Periocular Recognition Using CNN Features Off-the-Shelf

Kevin Hernandez-Diaz[1], Fernando Alonso-Fernandez[1] and Josef Bigun[1]

Abstract: Periocular refers to the region around the eye, including sclera, eyelids, lashes, brows and skin. With a surprisingly high discrimination ability, it is the ocular modality requiring the least constrained acquisition. Here, we apply existing pre-trained architectures, proposed in the context of the ImageNet Large Scale Visual Recognition Challenge, to the task of periocular recognition. These have proven to be very successful for many other computer vision tasks apart from the detection and classification tasks for which they were designed. Experiments are done with a database of periocular images captured with a digital camera. We demonstrate that these off-the-shelf CNN features can effectively recognize individuals based on periocular images, despite being trained to classify generic objects. Compared against reference periocular features, they show an EER reduction of up to ~40%, with the fusion of CNN and traditional features providing additional improvements.

Keywords: Periocular recognition, deep learning, biometrics, Convolutional Neural Network.

1   Introduction

Periocular biometrics represent a trade-off between using the iris (which may not be available in long distances due to low resolution, e.g. surveillance) and the face (which may be partially occluded, either accidentally or intentionally). It has attracted noticeable interest in recent years, being one the most discriminative regions of the face [AFB16]. It is also very suitable for non-cooperative biometrics, since it can be captured largely relaxing the acquisition conditions. In addition, it has shown to be more resistant to expression variation [SBK15], aging [Ju11], plastic surgery [JR12], or gender transformation [MRA14], as compared with the entire face. Apart from serving as a stand-alone modality, it can also be combined with iris or face [AMB15] to improve the overall performance.

Convolutional Neural Networks (CNNs) have become a very popular tool in many vision tasks. Their application to biometrics is however limited, with recent works on face recognition [PVZ15] and detection [Li15], iris recognition [Ng18], soft-biometrics, and image segmentation [BK17]. One reason of such limited research is the lack of big amounts of training data, as required by deep learning methods. Inspired by the work of Nguyen *et al.* in iris [Ng18], this paper leverages the power of existing pre-trained architectures which have proven to be successful in very large recognition tasks. This eliminates the necessity of designing and training new CNNs for the specific task of periocular recognition, which may be infeasible given the mentioned lack of large-scale data. In particular, we choose those from the series of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). These are trained on more than a million images of the ImageNet database[3],

---

[1] School of Information Technology (ITE), Halmstad University, Box 823, 30118 Halmstad, Sweden, Email: kevin.hernandez-diaz@hh.se, feralo@hh.se, josef.bigun@hh.se
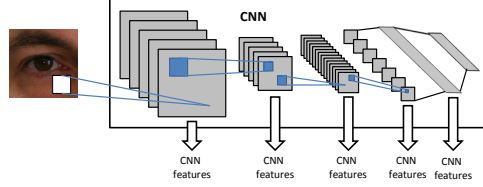
[3] http://www.image-net.org

Fig. 1: Extraction of periocular features from different CNN layers.

and it can classify images into 1000 object categories. Apart from obtaining leading positions in ILSVRC, these off-the-shelf networks have also proven very successful in many other tasks apart from the object detection and classification for which they are designed [Ra14]. Here, we investigate those architectures in the context of periocular recognition. We compare them against traditional baseline Local Binary Patterns (LBP) [OPM02], Histogram of Oriented Gradients (HOG) [DT05] and Scale-Invariant Feature Transform (SIFT) key-points [Lo04]. We employ a database of 1,718 periocular images captured with a digital camera at several distances. Our experiments show that Off-the-Shelf CNN features from these popular architectures can outperform traditional features, with Equal Error Rate (EER) reductions of up to ∼40%. We also show that fusion of CNN and traditional features provides additional performance improvements.

2   Periocular Recognition Using CNN Features

This section describes the basics of the CNN architectures used. Features are extracted from the different layers (Figure 1). Apart from convolutional layers, we also extract features from max pooling, ReLU, dropout, and fully connected layers. For this purpose, we employ the pre-trained models in Matlab r2018a.

