Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 143 (2018) 998–1006

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Computing and Communication (ICACC-2018)

Aggregation of Deep Local Features using VLAD and Classification using R^2 Forest

Vinay A, Harsh Garg, Ankit Anand, Rajat Nigam, Abhijay Gupta, K N Balasubramanya Murthy, S Natarajan

Center for Pattern Recognition and Machine Intelligence, PES University, Bangalore 560085, India

Abstract

The paper proposes an efficient and accurate model for face recognition using an attentive local feature descriptor extracted from Convolutional Neural Network referred to as DEep Local Feature (DELF). The algorithm mentioned formerly is used for extracting descriptors from the images using a fully convolutional network which are trained with weak supervision and using image level classes, neglecting the usage of patch and object level annotations. The physical characteristics such as colour, texture, etc are represented in the form of 40 dimensional vectors using DELF. Further, such descriptors are quantized to represent them into the compact form using Vector of Locally Aggregated Descriptors and Fisher kernels. Subsequently, such vectors are used for multi-class image classification using ensemble learning methods including Rotation Forest and Random Forests. Comparative study between both the classifiers and feature aggregation methods are performed and tabulated in the paper.

Keywords: DELF; Rotation Forest; Fisher Kernel; VLAD; Face Recognition

1. Introduction

Face Recognition (FR) has been one of the pre-eminent fields in the domain of computer vision and image analysis. It has been used in many different sectors by enterprises or government for different applications. Every year a significant amount of improved methods and algorithms are proposed by computer vision enthusiasts for efficient and accurate models for face recognition. Dealing with such a great number of algorithms and techniques on different hardware platforms has also been one of the important aspects of this field. Some algorithms prove to be better when

[*] Corresponding author. Tel.: 08026721983 ; fax: 08026720886.
E-mail address: a.vinay@pes.edu, harshagarwal76@gmail.com

10.1016/j.procs.2018.10.334

ran on laptops / personal computers for face authentication, whereas some other prove to be better when ran on embedded devices like mobile phones for the same purpose. These different aspects to the problem of face recognition has given it widespread applications ranging from authentication systems at public places to authorization mechanisms in mobile phones and personal computers. Such usage of FR applications has overpowered the use of traditional authorization mechanisms like password, pin or patterns mainly because of its ease of usage.

There has always been a need of efficient and accurate models of image recognition which can be used for its applications like Face Recognition. The work proposed in the paper involves an efficient model for face recognition which can be used in different constrained scenarios like pose, scale, expression and illumination. One of the most important steps in face recognition, i.e., feature description, is performed by DEep Local Feature (DELF) [19]. DELF is an attentive local feature descriptor which can be used for large-scale image retrieval. A single pass over the CNN network enables the extraction of local descriptors and keypoints. Further the descriptors obtained from DELF are quantized into a compact form using Fisher Kernels and VLAD. Subsequently, the quantized vector is sent for multi-class image classification using Rotation Forest and Random Forest.

2. Related Work

In Face Recognition, the first and most integral step is to extract a vector of features that represent the important aspects of the face present in the image. Then, the key attributes of these features have to be determined and stored in a vector. Once we have obtained the features of the image, it becomes easy to compare and classify images.

Earlier, handcrafted models such as SIFT [2], SURF [3], ORB [4], etc were extensively used for the task of keypoint detection and description. With the advancements in deep learning, a lot of architectures have been developed ([5] - [7]). DELF [8] proposes the use of an attention mechanism on its deep learning architecture to detect and describe keypoints in a facial image.The method discussed in [8] is the state of the art model and outperforms all handcrafted as well as deep learning models by a huge margin.

Feature quantization is an important step towards person identification as it retains only the most important aspects of the descriptor matrix. A number of techniques for feature aggregation have been proposed in the literature such as the bag of visual words [9], fisher vectors [10], VLAD [11] etc. [10] describes patch encoding through derivation of a generative Gaussian mixture model making it computationally inexpensive. [11] is the state of the art feature aggregator for creating visual codebooks.

