A RELATION NETWORK EMBEDDED WITH PRIOR FEATURES FOR FEW-SHOT CARICATURE RECOGNITION

Wenbo Zheng[1,2], *Lan Yan*[2,3], *Chao Gou* [2,4], *Wenwen Zhang*[1,2],*Fei-Yue Wang* [2,4]

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[4] Qingdao Academy of Intelligent Industries, Qingdao 266000, China
{*zwb2017@stu.xjtu.edu.cn;yanlan2017@ia.ac.cn;chao.gou@ia.ac.cn;zhangwenwen@stu.xjtu.edu.cn;feiyue.wang@ia.ac.cn*}

ABSTRACT

Caricature is a simple and abstract description of a person using her/his exaggerated characteristics. Due to amplified facial variations in the caricatures and significant differences among caricature and real face modalities, building vision models for recognizing each other between these modalities is an extremely challenging task. In addition, it is not easy to collect abundant samples of real faces and corresponding caricatures for training vision models, which makes the recognition more difficult. In this paper, we propose a novel relation network via meta learning to address the problem of few-shot caricature face recognition. In particular, we present a deep relation network to capture and memorize the relation among different samples. To employ the prior knowledge, we combine learned deep and handcrafted features to form the hybrid-prior representation via joint meta learning. Final recognition is derived from our relation network by learning to compare between the hybrid-prior features of samples. Experimental results on three caricature datasets of WebCaricature, IIIT-CFW, and Caricature-207 demonstrate that our method performs better than many existing ones for few-shot caricature recognition.

Index Terms— Caricature Recognition, Meta-Learning, Few-Shot Learning, Prior Feature.

1. INTRODUCTION

Caricatures are drawings with extreme distortion of a person's facial features. Caricatures recognition is not an easy task for computers. There are two main reasons: On one hand, a caricature cannot preserve facial structure and features to a large extent. On the other hand, a caricature may have variations with respect to the expressions, point of view, appearance, and also the underlying artistic style. Studying caricature recognition may help machine/computer to understand how human beings recognize faces well, and it may help the effective algorithms for face recognition in the real-world to improve their accuracy.

While many methods such as feature-based method, holistic methods, hybrid methods combining holistic and feature-based methods, geometry-based methods, and deep-learning-based methods have been successful at face recognition in the wild, the same cannot be said with respect to caricatures. In addition, there are only a few caricatures or one caricature of each person available in exacting caricatures dataset. Further, there are similar scenarios exist in real-world. For example, a police force may have only one single caricature or exaggerated sketch of a suspect when searching the massive videos. Due to the limitation of training samples in this kind of scenarios and exacting dataset, it is important to study the few-shot problem for caricatures recognition.

Caricature recognition belongs to a face recognition paradigm known as heterogeneous face recognition. Heterogeneous face recognition is the task of matching two faces from alternate modalities. While traditional face recognition has been explored to a large extent, there has been limited work on caricature recognition. In the work of caricature recognition, Abaci et al. [1] proposed a method to extract facial attribute features for photos, and manually labeled attribute features for caricatures recognition. Huo et al. [2] proposes an extension of the facial landmark based feature extraction scheme, and a variation robust cross-modal metric learning method for caricature recognition. However, in few-shot learning, this two methods are able to deal with the low degree of distortion, they are not effective to tackle with a lot of more curving distortion.

In general, there are three main challenges for caricature recognition:

1) Faces in photos and caricatures are represented heterogeneously. Photos are objective representations of subjects' faces captured directly from cameras, while caricatures are

This work is supported by National Natural Science Foundation of China (61806198, 61533019, U1811463). *(Corresponding author: Chao Gou).*

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00261

artists' subjective impressions of subjects.

2) Even facial appearance exaggerations are identified and artistic styles are removed, there are still other variations (e.g., point of view, appearance, and also the underlying artistic style) that can influence the recognition.