AlexNet [KSH12] obtained the 1st position in ILSVRC 2012. The network achieved a breakthrough in this competition, with a top-5 error of 15.3%, more than 10.8% percentage points ahead of the runner up. It has 25 layers (including 5 convolutional layers).

GoogLeNet/Inception v1 [Sz15] is the winner of ILSVRC 2014 with a top-5 error of 6.7%. Its novelty is the use of inception modules. These employ convolution filters of different sizes in the same layer, allowing multi-level feature extraction. To reduce the number of parameters, convolutions of 1×1 are first applied. Improvements were added in later versions by redesigning the filter arrangement in the inception module, and by employing more inception modules [Sz16, Sz17]. Here, we use GoogLeNet/Inception v1, with 144 layers (including 22 convolutional layers).

ResNet [He16] introduced the concept of residual connections to ease the training of CNNs. By reducing the number of training parameters, they can be substantially deeper. The key idea is to make available the input of a lower layer to a higher layer, bypassing intermediate ones. With an ensemble of residual networks, they won ILSVRC 2015, with

a top-5 error of 3.57%. Here, we employ ResNet50 and ResNet101, having 177 and 347 layers (including 50 and 101 convolutional layers respectively).

VGG and VGG-Face. Runner-up of ILSVRC 2014, VGG [SZ14] is a deeper network (in comparison to AlexNet in 2012) in which complexity is kept tractable by using very small 3×3 convolutional filters sequentially that emulate larger ones. The most popular versions are VGG16 and VGG19 (with 16 and 19 convolutional layers respectively). Here, we employ VGG16, which has 41 layers in total. They later presented VGG-Face [PVZ15], based on VGG16, evaluated with ~1 million images from the Labeled Faces in the Wild [Hu07] and YouTube Faces [WHM11] datasets.

3   Experimental Framework

3.1   Baseline Systems

As baseline descriptors, we employ the most widely used features in periocular research [AFB16]: LBP, HOG, and SIFT. In **HOG** and **LBP**, the image is decomposed into 8×8 non-overlapped regions. Then, HOG and LBP features are extracted from each block. In HOG, the histogram of gradient orientations is built, with each bin accumulating corresponding gradient magnitudes. In LBP, each pixel is assigned a 8-bits label by thresholding the intensity with each pixel in the 3×3 neighborhood. The binary labels are then converted into decimal values and accumulated into an histogram. Both HOG and LBP are quantized into 8 bins histogram per block. Histograms from each block are then normalized to account for local illumination and contrast variations, and concatenated into a single descriptor of the whole image. Authentication with HOG and LBP is done by simple distance measures. Euclidean distance is usually used [Pa11], but it has been observed that the χ^2 distance gives better results with normalized histograms, therefore we will use it in this work. Regarding the **SIFT** descriptor, key-points extraction is done first using difference of Gaussians functions in the scale space. A feature vector of $4 \times 4 \times 8 = 128$ is then obtained by computing 8-bin gradient orientation histograms (relative to the dominant orientation to achieve rotation invariance) in 4×4 sub-regions around the key-point. The recognition metric is the number of paired key-points between two images. We employ a free implementation of the SIFT algorithm[4], with the adaptations described in [Al09]. Particularly, it includes a post-processing step to remove spurious pairings by imposing additional constraints to the angle and distance of paired key-points.

3.2   Database and Protocol

We employ the UBIPr periocular database [PP12]. It was acquired with a CANON EOS 5D digital camera in 2 sessions, with distance, illumination, pose and occlusion variability. The distance varies between 4-8m in steps of 1m, with resolution from 501×401 pixels (8m) to 1001×801 (4m). For our experiments, we select 1,718 frontal-view images from

---

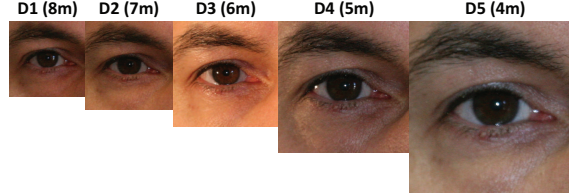[4] http://vision.ucla.edu/ vedaldi/code/sift/assets/sift/index.html

Fig. 2: Images from the UBIPr database (relative scale difference between images is shown).

