# Deep Residual Feature Quantization for 3D Face Recognition

**walid Hariri** ( hariri@labged.net )

Labged Laboratory, Computer Science Department, ,Badji Mokhtar Annaba University

https://orcid.org/0000-0002-5909-5433

**Marwa Zaabi**

CEM Laboratory ENIG, Gabes University, Tunisia

**Research Article**

**Keywords:** Bag-of-Features, CNN, Codebook, Residual feature, Feature map

**Posted Date:** November 23rd, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1103780/v1

# Deep Residual Feature Quantization for 3D Face Recognition

1st Walid Hariri
*LABGED Laboratory, Computer Science Department*
*Badji Mokhtar Annaba University, Algeria*
*hariri@labged.net*

2nd Marwa Zaabi
*CEM Laboratory*
*ENIG, Gabes University, Tunisia*
*marwa.zaabi@isimg.tn*

***Abstract*—3D face recognition (FR) has been successfully applied using Convolutional neural networks (CNN) which have demonstrated stunning results in diverse computer vision and image classification tasks. Learning CNNs, however, need to estimate millions of parameters that expect high-performance computing capacity and storage. To deal with this issue, we propose an efficient method based on the quantization of residual features extracted from ResNet-50 pre-trained model. The method starts by describing each 3D face using a convolutional feature extraction block, and then apply the Bag-of-Features (BoF) paradigm to learn deep neural networks (we call it Deep BoF). To do so, we apply Radial Basis Function (RBF) neurons to quantize the deep features extracted from the last convolutional layers. An SVM classifier is then applied to classify faces according to their quantized term vectors. The obtained model is lightweight comparing to classical CNN and it allows classifying arbitrary-sized images. The experimental results on the FRGCv2 and Bosphorus datasets show the powerful of our method comparing to state of the art methods.**

***Index Terms*—Bag-of-Features, CNN, Codebook, Residual feature, Feature map.**

## I. Introduction

Face recognition in unconstrained scenarios has achieved impressive prosperity due to the explosion of CNNs compared to previous methods using hand-crafted feature extractors, like Local Binary Pattern and SIFT [1]. Hence, deep learning has been found to be more robust to classify faces in the presence of many variations like expression, rotation, and scale. Their outstanding performance has been shown in challenging classification tasks of a large amount of labeled image databases such as ImageNet [2]. In order to minimize the number of parameters in the network, CNN can be used with other models such as the Bag-of-Features model [3]. We have structured the remaining parts of the paper as follows, Section II gives an overview of the related works, and Section III describes the method. The experimental findings on FRGCv2 and Bosphorus datasets are presented in Section IV. Conclusions end the paper.

## II. Related works

Various methods have been suggested in the literature to minimize the model size of CNN, with the aim of improving CNN-based image classification methods. We may divide them into three groups:

**a) Global pooling strategies:** makes it possible to deal with arbitrary size as Spatial Pyramid Pooling, for example, He et al. [4] applied this technique which can produce a fixed-length representation and provide more robustness to object deformations. Malinowski and Fritz [5] have suggested a pooling operator parameterization. They then analyzed the impact of different regularizerson the pooling regions. Lastly, approximations to the proposed model are applied to allow efficient training.

**b) Compression and pruning techniques:** they focus on compressing an already trained CNN. They are also unable to manage images of varying sizes, resulting in a fixed due for feed-forwarding the CNN layers. As an example, Chen et al. [6] proposed a frequency-sensitive hashed nets that exploits the accrued tautology in the CNN layers. Thus, they realized high savings in memory and storage consumption. Han et al. [7] applied a pipeline pruning, quantization and Huffman encoding simultaneously. (1) Pruning the network by learning only the more relevant connections. (2), Quantization of the weights is applied to enforce weight sharing to be able to apply Huffman encoding. Further, to be able to deploy CNN on different systems with insufficient computing power, Wu et al. [8] proposed a quantized CNN framework, to jointly accelerate the computing power and reduce storage requirements.

**c) Feature vector quantization techniques:** the classical BoF paradigm aims to extract a number of segregated patches from the images, sampling a representative set of patches from the image, then giving a geometrical or visual descriptor vector, and using the resulting distribution of samples as a characterization of the image. BoF paradigm has been widely employed to quantize shallow features [9]. When dealing with trainable convolutional layers, BoF is considered as a pooling layer and makes it possible to classify different sized-images [10].

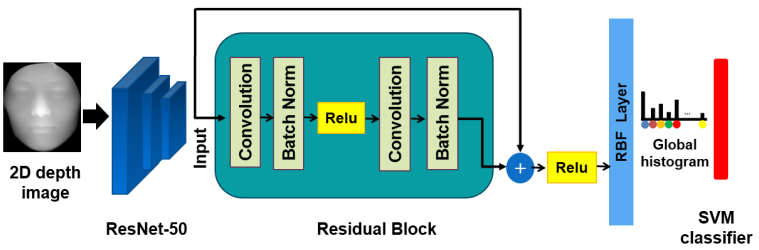


Fig. 1: The different steps of our proposal.