3) The number of photos and caricatures in the existing dataset limits the generalization of learning models. In addition, only photos or caricatures in one domain are available during training.

In contrast, even though there is a lot of exaggeration and distortion in caricatures beyond realism, humans are very good at recognizing the subjects. Why can human beings recognize caricatures quickly and accurately with very little direct supervision, or none at all? Probably because human beings can use the our past experience to learn, and the network can't. And isn't this one of the mechanisms of meta-learning [3]? *So why don't we use the principle of meta-learning to build a network to capture the relation of photos and caricatures?*

Because of the different cognitive mechanisms between the network and human beings, the feature extraction method of the network and human beings' handcrafted model are different for caricatures. The network requires a large amount of data to get a more reasonable feature representation, while the handcrafted model does not. Therefore, it is necessary to introduce handcrafted models into network model to enhance the network model or form hybrid-prior features. *Why not using hybrid-prior features during training network model for caricature recognition?*

Therefore, in this paper, we propose a novel few-shot learning approach via meta-learning to address the problem of few-shot caricature recognition. We propose an effective feature fusion approach to combine the extracted feature using network and the handcrafted feature for getting the better results of feature-based representative learning. We build the two-branch relation network via meta-learning. First, we use the embedding approach to do feature extraction of the training images. In this process, we introduce our proposed feature fusion approach to the learning of our network. Then, to compare the features, we design relation model that determines if they are from matching categories or not. We conduct experiments on the WebCaricature dataset [2, 4], IIIT-CFW dataset [5] and Caricature-207 dataset [6]. Experimental results show that the proposed algorithm performs better than similar works.

In short, the main contributions of this work are in three-fold.

1) We present a relation network with meta-learning to compare between the features of different samples. It allows to capture and memorize the relation between different caricatures.

2) In order to learn inductive prior knowledge for relation network, we further propose to embed handcrafted features into the relation network to enhance the recognition performance.

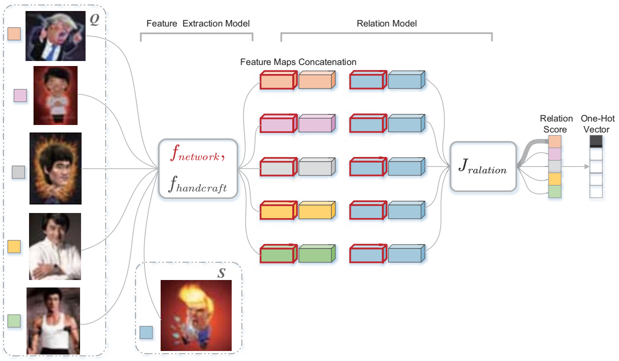


Fig. 1. The Pipeline of The Proposed Relation Network

3) Our network is effective, and experimental results show this method leads to better recognition performance than other start-of-the-art algorithms.

2. PRIOR FEATURES EMBEDDING RELATION NETWORK

2.1. Problem Setup

We consider the problem of caricature recognition as few-shot classifier learning. There are three datasets: a training set, a support set and a testing set. Note that the support set and testing set share the same label space, but the training set has its own label space. In other words, the training set is disjoint with support set and testing set.

We can in principle train a classifier to assign a class label $\hat{y}$ to each sample $\hat{x}$ in the test set while we only use the support set. However, in most cases, the performance of such a classifier is usually not good, because of the lack of the labeled samples in support set. Therefore, we use the meta-learning on the training set to transfer the extracted knowledge to on the support set. It aims to perform the few-shot learning on support set better and classify the test set more successfully.

We propose a novel matching networks [3] to solve the problem of caricature recognition. Suppose there are m labeled samples for each of n unique classes in support set. We select randomly n classes from the training set with m labeled samples from each of the n classes to conduct the sample set $S = \{(x_i, y_i)\}^z{}_{i=1} (z = m \times n)$, and we select the remaining samples to conduct the query set $Q = \{(x_j, y_j)\}^v{}_{j=1}$. This split strategy of sample and query set aims to simulate the support and test set that will be encountered at test time.