86 individuals (corresponding to users with 2 sessions), having 86×2=172 available eyes. Two images are available per eye and per distance, resulting in 172×2=344 images per distance. All images have been annotated manually, so radius and center of the pupil and sclera circles are available. Images of each distance group are resized via bicubic interpolation to have the same sclera radius (we choose the average sclera radius R_s of each distance group, given by the ground truth). We use the sclera for normalization since it is not affected by dilation (as it is the case of the pupil). Then, images are aligned by extracting a square region of $7.6R_s \times 7.6R_s$ around the sclera center. This size is set empirically to ensure that all images have sufficient margin to the four sides. Further, images are converted to gray-scale, and contrast is enhanced by Contrast-Limited Adaptive Histogram Equalization (CLAHE) [Zu94] to compensate variability in local illumination. We employ CLAHE since it is usually the preprocessing choice with images of the eye region [RU10]. An example of normalized images is shown in Figure 2. Additionally, we position a mask on the whole iris region, in line with [Pa11]. This is to ensure that periocular recognition is not achieved by using information from captured high quality iris patterns [Se16]. Finally, we resized the images to match the input size of each CNN, and subtract the pixel-wise mean image of the whole database to improve recognition.

We carry out verification experiments, with each eye considered a different user. For genuine trials, we compare all images of a user among them, avoiding symmetric comparisons. This results in 7,722 user scores. Concerning impostor experiments, the first image of a user is used as enrolment image, and it is compared with the second image of the remaining users, resulting in 172×171=29,412 scores. Fusion experiments are also done between different comparators. We use linear logistic regression fusion. Given N comparators which output the scores $(s_{1j}, s_{2j}, ... s_{Nj})$ for an input trial j, a linear fusion of these scores is: $f_j = a_0 + a_1 \cdot s_{1j} + a_2 \cdot s_{2j} + ... + a_N \cdot s_{Nj}$. The weights $a_0, a_1, ... a_N$ are trained via logistic regression [Al08] using two-fold cross validation. We use this trained fusion approach because it has shown better performance than simple fusion rules (like the mean or the sum rule) [Al08]. Nonetheless, this is a weighted sum rule, though the coefficients are optimized by a specific rule [Bi97].

3.3   Results

Normalized periocular images are fed into the feature extraction pipeline of each CNN. To investigate the representation capability of each layer, we report the corresponding recognition accuracy using features from each layer. The extracted CNN feature vectors are
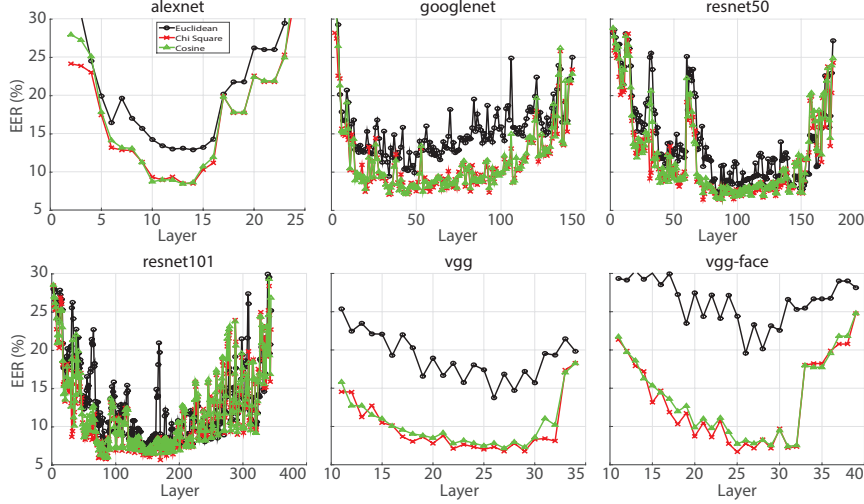
Fig. 3: Verification results (EER) of different CNN layers.