Rotation forest [12] and random forest [13] are methods based on ensemble learning using decision trees. [12] finds its use in human pose recognition [14], gene selection and classification [15] and estimating carbon dioxide emission from deforestation [16]. [13] is a PCA [33] based decision tree learning method used for classification of Alzheimer disease [17], bankruptcy detection and credit scoring [18].

3. Proposed Work

The face matching pipeline follows these 3 phases: (i) Deep local feature extraction from images & dimensionality reduction (ii) Feature descriptor aggregation (iii) Feature matching. Fig.1 given below depicts the Feature matching methodology.

3.1. Deep Local feature extraction & Dimensionality reduction

The feature extraction method uses Fully Convolutional Network (FCN) to extract the dense grid of local descriptors. The FCN is taken from the ResNet50 [1] model, using the output of conv4_x convolutional block. For the purpose of scale invariance in different regions of an image, the image pyramid is constructed and further used by FCN which is applied to each level independently. The image pyramid represents features that describe different locations in the image. The feature maps resulted after application of FCN to each layer of the pyramid form a dense grid of local descriptors. The localisation of the features is done by considering their corresponding receptive fields in the image, which is obtained by the configuration of the convolution and pooling layers of the FCN.

To make the algorithm more robust, fine tuning is done to the local descriptors which further increase the discriminativeness in them. For the purpose of image classification, ResNet50 model is trained using standard cross-entropy

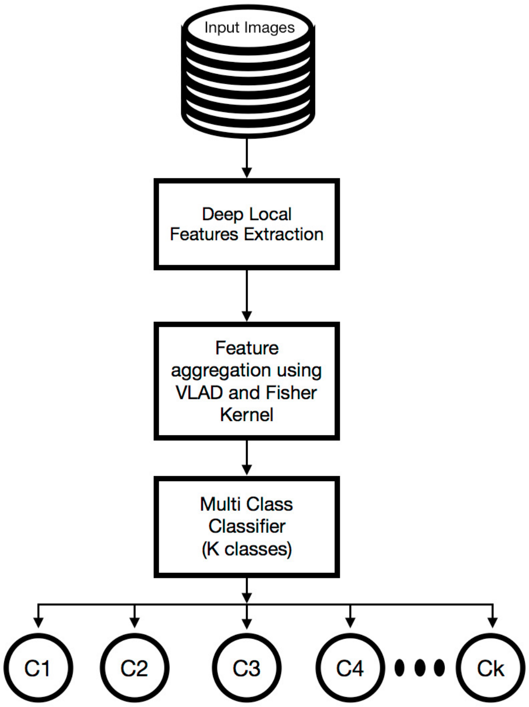
Input Images
Deep Local Features Extraction
Feature aggregation using VLAD and Fisher Kernel
Multi Class Classifier (K classes)
C1
C2
C3
C4
Ck

Fig. 1. Proposed Approach

loss. The input to the model are images which are center cropped and rescaled to 250 × 250. A set comprised of 224 × 224 crops randomly are used for training. The method provides local descriptors which learn representations of images which can also be obtained by using object or patch-level labels. The dimensionality of the obtained feature descriptors is reduced and filtered by first performing L2 normalization and then using PCA to reduce it to 40. The 40 dimensions of the descriptors provide good discriminativeness and compactness. The descriptors obtained are again L2 normalised after PCA. Further, the descriptors obtained are fed into the model for feature quantization. The proposed architecture of DELF is shown in Fig.2. The results of DELF have been compared in [19] with several other recent algorithms which provide local and global descriptors for the given image.

Four different algorithms are taken in for comparison with DELF. Deep Image Retrieval (DIR) [28] provides multi-resolution descriptors and are 2048 dimensional. siaMAC [29] extracts 512 dimensional global descriptor using a VGG16 [30]. CONGAS [31] uses a Laplacian of Gaussian keypoint detector after which Gabor wavelet responses is applied at the output of aforementioned keypoint detector which extracts a 40 dimensional feature descriptor. LIFT [32] which jointly learns keypoint detection, description and orientation estimation, extracts 128 dimensional features. For Performance comparison a new version of Precision and recall is considered as given below:

$$P_{RE} = \frac{\sum_q |R_q^{TP}|}{\sum_q |R_q|} \quad and \quad R_{EC} = \sum_q |R_q^{TP}| \tag{1}$$

where R_q is the set of images retrieved for query instance q for a threshold given and $R_q^{TP} (\subseteq R_q)$ denotes the set of true positives. The results comparison of DELF studied in [19] with all the other 4 algorithms shows that DELF outperforms these algorithms even in a challenging dataset having partial occlusion in the background, whereas the other