## III. METHOD

Fig. 1 depicts a summary of our FR proposal. The first step is the preprocessing. We apply different filters to remove facial deficiencies and noise (smoothing filter and median filter). Also discarding the unnecessary parts of the body is needed via the cropping filter. Next, each 3D face model is transformed to 2D depth image which takes into account the gray value of each image pixel in the following step. All 2D faces are normalized to 224 $\times$ 224 pixels which is the input size of the CNN models. After this stage, we extract deep features using ResNet-50 pre-trained network from the last convolutional layer as follows:

### A. Residual feature extraction layer

We extract residual deep features using ResNet-50 model [11]. It is a variant of ResNet pre-trained model on ImageNet dataset which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. It has 3.8 $\times$ $10^9$ Floating points operations. Fig. 2 shows in detail the architecture of this model. We only employ the Residual features shown in the same Figure. These features will be utilized in the next layer computation.

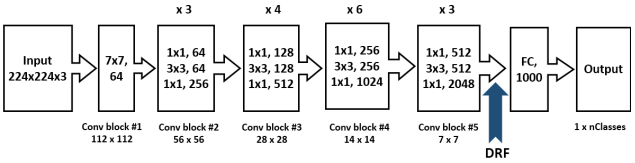


Fig. 2: ResNet-50 network architecture.

### B. Deep BoF layer

Using the obtained residual layer mentioned earlier, we extract feature maps (FMs) from the $i^{th}$ image. We used the RBF kernel as a similarity measure to estimate the similitude among these features and the *codewords*, also known as *term vector* as proposed in [12]. As a result, the first sublayer would be made up of RBF neurons, each of which is assigned to a codeword. The number of feature vectors acquired from the $i^th$ image is denoted by $p_i$. The initial setting of the RBF neurons can be done manually or automatically while creating the **codebook**. K-means is the most widely used automated algorithm. Let $P$ denote the set of all the feature vectors, identified by: $P = \{p_{ij}, i = 1 \dots p, j = 1 \dots p_i\}$ and $p_k$ denote the number of the RBF neurons centers referred by $n_k$. It's important to note that these RBF centers are learned afterward in order to obtain the final codewords. After that, the quantization is used to retrieve the histogram with a predetermined number of bins, each of which is referred to a *codeword*. The RBF layer, which has two sublayers, is then utilized as a similarity measure.

**(1) RBF layer**: determines how close the probe faces' input features are to the RBF centers.

The $j^{th}$ RBF neuron $\phi(X_j)$ is determined by:

$$\phi(X_j) = \exp(\|x - n_j\|_2/\sigma_j), \quad (1)$$

In this equation, $x$ denotes the feature vector where $n_j$ refers to the center of the $j^{th}$ RBF neuron.

**(2) Quantization layer:** this layer aims to combine the result of each neuron of RBF. We then obtain a global histogram that will be employed in the following for the classification stage. It is given by:

$$h_i = p_j \sum_{k}^{N^k} \phi(p_{jk}) \quad (2)$$

Where $\phi(p)$ refers to the yielded vector from the RBF layer through the $n_k$ bins.

### C. Dense layer and classification

The classification of the faces is carried out using an SVM classifier, with each face identified per a term vector. The back-propagation method is applied to train the deep BoF network using gradient descent. The 10-fold cross-validation method is carried out in our proposal on FRGCv2 and Bosphorus datasets. We refer to the term vector of each face by $P = [p1, \dots, p_k]$, where each $p_i$ denotes the number of the term $i$ in the given face. The codeword $P$ is used to identify test faces.

## IV. EXPERIMENTAL RESULTS

The proposed approach was evaluated using the FRGCv2 [13] and Bosphorus [14] datasets. We used Matlab R2017b on Win 7 with i7 CPU.

### A. Datasets description

**FRGCv2 dataset:** is a very challenging dataset for FR task. It includes 4007 3D face scans of 466 different subjects. The scans have been acquired in unconstrained scenario. The subjects are of different gender and age.

**Bosphorus dataset:** consists of 4,666 images attributed to 105 subjects. In addition to Expression variations, pose variations and occlusions are available in this dataset which make it more challenging than FRGCv2.

### B. Results and discussion

*1) **Setting**:* The FRGCv2 and Bosphorus faces has been adjusted as presented in the Section III. Using the normalized 2D depth faces of three sizes (i.e. 200 $\times$200, 224 $\times$ 224, 250 $\times$ 250 pixels), we use ResNet-50 pre-trained model to extract Residual features as presented in Section III-A. In this layer the FMs are of 7 $\times$ 7 dimension with 2048 channels. The global histogram is then extracted using the quantization-based method as shown in Section III-B. Finally, An SVM classifier is employed to attribute each face to its possible identity as presented in Section III-C. In this experiment, to validate the model, the 10 fold cross-validation technique is used, mini batches of 50 along with 50 iterations are applied. Note that all the steps of the proposed method are fully automatic. Note that we have followed the same protocol applied in [15] in order to make a fair comparison.

