Pattern Recognition Letters 000 (2017) 1–11

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Deep periocular representation aiming video surveillance

Eduardo Luz[a], Gladston Moreira[a], Luiz Antonio Zanlorensi Junior[b], David Menotti[a,b,*]

[a] Universidade Federal de Ouro Preto, Computing Department, Ouro Preto, MG, Brazil
[b] Universidade Federal do Paraná, Department of Informatics, Curitiba, PR, Brazil

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Deep learning
Transfer learning
VGG Periocular region
Video surveillance
UBIRIS.v2
MobBIO

ABSTRACT

Usually, in the deep learning community, it is claimed that generalized representations that yielding outstanding performance / effectiveness require a huge amount of data for learning, which directly affect biometric applications. However, recent works combining transfer learning from other domains have surmounted such data application constraints designing interesting and promising deep learning approaches in diverse scenarios where data is not so abundant. In this direction, a biometric system for the periocular region based on deep learning approach is designed and applied on two non-cooperative ocular databases. Impressive representation discrimination is achieved with transfer learning from the facial domain (a deep convolutional network, called VGG) and fine tuning in the specific periocular region domain. With this design, our proposal surmounts previous state-of-the-art results on NICE (mean decidability of 3.47 against 2.57) and MobBio (equal error rate of 5.42% against 8.73%) competition databases.

1. Introduction

Motivated by security reasons, surveillance equipment is rising around the globe. Major cities have thousands or even millions of security cameras spread over the streets, which makes it impossible to analyze all video content manually [1]. In this context, automated surveillance systems are mandatory to assist security agents in locating events of interest in real time.

Automated surveillance systems should be able to analyze a scene, detect suspicious activity, and desirably identify the involved individuals. Several tasks have to be integrated into the system to accomplish that, such as person detection and tracking, person re-identification, action recognition and finally person recognition/identification. This work is especially interested in the last stage of surveillance systems, i.e., the biometric stage.

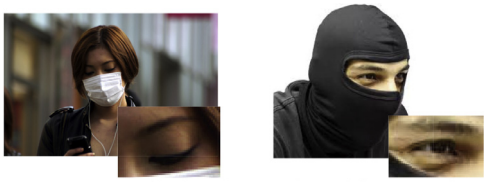
According to [1], no automated surveillance system has yet been able to perform reliable biometric recognition. However, we believe that this reality can change with the usage of deep learning on surveillance systems or at least diminish this room in the surveillance scenario. Deep learning has achieved impressive results in face recognition task including on unconstrained databases [2–8], a.k.a *in the wild*, considered as the closest to images acquired by surveillance security systems. The volume of images and videos available on the Internet favours creation of large and representative databases [6,7,9–11] and these databases allowed the advance of several feature extraction and learning representation techniques and lately those based on convolutional networks (CNN) that today represent the state-of-the-art for the face recognition problem.

Another factor that could add robustness to biometric techniques and make them more reliable to surveillance systems is to consider multimodalities [12–16]. In addition to face recognition, surveillance systems can benefit from other modalities present within the face itself, such as the ocular region (See Fig. 1). The iris, for example, is considered as the most reliable and accurate biometric trait and it is stable along aging of individuals [17]. The periocular region, which includes the iris as well, has also been the subject of recent studies [18–23] (See Fig. 2). Both iris and periocular region can be used together with the face to aid subject recognition. However, robust feature extraction/representation methods should be considered for those modalities, such as found for face modality.

To the best of our knowledge, except for the face modality, few investigations have successfully used deep learning to represent other biometric modalities [25]. For the iris modality, there are two works, DeepIris [26], DeepIrisNet [27], and both did their investigation on controlled and on near infrared spectrum (NIR) databases. For the periocular modality, there is one approach based on Semantics-Assisted Convolutional Neural Network [28] but this particular work does not overcome the current state-of-the-art [29].

* Corresponding author.
E-mail addresses: eduluz@iceb.ufop.br (E. Luz), gladston@iceb.ufop.br (G. Moreira), lazjunior@inf.ufpr.br (L.A. Zanlorensi Junior), menotti@inf.ufpr.br (D. Menotti).

https://doi.org/10.1016/j.patrec.2017.12.009
0167-8655/



(a) Woman on medical mask. (b) Soldier on mask.

Fig. 1. Scenarios where face recognition could fail.