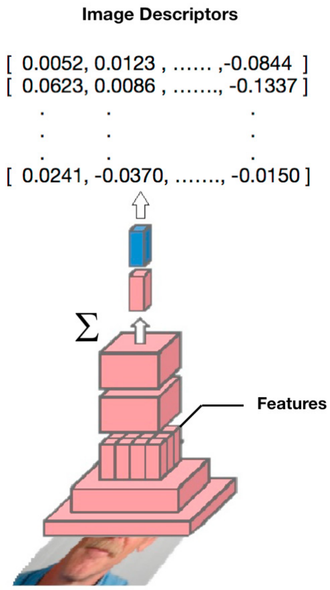

Fig. 2. DELF Architecture

algorithms for obtaining the global feature descriptors such as DIR and CONGAS perform poorly in the challenging dataset. The results obtained show that DELF is accurate for application in the domain of face recognition.

3.2. Feature quantization using VLAD

After extracting the descriptors from the image, we aggregate them into a compact vector representation using the locality criterion in the feature space. For each visual word c_i the VLAD [20] aggregator accumulates the difference between the local descriptor x and the assigned nearest visual word c_i as $x - c_i$. This is done for characterizing the vector distribution with respect to the center. The dimension D of the model is $k \times d$ where the dimension of the local descriptor is denoted by d and k is the number of visual words. v is represented by:

$$v_{i,j} = \sum_{\text{x such that NN(x) is } c_i} x_i - c_{i,j} \tag{2}$$

where, the descriptor is given by $v_{i,j}$, i is the visual word and j is the local descriptor component. Also, x_j represents the j^{th} component of the descriptor x and $c_{i,j}$ represents its corresponding visual word c_i. The descriptors obtained from VLAD are structured and relatively sparse, concluding that the higher values of the descriptor are positioned in the same cluster. The quantized vector is further fed into the model for classification using supervised algorithms.

3.3. Feature quantization using Fisher Kernel

After extracting the descriptors from the previous step, we aggregate them to represent them in a compact vector using Fisher Kernels [21]. [21] is used to convert the incoming set of variable size descriptors independent of each other into a fixed size vector representation. The aggregated vector is the slope of descriptors likelihood computed on the distribution parameters. The slope gives the direction in the space of parameters using which, the distribution should be changed to fit in a better way for the observed data. It has been used in [22] for image classification using Gaussian mixture model. The vector representation of dimension $(2d + 1) \times k - 1$ is obtained for an image feature set, where k is the number of components of diagonal variance matrices in mixture and d is the dimension of local descriptor. Such aggregated form of vectors is sent further in the model for classification.

3.4. Feature matching

3.4.1. Random Forest

Random Forest [24] is used as a supervised learner which applies bagging/bootstrap aggregating technique to tree learners. The bootstrapping leads to a better performance of the model since it decreases the variance without increasing the bias. This makes it less sensitive to noise. The difference from normal tree learning algorithm is that, here random subspace method is used, where at each candidate split, a random subset of the features is used. This randomness helps in removing the correlation between any two trees in the forest, which is major factor affecting the forest error rate. For random forests, $h_k(X) = h(X, \Theta_k)$ where $h_k(X)$ represents the k^{th} classifier in the forest and Θ_k represents a random vector for the tree. In accordance to the strong Law of Large numbers, with increasing number of trees, for almost all sequences Θ_1.., the generalisation error PE^* converges to

$$P_{X,Y}\left(P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0\right) \tag{3}$$

where, X, Y are indicative of the probability over the space X, Y. This shows why random forests do not overfit when the number of trees are increased. Random forests also provide an effective method to estimate missing values. Few of the major advantages of random forests are that they do not overfit, they dont need cross-validation to get an unbiased estimate of the test set error and they are fast.