compared by simple distance or similarity measures. In particular, we evaluate the Euclidean distance, the χ^2 distance, and the cosine similarity. The EER is given in Figure 3. As it can be observed, the χ^2 distance and the cosine similarity consistently provide better results than the Euclidean distance. This is an indication that the norms of the feature vectors are not informative, because cosine similarity and Euclidean distance are mathematically equivalent if vectors have the (Euclidean) norm 1. It is also worth noting that EER has the minimum in some intermediate layers for all CNNs. The lowest EER is obtained at (refer to Figure 3): layer 13 with cosine similarity (AlexNet), layer 34 with χ^2 distance (GoogLeNet), layer 73 with χ^2 distance (ResNet50), layer 170 with χ^2 distance (ResNet101), layer 27 with χ^2 distance (VGG16), and layer 25 with χ^2 distance (VGG-Face). The more intricate architectures of GoogLeNet (with inception layers which act as small networks inside a larger network) and Resnet (which skips connections to bypass nodes) may explain the oscillations observed between consecutive layers. The simpler architecture of AlexNet and VGG (which consists of single layers stacked one after the other) results in a smoother behavior. Also, according to [Ng18], the inception modules in GoogLeNet are able to quickly converge to the minimum, as can be seen in Figure 3. ResNet, on the other hand, results in a minimum at a deeper depth (specially ResNet50) due to skipping connections between layers. AlexNet and VGG also converge at deeper depths. AlexNet has its minimum around layer 13, which in the Matlab implementation employed is after the 4th convolutional layer (out of 5 convolutional layers). VGG networks have the minimum in layers 25-27, which is after the 11th convolutional layer (out of 16 convolutional layers). Since VGG filters are smaller, it is expected that it takes more convolutional layers than other networks until the optimum performance is achieved.

In the remainder of this paper, we employ the best configuration of each CNN mentioned above. DET curves and performance figures of those cases are shown in Figure 4 (left) and Table 1 (center). We also provide the performance of LBP, HOG and SIFT descriptors,

and the EER of the SIFT+LBP+HOG combination from a previous study using the same database with only frontal images [PP12]. It should be noted that another setup of the matchers is used in that study which may explain the different results, including different ROI selection, a different fusion scheme and the use of Euclidean distance with LBP and HOG. As it can be observed, the CNNs outperform the EER of the baseline features in all cases, as well as the FRR in most cases. The best results are given by the ResNet architectures, which also happens to be the deepest and the best performing ones at the ILSVRC challenges. The best CNN is ResNet101, with an EER of 5.6% and a FRR of 14% at FAR=1%. This is about 40% less than the performance of the baseline features. GoogLeNet and VGG also provide good FRR performance. AlexNet, on the contrary, is the worst performing CNN, which may be attributed to its simpler and shallower structure. It is also interesting to note that VGG-Face does not perform as well as VGG or other CNNs, despite being tuned to work with face images.

To study complementarity between CNN and baseline features, we also carry out fusion experiments, with results given in Figure 4 (right) and Table 1 (right). The performance in all cases is further improved w.r.t. the individual systems. The best combination involves (again) ResNet101, with an EER of 5.1% and a FRR of 11.3%. This is only a 8% improvement in the EER, but nearly a 20% improvement in the FRR. This is interesting for high security applications, since operating points are usually defined at small FAR values.