2.2. Network Formulation

Few-Shot Learning: As illustrated in Figure 1, our matching network consists of two branches: a feature extraction model and a relation model. Suppose sample x_j in the query

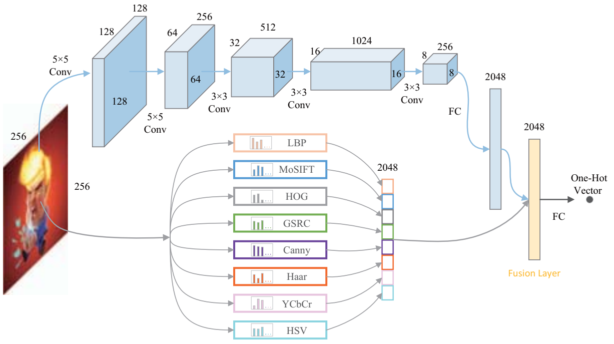


Fig. 2. The Structure of The Proposed Relation Network

set Q and sample x_i in the sample set S, we define the function $f_{network}$ which represents feature extraction function using network to produce feature maps $f_{network}(x_j)$ and $f_{network}(x_i)$, and define the function $f_{handcraft}$ which represents feature extraction function using handcrafted feature descriptors to produce handcrafted feature maps $f_{handcraft}(x_j)$ and $f_{handcraft}(x_i)$. The feature maps are combined using the function $C_{network}$, and the handcrafted feature maps are combined using the function $C_{handcraft}$. The feature maps and the handcrafted feature maps are combined using the function $C_{n\&h}$. In this work, we assume the $C_{handcraft}(\cdot,\cdot)$, $C_{network}(\cdot,\cdot)$, $C_{n\&h}(\cdot,\cdot)$ to be concatenation of corresponding feature maps in depth.

The combined feature map of the sample and query is used as the relation model $J_{relation}(\cdot)$ to get a scalar in range of 0 to 1 representing the similarity between x_i and x_j, which is called relation score. Suppose we have one labeled sample for each of n unique classes, our model can generate n relation scores $Judge_{i,j}$ for the relation between one query input x_j and training sample set examples x_i:

$$\begin{aligned} Judge_{i,j} = J_{relation}(C_{n\&h}(C_{network}(f_{network}(x_i), f_{network}(x_j)),\\ C_{handcraft}(f_{handcraft}(x_i), f_{handcraft}(x_j))))\\ i = 1, 2, \cdots, n \end{aligned} \tag{1}$$

Furthermore, for m labeled samples for each of n unique classes, we can element-wise sum over our feature extraction model outputs of all samples from each training class to form this class's feature map. And this pooled class-level feature map is combined with the query image feature map as above.

Objective Function: We use mean square error (MSE) loss to train our model, regressing the relation score $Judge_{i,j}$ to the ground truth: matched pairs have similarity 1 and the mismatched pair have similarity 0.

$$Loss = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{m} (Judge_{i,j} - (y_i == y_j))^2 \tag{2}$$

2.3. Network Architecture

Our network architecture is shown in Figure 2. Our network consists of two parts. The first part deals with traditional convolution, pooling and activation neurons for input images; the second part processes additional handcrafted feature representations of the same image. These two sub-networks are finally linked together to produce a full-fledged image description, so the second part will regularize the first part during learning. Finally, our hybrid-prior feature is extracted from the last fusion layer.

Network Features: The upper part of Figure 2 describes a traditional process of convolution and pooling. We use the 6-layer network architecture. Taking an image as input, the output of the 6-th pooling layer is a 2048-dimensional vector, which we regard as network features. The kernels of network change in turns: $3\times256\times256 \rightarrow 128\times128\times128$ (Convolution, kernel size: 1×1) $\rightarrow 256 \times 64 \times 64$ (Convolution, kernel size: 3×3) $\rightarrow 512 \times 32 \times 32$ (Convolution, kernel size: 3×3) $\rightarrow 1024 \times 16 \times 16$ (Convolution, kernel size: 3×3) $\rightarrow 256 \times 8 \times 8$. Then, we apply the fully connected layer to change into 2048-dimensional vector.