3.4.2. Rotation Forest

Rotation forest [25] is a technique for generating classifier ensembles based on feature extraction. To obtain the training data for the base classifier, the feature set is split randomly into K subsets. Following this, PCA is applied to each subset. In order to generate new features for the base classifier, K axis rotations are applied. This rotation helps promote diversity and individual accuracy within the ensemble model.

Given a data point $x = [x_1, x_2..x_n]^T$, described by n features and a $N \times n$ matrix(X), representing the dataset containing training objects. Also given $Y = [y_1, y_2, y_n]^T$ is a vector of class labels for the data, where y_i takes a value from the set of class labels. $D_1...D_L$ represent the classifiers in the ensemble. For constructing the training set for classifier D_i, these steps are followed:

First, the feature set represented by F is randomly divided into K subsets, where K is a parameter of the algorithm. To maintain simplicity, K is chosen to be a factor of n so that each individual feature subset consists of $M = n/K$ features. In order to maintain high variety, disjoint subsets are chosen. Then, for every subset $F_{i,j}$ (where $F_{i,j}$ represents the j^{th} feature subset for training set of classifier D_i), a subset of classes is randomly selected, (making sure they are not empty) and then a bootstrap sample of objects is drawn of size equal to 75% of the data. Now, the PCA algorithm is run using the M features in $F_{i,j}$ and the subset of X which was selected. The coefficients of the principal components represented by $a_{i,j}^{(1)},...a_{ij}^{M_j}$,(each is of size $M \times 1$) are stored. PCA is run on a subset of classes to avoid the coefficients from being identical in the case where the same feature subset is chosen for different classifiers. The obtained vectors with coefficients are arranged in a sparse rotation matrix denoted by R_i with dimensions $n \times \sum_j M_j$.

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, ..a_{i,1}^{(M_1)} & [0] & [0] \dots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, ..a_{i,2}^{(M_2)} & [0] \dots & [0] \\ \vdots & \vdots & \vdots \ \ddots & \vdots \\ [0] & [0] & [0] \dots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, ..a_{i,K}^{(M_K)} \end{bmatrix}$$

In order to calculate the training set for classifier D_i, the columns of R_i (which represents the features) are rearranged, so that they correspond to the original features. This rearranged rotation matrix is denoted by R_i^a which has dimensions $N \times n$. Now, for the classifier, the training set is given by XR_i^a.

For classification, for a given x, for each class w_j, the confidence is calculated using average combination method and the class with the maximum confidence is assigned to x.

4. Dataset Description

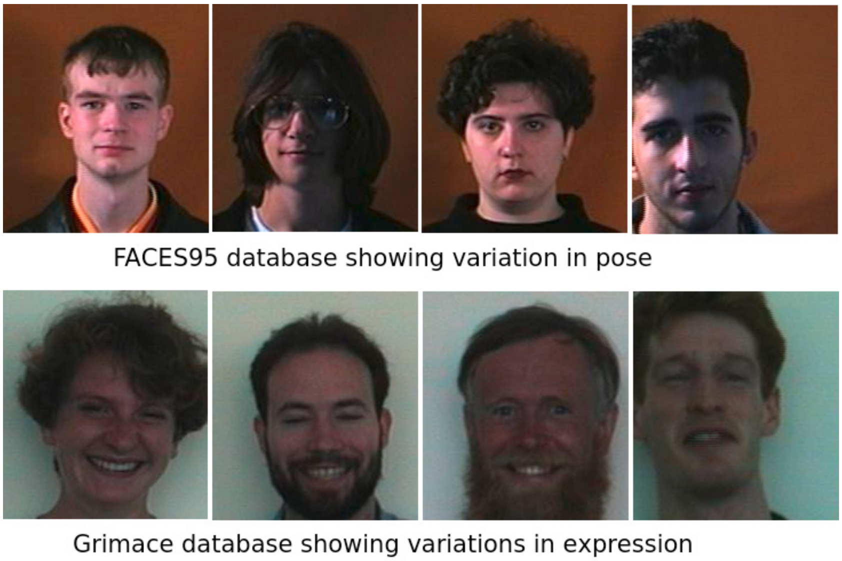
FACES95 database showing variation in pose
Grimace database showing variations in expression

Fig. 3. Datasets reflecting variation in pose, expression, illumination

To test our proposed methodology we use FACES95 [26], GRIMACE [27] as standard datasets and executed our model. Fig.3 shows the datasets having variance in pose, expression, scale and illumination.