baseline			CNNs			baseline+CNNs		
system	**EER**	**FRR**	**system**	**EER**	**FRR**	**system**	**EER**	**FRR**
LBP	17.8%	48.0%	alexnet	8.4%	26.1%	alexnet	6.9%	17.5%
HOG	11.3%	31.2%	googlenet	7.0%	15.9%	googlenet	6.0%	13.3%
SIFT	16.6%	30.1%	resnet50	6.4%	15.4%	resnet50	5.8%	12.7%
all	9.1%	21.8%	resnet101	**5.6%**	**14%**	resnet101	**5.1%**	**11.3%**
all*	16.0%	n/a	vgg16	6.8%	16.6%	vgg16	6.1%	13.2%
*According to [PP12]			vgg-face	6.7%	17.9%	vgg-face	5.5%	13.5%

Tab. 1: Verification results of the baseline systems (left), of the best layer of each CNN (center), and of the fusion experiments (right). FRR is given at FAR=1%.

4   Conclusions

This paper deals with the task of periocular recognition using off-the-shelf CNN architectures, pre-trained in the context of the ImageNet Large Scale Visual Recognition Challenge. Originally trained for generic object recognition, they have also proven to be very successful for many other vision tasks [Ra14]. In this paper, we test the popular AlexNet [KSH12], GoogLeNet/Inception [Sz15], ResNet [He16], and VGG [SZ14, PVZ15] architectures. Features are extracted from different layers, with comparison between feature vectors done by simple distance or similarity measures. They are evaluated with a database of frontal periocular images taken with a digital camera. Our experiments show that these popular architectures can outperform traditional features encompassing Local Binary Patterns, Histogram of Oriented Gradients, and Scale-Invariant Feature Descriptors
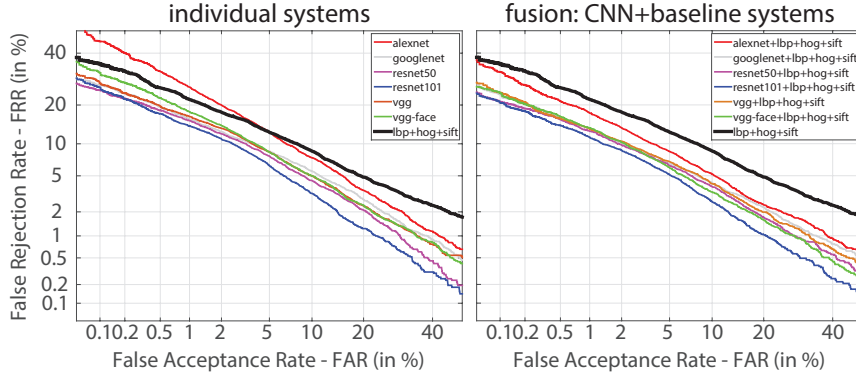
Fig. 4: Verification results (DET curves). Left: individual systems (baseline systems and best layer of each CNN). Right: fusion experiments (CNN+baseline systems).

[Pa11], with EER reductions of up to ∼40%. This shows the effectiveness of those generic objects classifiers to the periocular recognition problem as well.

Future work includes extending our framework to another periocular databases, both with visible and near-infrared illumination, and to non-frontal images. Another approach to cope with the lack of large databases in the periocular domain is to fine-tune off-the-shelf pre-trained architectures as the ones employed here. A common approach to reduce complexity is to keep initial layers fixed and only fine-tune some high-level layers, because earlier layers usually contain generic features useful to many related tasks, but later layers become progressively more specific to the problem at hand. Exploiting features from video frames to improve performance is another avenue, thanks to architectures capable to cope with dynamic information, such as Recurrent Neural Networks.

References

[AFB16]   Alonso-Fernandez, F.; Bigun, J.: A survey on periocular biometrics research. Pattern Recognition Letters, 82:92–105, 2016.

[Al08]   Alonso-Fernandez, F. *et al.*: Dealing With Sensor Interoperability in Multi-biometrics: UPM Experience at the Biosecure Multimodal Evaluation 2007. Proc SPIE BTHI, 2008.

[Al09]   Alonso-Fernandez, F. *et al.*: Iris Recognition Based on SIFT Features. Proc BIDS, 2009.