*2) **Results on FRGCv2 dataset**:* Table I shows the classification rates on the FRGCv2 dataset. Three different dimensions of codebooks are tested including fifty, sixty and seventy term vectors per image. It can be noticed that using the third FMs in the fifth block layer with sixty term vectors, we get the highest recognition rate by 98.9%.

*3) **Results on Bosphorus dataset**:* The classification rates using 3 codebooks on the Bosphorus dataset are shown in Table II. We can see that the highest performance has been achieved is 97.3% using the third FMs.

Table III and IV present the comparison between the recognition rate of the obtained method and state of the art accuracy on FRGCv2 and Bosphorus datasets respectively. We can see that our proposed approach outperforms handcrafting and classical CNN-based methods. This accuracy is achieved due to the Residual deep features extracted from the ResNet-50 model and their effective quantization.

TABLE I: Classification performance on FRGCv2 dataset.

| **Feature** term vectors | **FM-Size** 50 | **FM-Size** 60 | **FM-Size** 70 |
|---|---|---|---|
| Conv5 FM1 | 91.2% | 93.9% | 96.1% |
| Conv5 FM2 | 93.8% | 94.8% | 95.4% |
| Conv5 FM3 | 95.4% | **98.9%** | 98.1% |

TABLE II: Classification performance on Bosphorus dataset.

| **Feature** term vectors | **FM-Size** 50 | **FM-Size** 60 | **FM-Size** 70 |
|---|---|---|---|
| Conv5 FM1 | 93.2% | 91.7% | 91.3% |
| Conv5 FM2 | 95.2% | 96.0% | 95.3% |
| Conv5 FM3 | 95.4% | 97.0% | **97.3%** |

TABLE III: Comparison with the FRGCv2 dataset.

| **Method** | **Classification rate** |
|---|---|
| Alyuz et al. [15] | 97.5% |
| Elaiwat et al. [16] | 94.4% |
| Drira et al. [17] | 97.0% |
| Faltemier et al. [18] | 97.2% |
| Chong et al. [19] | 90.87% |
| **Our method** | **98.9%** |

TABLE IV: Comparison with the Bosphorus dataset.

| **Method** | **Classification rate** |
|---|---|
| Alyuz et al. [15] | 95.9% |
| Li et al. [20] | 95.4% |
| **Our method** | **97.3%** |

## V. Conclusion

A new methodology using a deep bag-of-features paradigm for 3D face recognition is proposed. Between the last convolutional block and the dense layer, a BoF-based pooling layer is used instead of a traditional CNN that feeds the extracted vectors directly into the fully connected layer. Therefore, we extract a global histogram quantization from the residual features of the ResNet-50 model. The obtained representation is independent of the image size and reduces considerably the number of parameters that we can find in a classical CNN. High recognition accuracy is achieved. Note that our method is generic so that other 2D map images can be used (e.g. Shape index and curvature maps). As a limitation, our method is based only on residual features which cannot provide a high generalization power in the presence of many variations. Hence, other pre-trained models can be added to the framework (e.g. Alexnet, ExpNet).

## References


[1] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B.-L. Lu, "Person-specific sift features for face recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–593.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision*. Springer, 2006, pp. 490–503.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[5] M. Malinowski and M. Fritz, "Learnable pooling regions for image classification," *arXiv preprint arXiv:1301.3516*, 2013.

[6] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing convolutional neural networks," *arXiv preprint arXiv:1506.04449*, 2015.

[7] S. Han, H. Mao, and W. J. Dally, "A deep neural network compression pipeline: Pruning, quantization, huffman encoding," *arXiv preprint arXiv:1510.00149*, vol. 10, 2015.

[8] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4820–4828.

[9] H. Lobel, R. Vidal, D. Mery, and A. Soto, "Joint dictionary and classifier learning for categorization of images using a max-margin framework," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2013, pp. 87–98.

[10] N. Passalis and A. Tefas, "Neural bag-of-features learning," *Pattern Recognition*, vol. 64, pp. 277–294, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] W. Hariri, "Efficient masked face recognition method during the covid-19 pandemic," *arXiv preprint arXiv:2105.03026*, 2021.

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE computer society conference on Computer vision and pattern recognition, 2005. CVPR 2005.*, vol. 1. IEEE, 2005, pp. 947–954.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*. Springer, 2008, pp. 47–56.

[15] N. Alyüz, B. Gökberk, and L. Akarun, "Regional registration for expression resistant 3-d face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 425–440, 2010.

[16] S. Elaiwat, M. Bennamoun, F. Boussaid, and A. El-Sallam, "3-d face recognition using curvelet local features," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 172–175, 2014.

[17] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3d face recognition under expressions, occlusions, and pose variations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2270–2283, 2013.

[18] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn, "A region ensemble for 3-d face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 62–73, 2008.

[19] L.-Y. Chong, A. B. J. Teoh, and T.-S. Ong, "Range image derivatives for grcm on 2.5 d face recognition," in *Information Science and Applications (ICISA) 2016*. Springer, 2016, pp. 753–763.

[20] H. Li, D. Huang, J.-M. Morvan, L. Chen, and Y. Wang, "Expression-robust 3d face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns," *Neurocomputing*, vol. 133, pp. 179–193, 2014.