- FACES95 - FACES95 contain 1440 images of 72 individuals where each class has 20 images. The mentioned dataset was constructed by repeatedly taking snapshots of 72 individuals with the delay of 0.5 seconds between successive frames. The dataset contains significant head pose variations between successive images of the same individual.
- GRIMACE - GRIMACE contain 360 images of 18 individuals. The images contained in the dataset consists of significant variation in scale, lighting and position of face in the image. Also, the grimaces were made after moving the head of the individuals which gets extreme towards the end of the sequence.

5. Results

For the purpose of validation of results, the dataset of facial images were split into training and validation sets. To check the accuracy of our proposed approach, we executed our model on two benchmark datasets, i.e, FACES95 and GRIMACE. The datasets considered for our testing purpose exhibited variance in head pose, illumination, expression, and translation of faces. The classifiers performed well on both the datasets and proved to be 83.70% and 95.41% accurate on FACES95 and GRIMACE respectively. To perform detailed experimentation we alternated the feature aggregation techniques between Fisher kernels and VLAD with similar change in classifiers between Rotation and Random forest. Such changes in quantization techniques and classifiers has given promising results. When tested on FACES95 with Fisher kernels, the Random Forest classifier turned to be 82.96% accurate, Fisher kernels and Rotation

Forest classifier turned to be 75.00% accurate. Upon changing the aggregation method from Fisher kernels to VLAD, the classifier as Random Forest and Rotation Forest it gave an accuracy around 83.70% and 72.60% respectively. Similar but enhanced results are achieved upon testing the model on GRIMACE dataset which bear significant variation in illumination and expression. The model when tested with Fisher Kernels as the aggregator and Random Forest as the classifier proved to be 95.41% accurate and with Rotation Forest to be 77.8% accurate. Upon testing the model on GRIMACE with VLAD as aggregator and Rotation forest as classifier, it gave an accuracy of 96.50% whereas with Random Forest, it gave an accuracy of 93.33%. The results for the same are shown in Table 1 and Table 2 for Fisher kernels and VLAD respectively.

Table 1. Fisher Kernel Results

Dataset	Random Forest (*Accuracy in %*)	Rotation Forest (*Accuracy in %*)
FACES95	82.9%	75.00%
GRIMACE	95.41%	77.78%

Table 2. VLAD Results

Dataset	Random Forest (*Accuracy in %*)	Rotation Forest (*Accuracy in %*)
FACES95	83.70%	72.60%
GRIMACE	93.33%	96.50%

Further, a 10 fold cross validation is performed on training dataset to obtain mean accuracy over all the 10 subsets of the training sets. Adding to the same, the hyperparameters like class_weight, split criterion, max_depth of trees and number of estimators are tuned for better results in the ensemble model. Evaluation metrics like accuracy, precision, recall and F1-score are computed from the classified results. The metrics computed previously are shown and compared in Fig 4,5,6 and 7.

