSTM [22]	97.2	-	-
FP+SAE [31]	-	90.95	-
LSVM [32]	-	-	69.3
Inception+NIN [3]	-	-	66.4

5 Conclusion and Future Work

This paper presents a double-input Inception-ResNet block for facial expression recognition in static images. The network explores the texture information in order to improve the architecture performance. Considering the depth of the network, we first increase the number of images to prevent overfitting, then put the facial images into the network and classify them into either of the six basic expressions or the neutral one. The proposed method is evaluated on three well-known datasets: the CK+ dataset, the JAFFE dataset, and the FER 2013 dataset. Our experimental results show that the proposed method is more superior to many state-of-art methods when performing in FER2013.

Acknowledgments. This work is supported by National Natural Science Foundation of China under Grant 61463032 and 61703198, Natural Science Foundation for Distinguished Young Scholars of Jiangxi Province under Grant 2018ACB21014, Open Fund of State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences under Grant 20180109.

References

1. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint arXiv:1502.03167 (2015)
2. Shimodaira, H.: Improving Predictive Inference under Covariate Shift by Weighting the Log Likelihood Function. J. Stat. Planning Infer. **90**(2), 227–244 (2000)
3. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going Deeper in Facial Expression Recognition using Deep Neural Networks. IEEE Winter Conference on Applications of Computer Vision. IEEE, 1-10 (2016)
4. Szegedy, C., Liu, W., Jia, Y., et al.: Going Deeper with Convolutions. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 1-9 (2015)
5. He, K., Zhang, X., Ren, S., et al.: Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 770-778 (2016)
6. He, K., Zhang, X., Ren, S., et al.: Identity Mappings in Deep Residual Networks. European Conference on Computer Vision. pp. 630—645. Springer, Cham, (2016)
7. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261 (2016)
8. Hasani, B., Mahoor, M.H.: Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields. IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 790-795 (2017)
9. Hasani, B., Mahoor, M.H.: Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. Computer Vision and Pattern Recognition Workshops, 2278—2288 (2017)

10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. International Conference on Neural Information Processing Systems. Curran Associates Inc, 1097-1105 (2012)
11. Ojala, T., Harwood, I.: A Comparative Study of Texture Measures with Classification Based on Feature Distributions. Pattern Recognition. 29(1), 51-59 (1996)
12. Ali, A., Hussain, S., Haroon, F., et al.: Face Recognition with Local Binary Patterns. Bahria University Journal of Information & Communication Technologies. 5(1), 46-50 (2014)
13. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution Grayscale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE PAMI. 24(7),971-987 (2002)
14. Lucey, P., Cohn, J.F., Kanade, Y., Saragih, J., Ambadar, Z., Matthews. I.: The Extended Cohn-kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In: 2010 IEEE Computer Society Conference on, pp. 94–101 (2010)
15. Tian, Y., Brown, L., Hampapur, A., Pankanti, S., Senior, A., Bolle, R.: Real World Real-Time Automatic Recognition of Facial Expression. Proceedings of IEEE Workshop on, (2003)
16. Lyons, M., Akamatsu, S., Kamachi, M., et al.: Coding Facial Expressions with Gabor Wavelets. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp.200-205 (1998)
17. Song, Z., Ni, B., Guo, D., Sim, T., Yan, S.: Learning Universal Multi-View Age Estimator using Video Context. In: Proceedings of IEEE International Conference Comput. Vis., pp. 241–248 (2011)
18. Lucey, P., Cohn, J., Lucey, S., Matthews, I., Sridharan, S., Prkachin, K.: Automatically Detecting Pain using Facial Actions. In: IEEE International Conference on Affective Computing and Intelligent Interaction, pp.1–8 (2009)
19. Lecun, Y., Bottou, L., Bengio, Y.: Gradient-based Learning Applied to Document Recognition. Proceedings of the IEEE. 86(11), 2278-2324 (1998)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Communications of the Acm. 60(2), 1097-1105 (2012)
21. Happy, S.L., Routray, A.: Automatic Facial Expression Recognition Using Features of Salient Facial Patches. IEEE Transactions on Affective Computing. 6(1), 1-12 (2015)
22. Rodriguez, P., Cucurull, G., Gonzàlez, J.: Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. IEEE Transactions on Cybernetics, pp. 1-11 (2017)
23. Mehrabian, A.: Communication without words. Psychol. Today. 2(4), 53–56 (1968)
24. Picard R. W. Affective computing. MIT Press. 1(1), 71-73 (1997)
25. Viola, P., Jones. M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision. 57(2), 137-154 (2004)
26. Ekman, P., Friesen, W.V.: Constants across Cultures in the Face and Emotion. Personality Social Psychol. 17(2), 124–129 (1971)
27. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto. (1978)
28. Suwa, M., Sugie, N., Fujimora, K.: A Preliminary Note on Pattern Recognition of Human Emotional Expression. Proceedings of the Fourth International Joint Conference on Pattern Recognition, Kyoto, Japan, pp. 408–410 (1978)
29. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science, pp. 504-507 (2006)
30. Lecun, Y.: Generalization and Network Design Strategies. Connectionism in Perspective. (1989)
31. Lv, Y., Feng, Z., Xu, C.: Facial Expression Recognition via Deep Learning. International Conference on Smart Computing. IEEE, pp. 347-355 (2015)
32. Tang, Y.: Deep Learning Using Linear Support Vector Machines. Eprint Arxiv. (2013)

33. The FER2013 dataset, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
34. Zheng, L., Shen, L., Tian, L.: Person Re-identification Meets Image Search. Computer Science, pp. 1-1 (2015)