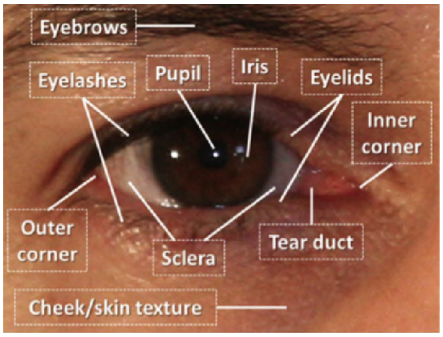


Fig. 2. Periocular region and its features. Source: [24].

Ghosh claims that the lack of large databases could be one major issue to the usage of deep learning on multimodal systems [25], since deep learning demands a large volume of data for training process [2,5]. It is known that deep network architectures often present poor performance when trained in small/reduced databases due to overfitting and techniques such as transfer learning and data augmentation [30] were proposed to help overcome this problem. In this direction, studies in the last few years have shown outstanding results using transfer learning and convolutional networks in several computational vision tasks [31,32] on reduced database. The aim of this work is to investigate deep learning for the biometric task in a modality where databases are restricted in size, such is the case of the periocular modality.

In the present work, we consider UBIRIS.v2 and MobBio databases since they are considered very challenging and acquired in such a manner that resembles a surveillance system, i.e., it is a non-cooperative database. The UBIRIS.v2 database, for example, contains 11,102 images from 261 individuals and it is one of the largest non-cooperative ocular databases available. The non-cooperative term implies that the photo was taken without awareness of the subject since images are captured under visible lighting and people were walking from 8 m to 4 m away from the camera.

Even knowing that the UBIRIS.v2 database is one of the largest available in the literature focusing ocular on the visible spectrum, it is still considered reduced in size for deep learning. MobBio is even more restricted concerning the number of images. Thus, our specific goal is to investigate and compare methods of transfer learning and data augmentation to train a very deep architecture, called VGG [5], for NICE and MobBio databases. We choose to investigate only the periocular modality since iris recognition often demands higher resolution [1] which hinders its use in surveillance systems.

Our experiments show that transfer learning [30,33] from a pre-trained CNN models can bring the advancement achieved in the face recognition problem to periocular recognition, which can favor the development of a multimodal system for images taken in-the-wild. We also show that the use of transfer learning and CNN outperforms state-of-the-art results on NICE and MobBio competition database, even considering only the periocular region. Furthermore, we explore modifications in the architecture of the pre-trained model aiming to improve performance and reduce computational cost for enrolling and test phase.

Fig. 3. UBIRIS.v2 images database.

This work is organized as follows. In Section 2, we present the non-cooperative database considered for evaluating the proposed method. A review of state-of-the-art methods aiming non-cooperative ocular recognition are described in Sections 3 and 4 presents the proposed methodology on transfer learning. In Section 5, we show our experimental setup and perform experiments comparing our method with benchmark methods in the literature. Finally, conclusions are presented in Section 6.

2. Periocular non-cooperative databases

A survey on biometric recognition in surveillance scenarios is presented in [1] and according to the authors, biometric systems must be robust and reliable in unconstrained and challenging scenarios to become useful for security systems. Thus, non-cooperative and in the wild databases must be chosen for the analysis of feature extraction and representation learning methods. In this section, we describe the two non-cooperative databases considered in the present work since they are acquired in such a manner that resembles a surveillance system.

2.1. NICE and UBIRIS.v2

The Noisy Iris Challenge Evaluation (NICE) database came from version 2 of the University of Beira Interior Iris database (UBIRIS.v2) [34] and then they follow the same pattern of image acquisition The UBIRIS.v2 database [34] has 11.102 images from 261 subjects and images were acquired to mimic an uncontrolled scenario, at different distances, angles, and lighting to simulate real noise conditions. The images have 400×300 resolution, 72 dpi and 24-bit color. Examples of database images are shown in Fig. 3. The image resolution of the NICE.II images are higher than that of UBIRIS.v2.

The NICE was the first competition created specifically to investigate the impact of ocular images acquired in uncontrolled environments aiming subject recognition and it was attended by 67 participants from 30 countries. The training set consists of 1000 images from 171 subjects poorly distributed, and another selection of 1000 images was reserved exclusively for the official evaluation. Nevertheless, for the test set those 1000 images came from 150 subjects.