Handcrafted Features: The lower part of Figure 2 extracts conventional handcrafted features widely used in caricature recognition. In this work, we use the strategy of ensemble of local features to extract LBP [7], HOG [8], MoSIFT [9], GSRC [10], Canny [11], Haar [12], HSV [13] and YCbCr [14] histograms of the input image. Therefore, input image is equally partitioned into 8 horizontal stripes, and our features are composed of color features including HSV and YCbCr and texture features including LBP, HOG, MoSIFT, GSRC, Canny, and Haar. A 8-dimensional histogram is extracted for each channel and then normalized by L_1-norm [15]. All histograms are concatenated together to form 2048-dimensional vector. In this work, we denote the above handcrafted features as prior features.

Hybrid-Prior Features: We aim to jointly map network features and prior features to a unitary feature space. This combined features are called hybrid-prior features. A feature fusion deep neural network is proposed in order to use hand-crafted features to regularize network features so as to make network extract complementary features. Our fusion layer uses full connection to provide self-adaptation on caricature recognition problems. This fusion layer follows both handcrafted features and network features to output a 2048-dimensional vector.

3. EXPERIMENTS AND RESULTS

We evaluated the performance of our algorithm in terms of its accuracy. All experiments were conducted using a 4-core PC with an NVIDIA GTX 970 GPU, 16GB of RAM, and Ubuntu 16.

3.1. Experimental Settings and Baselines

We evaluate our approach on two related tasks: few-shot photo to caricature recognition and few-shot caricature to photo recognition on WebCaricature dataset [4], IIIT-CFW dataset [5] and Caricature-207 dataset [6] respectively. Few-shot photo to caricature recognition here is: given few real face of a public figure, we can recognize all the caricature faces of that public figure from a dataset of caricature. Few-shot caricature to photo recognition here is: given few caricature faces of that public figure, we can recognize all real face of a public figure from a dataset of caricature.

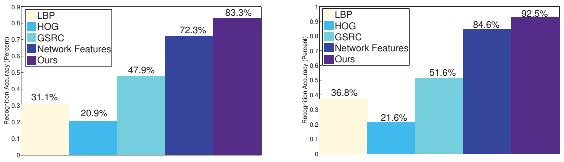
Specifically, the face images of the three datasets are resized to $256\times256\times3$. For the task of few-shot photo to caricature recognition, we choose the caricature images to conduct test set. For the task of few-shot caricature to photo recognition, we choose the real face images to conduct test set. On the WebCaricature dataset, we choose 80 real face images and 80 corresponding caricature images of the 80 individuals to construct training set, and the remaining real face (for few-shot photo to caricature recognition)/caricature (for few-shot caricature to photo recognition) samples of these 80 subjects are used as test set. On the IIIT-CFW dataset, we randomly choose 10 subjects for training and testing. For each subject, the selected randomly 10 real face images and 10 corresponding caricature images are used as the training sample, and the remaining selected randomly 10 real face images (for few-shot photo to caricature recognition)/10 corresponding caricature images (for few-shot caricature to photo recognition) are used for testing. On the Caricature-207 dataset, we randomly choose a subset of 100 individuals with one samples per person for training and testing. For these subjects, we choose 100 real face images and 100 corresponding caricature images for training and testing. We choose only 14 face images and of 14 persons to construct training set, and the remaining samples are test images. Note that we randomly choose 10 times as per the above strategy and take the average recognition accuracy for comparison.

Settings We use the Adam optimizer with a batch size of 1, for training where the learning rate was set to 0.00001 and momentums were set to 0.5 and 0.999.