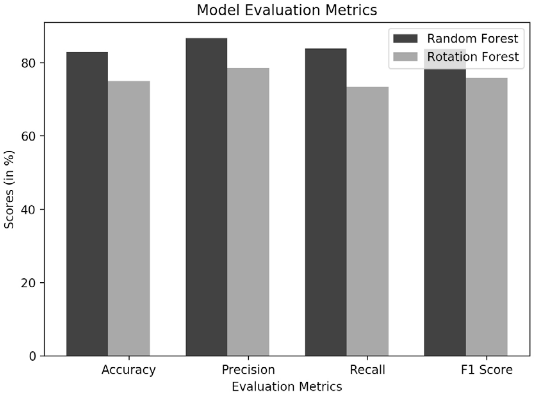

Fig. 4. Fisher Kernel on FACES95 dataset

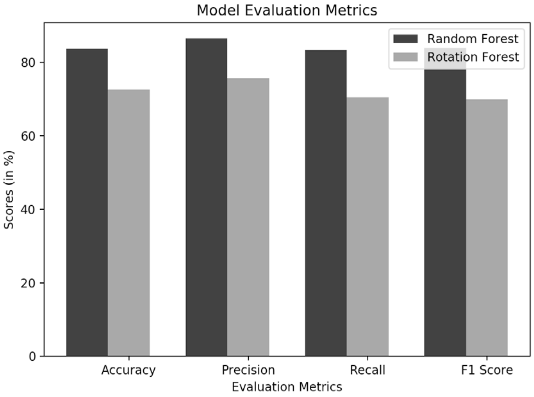

Fig. 5. VLAD on FACES95 dataset

However, we cannot perform a direct comparison between the state of art models available in literature. This is mainly because of the different reasons which are mentioned below in the context. One of the important reasons include the application of different preprocessing steps and application of the same on different results. Such a change might result into a difference in the accuracy and speed of the model. Also testing the model on different datasets might not result with the same accuracy and speed. In this paper the experiments cannot be directly compared on standard datasets like CelebA because it contain images for different range of classes like eye-glasses, wearing hat, wavy hair etc. Such datasets can be used for other computer vision tasks like face attribute recognition, facial part localization. At the same time, changing hardware configuration and implementation algorithms also affect the model

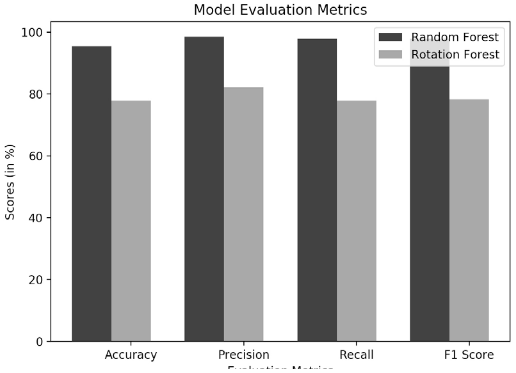

Fig. 6. Fisher Kernel on GRIMACE dataset

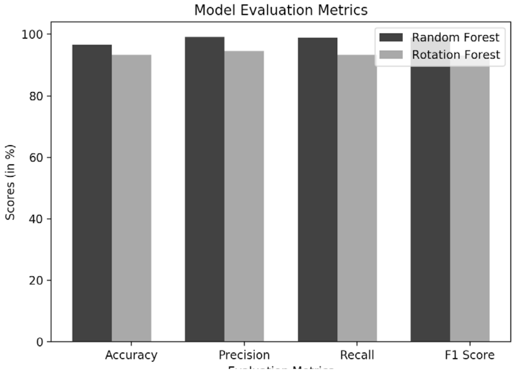

Fig. 7. VLAD on GRIMACE dataset

to some extent. However, [23] has compared the different techniques for FR applications including PCA, SVM, ICA, Gabor wavelet, LDA and ANN. Such a comparison has concluded that for enhanced face detection and recognition, hybrid methods for soft computing tools should be used such as ANN, Gabor Filter achieves better results in terms of accuracy. Also, for making the model more real-time and practical we need to consider cases for occlusions and speed of recognition.

6. Conclusion

A face recognition application using DEep local feature based descriptor is proposed in this paper, which is derived from a CNN based model. To properly evaluate the performance of the ensemble classifiers, we performed tests on FACES95 and GRIMACE datasets. It can be concluded that by keeping the aggregator as Fisher kernel, Random Forest performed better than Rotation Forest on both the benchmark datasets, whereas in case of VLAD, Rotation Forest gave better results than Random Forest.

References

[1] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[2] Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. 2004 Nov 1;60(2):91-110.

[3] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). Computer vision and image understanding. 2008 Jun 1;110(3):346-59.