The official competition metric was decidability (d), which measures how well intra-class (genuine) and inter-class (impostor) distribution scores are distant from each other [35]. The *decidability* index can be defined as follows

$$d = \frac{|\mu_E - \mu_I|}{\sqrt{\frac{1}{2}(\sigma_I^2 + \sigma_E^2)}} \quad (1)$$

where μ_I and μ_E are means and σ_I and σ_E stand for standard deviations of intra-class and inter-class distributions, respectively.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

Fig. 4. Mobbio: Face and eye images.

2.2. MobBio database

The Multimodal Database Captured with a Portable Handheld Device (MobBIO) [36] composed the first multimodal competition (2013) in which the database was fully built with a mobile device (ASUS tablet). The modalities are face, eye and voice. For each of the 105 volunteers, 16 eye images (8 per eye) and 16 face images were captured in two different lightening conditions, varying eye orientations and occlusion levels. In this work, only the iris images were used, which also include the periocular region, and were captured at a distance of approximately 10 cm. The iris images were obtained by cropping a single image containing both eyes. The resolution of each iris image is 300×200 pixels. Most of the subjects that belong to the database are of Portuguese origin, aged between 18 and 69 years, and 79% male. A selection of 406 images were made available for competition official evaluation, and 800 were previously made available for training. Examples of database images are shown in Fig. 4. The official metric used to evaluate verification mode was Equal Error Rate (EER). The database is available from the *Faculdade de Engenharia da Universidade do Porto*.[1]

3. Related works

In this section, the works considered state-of-the-art in ocular recognition for non-cooperative databases are described, especially those taking into consideration NICE, UBIRIS and MobBio databases.

Best result on NICE.II competition is achieved by the method proposed in [37], reaching a decidability of 2.57, in which authors fused ocular and iris biometrics. This method consists of four approaches for feature extraction, two on iris and two on periocular images. Feature extraction techniques for the iris are ordinal measures and color histograms. In order to extract features from the periocular region, dense Scale Invariant Feature Transforms [38] (SIFT) is used along with K-means for texton representations. Also, semantic information is extracted from the eye. Dissimilatiry score used for classification is calculated with chi-square distance metric. All the four approaches are combined with sum-rule at matching score level.

Wang et al. [39] proposed a method using only iris modality. First, the iris is normalized as proposed in [40] and partitioned into several segments. Robustness of the method comes from the partitioning scheme, which is dependent on the quality of iris segmentation. After partitioning, features are extracted with Gabor filters and later an adaptive boosting algorithm (Adaboost) is then used to select the best features and calculate similarity. A decidability of 1.82 was reported and with this result the method was awarded second place in the NICE.II competition. Note that this method is the best performing in competition, considering only iris modality.

Santos & Hoyle [41] used several techniques for feature extraction in both modalities (iris and periocular), such as SIFT, Local Binary Patterns (LBP) and texture descriptors based on wavelet transform, generating sets of different feature vectors. One distance metric is used for each feature set, such as Euclidean Distance metric, Distance-Ratio Based Scheme, Dissimilarity Using Correlation Coefficient, Dissimilarity Using Correlation Coefficients and Spatial and Frequency Analysis. The set of feature vectors is then merged by logistic regression model and the result in the NICE.II competition is a decidability of 1.78.

The work presented in [42], published after NICE.II competition, uses only normalized iris images for subject recognition. Two techniques were used to extract the features, one based on Log-Gabor, called global iris bits stabilization and another based on Zernike moments. Hamming distance metric is considered to score computation. An analysis is performed only on images reserved for training phase of the competition NICE.II, that is 1000 images from 171 subjects. Therefore images from the first 19 subjects were used for Log-Gabor parameter estimation and the remainder, 864 images associated with 151 individuals, were used for evaluation. Although, in our opinion, the presented result (decidability = 2.57) can not be directly compared with the methods that reported results in the official NICE competition test set, since they used slightly but different dataset, those authors claimed that their results are comparable to state-of-the-art methods to date.

In 2012, Proença & Alexandre [29] have merged information from iris and periocular region on matching score level for two databases on the visible spectrum. Although they have not reported results on the official NICE.II competition set of images, they used 2340 images from UBIRIS.v2 [34] which is a similar database. For evaluation, 50000 intra-class pairs and 250000 inter-class pairs were chosen at random. The result, in our opinion, is considered state-of-the-art for eye recognition in visible spectrum with an average decidability of 2.97. The techniques used to extract iris features (called Strong Biometric Traits) favors a robust representation of texture by means of a convolution of the normalized iris images against a Multi-Lobe Differential Filter (MLDF) bank. For periocular region (called Weak Biometric Trait), hand-crafted features are proposed to represent sclera color and geometry, as well as shape and texture of eyebrow.