Baselines We compare against various state of the art baselines for few-shot face recognition, including LBP+PCA [16], LBP+RPCA [17], VGG-16 [18], Resnet-50 [19], SEnet [20], VGG Features [18]+SDML [21], and VGG Features [18]+LSML [22].

3.2. Experiment on WebCaricature Dataset

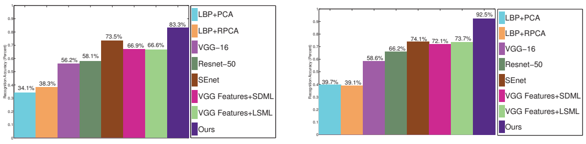
Ablation Study In order to verify the reasonableness and effectiveness of our hybrid-prior features on WebCaricature dataset, we design the Ablation experiment. We use the LBP [7], HOG [8], GSRC [10], and only network features and compared them with our proposed method. From Figure 3(a), for task of few-shot photo to caricature recognition on the Ablation experiment, the recognition accuracy of our method is 52.2%, 62.4%, 35.4%, and 11.0% higher than only using LBP, only using HOG, only using GSRC, and only network features, respectively. From Figure 3(b), for the task of few-shot caricature to photo recognition on the Ablation experiment, the recognition accuracy of our method is 55.7%, 70.9%, 40.9%, and 7.9% higher than only using LBP, only using HOG, only using GSRC, and only network features, respectively. It is clear that our method once again achieves the better results than the methods of only using one kind feature. This suggests that the design of hybrid-prior features is reasonable. The method of using only network features are less effective than using our hybrid-prior features for few-shot recognition. It shows our design of our hybrid-prior features are effective.



(a) Recognition Accuracies (Percent) on The Ablation Experiment for The Task of Few-Shot Photo to Caricature Recognition

(b) Recognition Accuracies (Percent) on the Ablation Experiment for The Task of Few-Shot Caricature to Photo Recognition

Fig. 3. Recognition Accuracies (Percent) of The Ablation Experiment



(a) Recognition Accuracies (Percent) on WebCaricature Dataset for The Task of Few-Shot Photo to Caricature Recognition

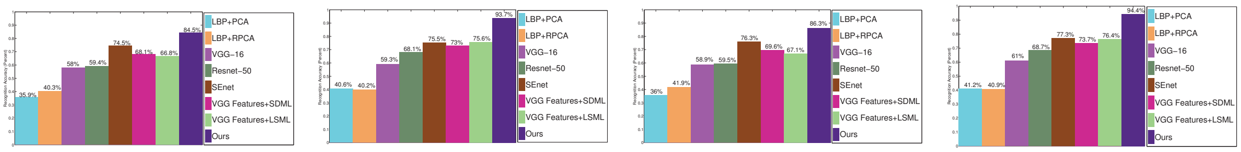
(b) Recognition Accuracies (Percent) on WebCaricature Dataset For The Task of Few-Shot Caricature to Photo Recognition

Fig. 4. Recognition Accuracies (Percent) of The Comparison Experiment on WebCaricature Dataset

In our first comparison experiment, we use the WebCaricature dataset for training and testing. According to Figure 4(a) and Figure 4(b), it can get the following two points:

Firstly, for task of few-shot photo to caricature recognition, the recognition accuracy of our method is 49.2%, 45.0%, 27.1%, 25.2%, 9.8%, 16.4%, and 16.7% higher than LBP+PCA [16], LBP+RPCA [17], VGG-16 [18], Resnet-50 [19], SEnet [20], VGG Features+SDML [18, 21], and VGG Features+LSML [18,22], respectively.