[AMB15]   Alonso-Fernandez, F.; Mikaelyan, A.; Bigun, J.: Comparison and Fusion of Multiple Iris and Periocular Matchers Using Near-Infrared and Visible Images. Proc IWBF, 2015.

[Bi97]   Bigun, E.S.; Bigun, J.; Duc, B.; Fischer, S.: Expert Conciliation for Multi Modal Person Authentication Systems by Bayesian Statistics. Proc AVBPA, 1997.

[BK17]   Bhanu, B.; Kumar, A., eds. Deep Learning for Biometrics. Springer, 2017.

[DT05]   Dalal, N.; Triggs, B.: Histogr of oriented gradients for human detect. Proc CVPR 2005.

[He16]   He, K. *et al.*: Deep Residual Learning for Image Recognition. Proc CVPR, 2016.

[Hu07]   Huang, G. B. *et al.*: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. TR 07-49, Univ Massachusetts, Amherst, 2007.

[JR12]   Jillela, R.; Ross, A.: Mitigating effects of plastic surgery: Fusing face and ocular biometrics. Proc BTAS, 2012.

[Ju11]   Juefei-Xu, F.; Luu, K.; Savvides, M.; Bui, T.; Suen, C.: Investigating age invariant face recognition based on periocular biometrics. Proc IJCB, 2011.

[KSH12]   Krizhevsky, A. *et al.*: ImageNet Classification with Deep Convolutional Neural Networks. Proc NIPS, 2012.

[Li15]   Li, H. *et al.*: A convolut neural network cascade for face detection. Proc CVPR, 2015.

[Lo04]   Lowe, D.: Distinctive image features from scale-invariant key points. IJCV, 60(2), 2004.

[MRA14]   Mahalingam, G.; Ricanek, K.; Albert, A.M.: Investigating the Periocular-Based Face Recognition Across Gender Transformation. IEEE TIFS, 9(12):2180–2192, Dec 2014.

[Ng18]   Nguyen, K.; Fookes, C.; Ross, A.; Sridharan, S.: Iris Recognition With Off-the-Shelf CNN Features: A Deep Learning Perspective. IEEE Access, 6:18848–18855, 2018.

[OPM02]   Ojala, T.; Pietikainen, M.; Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TIFS, 24(7):971–987, 2002.

[Pa11]   Park, U.; Jillela, R. R.; Ross, A.; Jain, A. K.: Periocular Biometrics in the Visible Spectrum. IEEE TIFS, 6(1):96–106, 2011.

[PP12]   Padole, C.N.; Proenca, H.: Periocular recognition: Analysis of performance degradation factors. Proc ICB, 2012.

[PVZ15]   Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep Face Recognition. Proc BMVC, 2015.

[Ra14]   Razavian, A. S.; Azizpour, H.; Sullivan, J.; Carlsson, S.: CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. Proc CVPRW, 2014.

[RU10]   Rathgeb, C.; Uhl, A.: Secure Iris Recogn Based on Local Intensity Variat. Proc ICIAR, 2010.

[SBK15]   Smereka, J. M.; Boddeti, V. N.; Kumar, B. V. K. Vijaya: Probabilistic Deformation Models for Challenging Periocular Image Verification. IEEE TIFS, 2015.

[Se16]   Sequeira, A. *et al.*: Cross-Eyed - Cross-Spectral Iris/Periocular Recognition Database and Competition. Proc BIOSIG, 2016.

[SZ14]   Simonyan, K.; Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556, 2014.

[Sz15]   Szegedy, C.; Liu, Wei; Jia, Yangqing; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.: Going deeper with convolutions. Proc CVPR, 2015.

[Sz16]   Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. Proc CVPR, 2016.

[Sz17]   Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Wojna, Z.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. Proc AAAI CAI, 2017.

[WHM11]   Wolf, L.; Hassner, T.; Maoz, I.: Face recognition in unconstrained videos with matched background similarity. Proc CVPR, 2011.

[Zu94]   Zuiderveld, Karel: Graphics Gems IV. pp. 474–485, 1994.