The work in [28] is the only in the literature using deep learning for recognition in the UBIRIS.v2 database. This work introduces a methodology called semantics-assisted convolutional neural networks (SCNNs), which aims to extract and explore discriminant information from the periocular region, using a limited number of training samples. The outputs of second last layer of each CNN were concatenated in a vector and used as features. Besides, PCA was applied to reduce the dimensionality of the vector. To predict the similarity between a pair of feature vectors, the joint Bayesian scheme was utilized. The authors did not use the decidability index but reported an equal error rate (EER) of 10.98% on a subset of NICE.I data (161 subjects).

In [43], an algorithm based on retinotopic sampling grids [44] and Gabor analysis on the spectrum is proposed. Sequential forward floating selection (SFFS) technique is used to find best features/regions for periocular region. Features are also extracted from iris with 1D Log-Gabor. These authors concluded that for the MobBio database, regions closer to the skin and eyebrow have resulted in a better region to extract features. Better results are achieved with the fusion of features from the iris and periocular region. An extension of this work is presented in [45], in which results were improved with the usage of eye detection techniques. Four techniques for feature extraction from iris modality are evaluated: 1D log-Gabor filter, local intensity variations [46], coefficients of the Discrete-Cosine Transform (DCT) [47] and cumulative-sum-Based gray change analysis proposed by Ko at al.[48]. Results show that the periocular region is a better modality for uncontrolled/non-cooperative databases in visible spectrum (EER = 12.32%) and

[1] https://web.fe.up.pt/~mobbio2013/database.html.

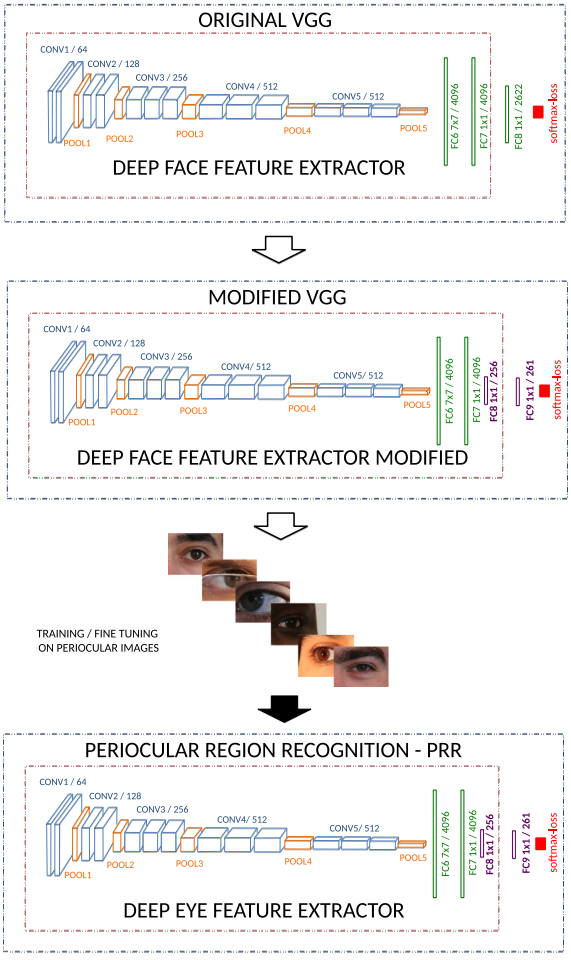


Fig. 5. Deep feature descriptor process for periocular modality.

among the iris features, Log-Gabor obtained best results for MobBio (EER = 18.81%). Fusion of features from both modalities resulted in a significant improvement (EER = 11.00%).

The state-of-the-art on ocular recognition in MobBio database, considered here, is the work presented in [49]. In that work, six feature extraction techniques are evaluated. Three techniques for iris: 1D log-Gabor filter, Discrete-Cosine Transform (DCT), SIFT and three periocular region techniques: Symmetry Assessment by Feature Expansion (SAFE) descriptors [50], log-Gabor [45] and SIFT. A different distance metric is used for each feature set, such as Hamming, Euclidean and also a customized metric based on SAFE descriptors. The best result for the periocular region is achieved by the SIFT descriptor (EER = 8.73%) and for iris modality by Log-Gabor based feature descriptor (EER = 18.81%). Fusion of all three periocular approaches and Log-Gabor for iris resulted on the best result to date (EER = 6.75%).