Secondly, for task of few-shot caricature to photo recognition, the recognition accuracy of our method is 52.8%, 53.4%, 33.9%, 26.3%, 18.4%, 20.4%, and 18.8% higher than LBP+PCA [16], LBP+RPCA [17], VGG-16 [18], Resnet-



(a) Recognition Accuracies (Percent) on IIIT-CFW Dataset for The Task of Few-Shot Photo to Caricature Recognition

(b) Recognition Accuracies (Percent) on IIIT-CFW dataset for The Task of Few-Shot Caricature to Photo Recognition

(c) Recognition Accuracies (Percent) on Caricature-207 Dataset for The Task of Few-Shot Photo to Caricature Recognition

(d) Recognition Accuracies (Percent) on Caricature-207 Dataset for The Task of Few-Shot Caricature to Photo Recognition

Fig. 5. Recognition accuracies (percent) of the Comparison experiments on IIIT-CFW dataset and on Caricature-207 dataset

50 [19], SEnet [20], VGG Features+SDML [18, 21], and VGG Features+LSML [18, 22], respectively.

From the above two points, it is clear that our method is more effective than LBP+PCA, and LBP+RPCA. Our proposed hybrid-prior features is more effective than the design of “one handcrafted feature + one of the family of principal component analysis” for few-shot caricature recognition. It is also clear that our method is more effective than VGG Features+SDML and VGG Features+LSML. This means our design of our hybrid-prior features is more effective than the design of “learning features + metric learning” for few-shot caricature recognition. In addition, our method is more effective than VGG-16, Resnet-50 and SEnet. It shows our design of our relation network is more effective than the design of other network-based methods for few-shot caricature recognition. All in all, it is obvious that our method is more robust than other caricature recognition methods for few-shot recognition.

3.3. Experiment on IIIT-CFW Dataset and on Caricature-207 Dataset

In our second comparison experiment and third comparison experiment, we use the IIIT-CFW dataset and Caricature-207 dataset for training and testing, respectively. From Figure 5(a), for task of few-shot photo to caricature recognition on IIIT-CFW dataset, the recognition accuracy of our method is 48.6%, 44.2%, 26.5%, 25.1%, 10.0%, 16.4%, and 17.7% higher than LBP+PCA [16], LBP+RPCA [17], VGG-16 [18], Resnet-50 [19], SEnet [20], VGG Features+SDML [18, 21], and VGG Features+LSML [18, 22], respectively. From Figure 5(b), for task of few-shot caricature to photo recognition on IIIT-CFW dataset, the recognition accuracy of our method is 53.1%, 53.5%, 34.4%, 25.6%, 18.2%, 20.7%,and 18.1% higher than LBP+PCA, LBP+RPCA, VGG-16, Resnet-50, SEnet, VGG Features+SDML, and VGG Features+LSML, respectively. According to Figure 5(c), for task of few-shot photo to caricature recognition on Caricature-207 dataset, the recognition accuracy of our method is 50.3%, 44.4%, 27.4%, 26.8%, 10.0%, 16.7% and 19.2% higher than LBP+PCA, LBP+RPCA, VGG-16, Resnet-50, SEnet, VGG Features+SDML, and VGG Features+LSML, respectively. According to Figure 5(d), for task of few-shot caricature to photo recognition on Caricature-207 dataset, the recognition accuracy of our method is 53.2%, 53.5%, 33.4%, 25.7%, 17.1%, 20.7% and 18.0% higher than LBP+PCA, LBP+RPCA, VGG-16, Resnet-50, SEnet, VGG Features+SDML, and VGG Features+LSML, respectively. In accordance with these points, our method achieves the best results among all methods on the two tasks. This means our method has the ability of using the network features or handcrafted features as other competing methods. This also shows our method is able to achieve robust performance on a wide range of distortion degrees and outperforms state-of-the-art methods for few-shot recognition.

4. CONCLUSION AND FUTURE WORK

We propose a few-shot learning approach for caricature recognition in this paper. Considering the strategy of learning inductive priors that it is necessary to introduce human prior models into artificial intelligence systems, we present an effective feature fusion approach to form hybrid-prior feature using network and the handcrafted feature for getting the better results of feature-based representative learning. Last but not the least, we build the two-branch relation network via meta-learning to compare between the features of different samples. In addition, we further introduce hybrid-prior feature into the relation network to enhance the recognition performance. Experimental results validate that the method achieves good recognition performance. It shows our method has strong robustness and high recognition accuracy.