[4] Rublee E, Rabaud V, Konolige K, Bradski G. ORB: An efficient alternative to SIFT or SURF. InComputer Vision (ICCV), 2011 IEEE international conference on 2011 Nov 6 (pp. 2564-2571). IEEE.

[5] Altwaijry H, Veit A, Belongie SJ, Tech C. Learning to Detect and Match Keypoints with Deep Architectures. InBMVC 2016 Sep.

[6] Simo-Serra E, Trulls E, Ferraz L, Kokkinos I, Fua P, Moreno-Noguer F. Discriminative learning of deep convolutional feature point descriptors. InComputer Vision (ICCV), 2015 IEEE International Conference on 2015 Dec 7 (pp. 118-126). IEEE.

[7] Yi KM, Trulls E, Lepetit V, Fua P. Lift: Learned invariant feature transform. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 467-483). Springer, Cham.

[8] Noh H, Araujo A, Sim J, Weyand T, Han B. Large-Scale Image Retrieval with Attentive Deep Local Features. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 Oct 1 (pp. 3456-3465).

[9] Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. InWorkshop on statistical learning in computer vision, ECCV 2004 May 15 (Vol. 1, No. 1-22, pp. 1-2).

[10] Snchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: Theory and practice. International journal of computer vision. 2013 Dec 1;105(3):222-45.

[11] Wang Y, Duan LY, Lin J, Wang Z, Huang T. Hierarchical multi-VLAD for image retrieval. InImage Processing (ICIP), 2015 IEEE International Conference on 2015 Sep 27 (pp. 4629-4633). IEEE.

[12] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[13] Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence. 2006 Oct;28(10):1619-30.

[14] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-time human pose recognition in parts from single depth images. InComputer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on 2011 Jun 20 (pp. 1297-1304). Ieee.

[15] Daz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006 Dec;7(1):3.

[16] Baccini AG, Goetz SJ, Walker WS, Laporte NT, Sun M, Sulla-Menashe D, Hackler J, Beck PS, Dubayah R, Friedl MA, Samanta S. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Climate Change. 2012 Mar;2(3):182.

[17] Liu M, Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Ensemble sparse classification of Alzheimer's disease. NeuroImage. 2012 Apr 2;60(2):1106-16.

[18] Nanni L, Lumini A. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert systems with applications. 2009 Mar 1;36(2):3028-33.

[19] Noh H, Araujo A, Sim J, Weyand T, Han B. Large-Scale Image Retrieval with Attentive Deep Local Features. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 Oct 1 (pp. 3456-3465).

[20] Jgou H, Douze M, Schmid C, Prez P. Aggregating local descriptors into a compact image representation. InComputer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on 2010 Jun 13 (pp. 3304-3311). IEEE.

[21] Jaakkola T, Haussler D. Exploiting generative models in discriminative classifiers. InAdvances in neural information processing systems 1999 (pp. 487-493).

[22] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. InComputer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on 2007 Jun 17 (pp. 1-8). IEEE.

[23] Javed M, Gupta B. Performance comparison of various face detection techniques. International Journal of Scientific Research Engineering & Technology (IJSRET) Volume. 2013 Apr;2:019-027.

[24] Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE transactions on pattern analysis and machine intelligence. 2006 Oct;28(10):1619-30.

[25] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[26] FACES95: http://cswww.essex.ac.uk/mv/allfaces/faces95.html

[27] GRIMACE: http://cswww.essex.ac.uk/mv/allfaces/grimace.html

[28] Gordo A, Almazn J, Revaud J, Larlus D. Deep image retrieval: Learning global representations for image search. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 241-257). Springer, Cham.

[29] Radenovi F, Tolias G, Chum O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 3-20). Springer, Cham.

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[31] Neven H, Rose G, Macready WG. Image recognition with an adiabatic quantum computer I. Mapping to quadratic unconstrained binary optimization. arXiv preprint arXiv:0804.4457. 2008 Apr 28.

[32] Yi KM, Trulls E, Lepetit V, Fua P. Lift: Learned invariant feature transform. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 467-483). Springer, Cham.

[33] Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010 Jul 1;2(4):433-59.