4. Method

In this section, we present the proposed method aiming subject recognition with deep learning on the periocular region (See Fig. 5). Firstly we present the pre-trained model and then the approach to transfer learning. Finally a modification on the deep architecture proposed in [5], called VGG, yielding our Periocular Region Recognition (PRR) architecture, aiming to improve performance is proposed.

4.1. VGG

First of all, a trained CNN architecture must be chosen for analysis, since our approach is based on transfer learning. Among best architectures for face recognition published in the literature, the state-of-the-art model proposed in [5] (called VGG) is publicly available and therefore it is considered by this work to favor reproducibility.

The VGG architecture consists of four operations: convolution, activation (ReLu), pooling and fully connected layers and all these operations are well described in the literature [33]. The network architecture is very deep, inspired by [51], and comprises a long sequence of convolutional operations. According to [51], increasing the depth of the architecture using very small filter kernels (3x3) significantly improve results on several tasks and in challenging datasets. During training, the input for the VGG is resized to 224x224 RGB image. Although in [5] all images are pre-processed by subtracting mean value, in this work this step is disregarded for the transfer learning. The convolutional stride is fixed to 1 and padding is used to preserve image size after convolution. Five max-pooling operations are employed over a 2x2 pixel window, with stride 2. Three Fully-Connected (FC) layers are proposed in [5]: the first two have 4096 channels each, the third contains 2062 channels (one for each class of the dataset proposed in [5]). Rectification (ReLU) are used after every convolutional layer and also after FC layers.

To allow the training of such a deep network (VGG), a large database with approximately 2.6 million images from 10.000 subjects was created and well described in [5].

According to [5], at first, a bootstrapping with a subset of the proposed database (containing data from 2622 subjects) is performed to solve a simple classification problem (with the softmax-log function as the last layer). After bootstrapping, the network is improved for verification mode with triplet loss as loss function [2]. For the latter process, the entire database is employed (10 thousand classes and over 2.6 mi images). Our approach starts from this pre-trained model to carry out transfer learning.

4.2. Fine tunning & transfer learning

Transfer learning process happens when a model trained for one domain or one task is used to accelerate or enhance learning in another domain/task [33]. As shown in [31], initial layers of a CNN can be used for extraction of generic features for an image representation. In this manner, the network can be pre-trained in one dataset, and re-used in another dataset with a different target task. The number of classes and distributions of the images in the source and target databases may be different, causing a problem. An approach that can be used to solve this issue is to design a network architecture that explicitly remaps class labels between two different datasets. Therefore, convolution and pooling layers are maintained and last fully connected layers responsible for classification are changed or remodeled for the new task.

According to [33], the transfer of initial layers favors problems that share same low-level features on the input images (edge filters, lighting, even geometric type changes), as well as the transfer of final layers favors problems or tasks with similarities regarding output semantics. Controlling which layer to transfer is done by freezing the learning rate of specific layers (setting it to zero). Considering different domains and the same task, for example, the transfer of all layers, except the last fully connected, are a very efficient strategy [52,53]. Therefore the transfer of any layer favors the learning of new models, especially when domains and tasks are similar [30].

In the present work, we use the pre-trained VGG model (see Table 1) without freezing any layer and let the whole model adapt

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

Table 1
Periocular region recognition (PRR) architecture, extended from VGG network, where *NF* and *NC* stand for the number of features and number and classes. Note that for the feature vector, layers 37 and 38 are not used.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	
name	n/a	$conv_1^1$	$relu_1^1$	$conv_2^1$	$relu_2^1$	pool1	$conv_1^2$	$relu_1^2$	$conv_2^2$	$relu_2^2$	pool2	$conv_1^3$	$relu_1^3$	$conv_2^3$	$relu_2^3$	$conv_3^3$	$relu_3^3$	pool3	$conv_1^4$	
support	-	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3	
filt dim	-	3	-	64	-	-	64	-	128	-	-	128	-	256	-	256	-	-	256	
num filts	-	64	-	64	-	-	128	-	128	-	-	256	-	256	-	256	-	-	512	
stride	-	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1	
pad	-	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1	
layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	*36*	*37*	38
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	*conv*	*conv*	softmax
name	$relu_1^4$	$conv_2^4$	$relu_2^4$	$conv_3^4$	$relu_3^4$	pool4	$conv_1^5$	$relu_1^5$	$conv_2^5$	$relu_2^5$	$conv_3^5$	$relu_3^5$	pool5	fc6	relu6	fc7	relu7	*fc8*	*fc9*	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1	1
filt dim	-	512	-	512	-	-	512	-	512	-	512	-	-	512	-	4096	-	4096	*NF*	-
num filts	-	512	-	512	-	-	512	-	512	-	512	-	-	4096	-	4096	-	*NF*	*NC*	-
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0