In future research, on the one hand, we consider to extend to the task of few-shot caricature-visual face verification and identification [23]. On the other hand, we plan to study how to implement an algorithm in parallel platform.

5. REFERENCES

[1] Bahri Abaci and Tayfun Akgul, “Matching caricatures to photographs,” *Signal, Image and Video Processing*, vol. 9, no. 1, pp. 295–303, Dec 2015.

[2] Jing Huo, Yang Gao, Yinghuan Shi, and Hujun Yin,

"Variation robust cross-modal metric learning for caricature recognition," New York, NY, USA, 2017, Thematic Workshops '17, pp. 340–348, ACM.

[3] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin, "WebCaricature: a benchmark for caricature recognition," in *British Machine Vision Conference*, 2018.

[5] Ashutosh Mishra, Shyam Nandan Rai, Anand Mishra, and C. V. Jawahar, "IIIT-CFW: A benchmark database of cartoon faces in the wild," in *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou, Eds., Cham, 2016, pp. 35–47, Springer International Publishing.

[6] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, "Towards automated caricature recognition," in *2012 5th IAPR International Conference on Biometrics (ICB)*, March 2012, pp. 139–146.

[7] N. Werghi, C. Tortorici, S. Berretti, and A. Del Bimbo, "Boosting 3D LBP-based face recognition by fusing shape and texture descriptors on the mesh," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 964–979, May 2016.

[8] H. Tan, B. Yang, and Z. Ma, "Face recognition based on the fusion of global and local HOG features of face images," *IET Computer Vision*, vol. 8, no. 3, pp. 224–234, June 2014.

[9] Y. Zheng and G. An, "A novel facial expression recognition approach based on MoSIFT," in *6th International Conference on Wireless, Mobile and Multi-Media (ICWMMN 2015)*, Nov 2015, pp. 223–227.

[10] Meng Yang and Lei Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios, Eds., 2010, pp. 448–461.

[11] Rojana Kam-art, T. Raicharoen, and V. Khera, "Face recognition using feature extraction based on descriptive statistics of a face image," in *2009 International Conference on Machine Learning and Cybernetics*, July 2009, vol. 1, pp. 193–197.

[12] Y. Pang, X. Li, Y. Yuan, D. Tao, and J. Pan, "Fast Haar transform based feature extraction for face representation and recognition," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 441–450, Sept 2009.

[13] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 264–277, Sept 1999.

[14] H. Demirel and G. Anbarjafari, "Pose invariant face recognition using probability distribution functions in different color channels," *IEEE Signal Processing Letters*, vol. 15, pp. 537–540, 2008.

[15] F. Zhong, J. Zhang, and D. Li, "Discriminant locality preserving projections based on L1-norm maximization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 11, pp. 2065–2074, Nov 2014.

[16] M. J. Nordin and A. A. K. A. Hamid, "Combining local binary pattern and principal component analysis on T-Zone face area for face recognition," in *2011 International Conference on Pattern Analysis and Intelligence Robotics*, June 2011, vol. 1, pp. 25–30.

[17] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, July 2018.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[19] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] Weidi Xie and Andrew Zisserman, "Multicolumn networks for face recognition," in *BMVC*, 2018.

[21] Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang, "An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization," New York, NY, USA, 2009, ICML '09, pp. 841–848, ACM.

[22] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang, "Metric learning from relative comparisons by minimizing squared residual," in *2012 IEEE 12th International Conference on Data Mining*, Dec 2012, pp. 978–983.

[23] Jatin Garg, Skand Vishwanath Peri, Himanshu Tolani, and Narayanan.C Krishna, "Deep cross modal learning for caricature verification and identification (CaVINet)," in *Proceedings of the 2018 ACM Conference on Multimedia*. 2018, ACM.