to the new problem. Thus, the transfer learning process also resembles a bootstrapping process although we have maintained transfer learning characteristics, such as low learning rates. For transfer learning, the last two layers of VGG are removed and three new layers are included: one fully connected layer for controlling the size of output feature vector, one layer to adjust the network for the new amount of training classes *NC* (each subject is a class), depending on the new domain database, and a softmax loss layer at the end, then generating our proposed Periocular Region Recognition (PRR architecture) as can be seen in Table 1.

Once the transfer is finished, i.e. the network is adapted to the new domain, the last two layers (37 and 38) are again removed and the new network output provides a feature vector of size *NF*, as represented by layer 36 of the architecture presented in Table 1. In this manner, the network can be seen as a feature extractor method and the network output as a feature, signature or representation vector.

Fig. 5 illustrates the steps for one single modality, i.e. the periocular region.

Evaluation of the proposed method follows the protocol adopted by the competitions NICE and MobBIO, i.e., biometric verification. In biometric verification mode, the system is considered an open gallery problem and the CNN is used to create representation vector for each image in the gallery. Finally, a distance metric is used to achieve the similarity scores among representation vectors. With similarity scores, the EER is determined from the Detection Error Tradeoff (DET) curve by the variation of a threshold with a resolution of $0.2x10^{-}3$.

5. Experiments

In this section, we detail the transfer learning for PRR construction and also experiments performed in the NICE and MobBio databases (Section 5.1). Biometric verification is performed in order to evaluate features extracted from PRR following established protocols of NICE.II and MobBio competitions. Thus, results achieved here can be direct compared to state-of-the-art methods in the literature. We perform a special analysis in the proposed PRR architecture (Section 5.2). We also evaluate techniques for data augmentation to train PRR from scratch (Section 5.3) and a robustness analysis of the proposed method is performed (Section 5.4). Finally, main results of this sections are briefly discussed (Section 5.5).

The computational resources used here include an Intel (R) Core i7-5820K CPU @ 3.30GHz 12 core, 64GB of DDR4 RAM and a GeForce GTX TITAN X GPU. Implementation is based on the MatConvNet toolbox [54] linked with NVIDIA CuDNN. The source code will be publicly available to allow easy reproducibility.[2]

5.1. Transfer learning

A fully trained VGG model was provided by authors [3] and we started the transfer learning from this point.

For transfer learning, the same approach proposed in [52,53] is used. Initially, no extra layer for control size of the feature vector is included in the PRR architecture. The rationale for this experiment is to evaluate the original VGG architecture on transfer learning. Therefore, the last layers of the VGG were removed and two new layers added. The new final layer is a softmax-loss layer, for simple classification, and the layer before that is a fully-connected layer which the whose number of neurons is equal to the number of class of new database (261 for UBIRIS.v2).

Two learning rates (*LR*) were used for 15 epochs, $LR = \{0.001, 0.0005\}$. We stopped the training after 5 epochs without improvement on validation error. The remainder parameters are inherited from VGG and were kept. Error curve and decay of cost function, for training on UBIRIS.v2 database, can be seen in Fig. 6.

We stress that we do not freeze the learning rate of any layer during fine-tuning, since our task is the same from original VGG (subject recognition). Our hypothesis is that both face and eye images share same low-level features (first layers of the network).

5.1.1. NICE database

As the UBIRIS.v2 database is balanced regarded number of images per subject it is more suitable for the training phase.

The results reported here are constructed with the official test set of the NICE.II competition for comparison purposes. The number of intra-class (genuine) pairs is 4.634 and inter-class pairs (imposters) is 494.866.

To evaluate the impact of the number of images during training, we performed three experiments by gradually adding images from the UBIRIS.v2 database in training step. By observing the curves in Fig. 7, one can conclude that adding more images results in better fine tuned model. The best results are achieved when all UBIRIS.v2 images are used for transfer learning. We stress that no segmentation or preprocessing is applied on the periocular image.

We also investigate the impact of distance metric for DET curve construction (see in top-right corner of Fig. 7). There is a significant difference between results on different distance metrics,

[2] http://www.decom.ufop.br/csilab.
[3] http://www.vlfeat.org/matconvnet/pretrained/.

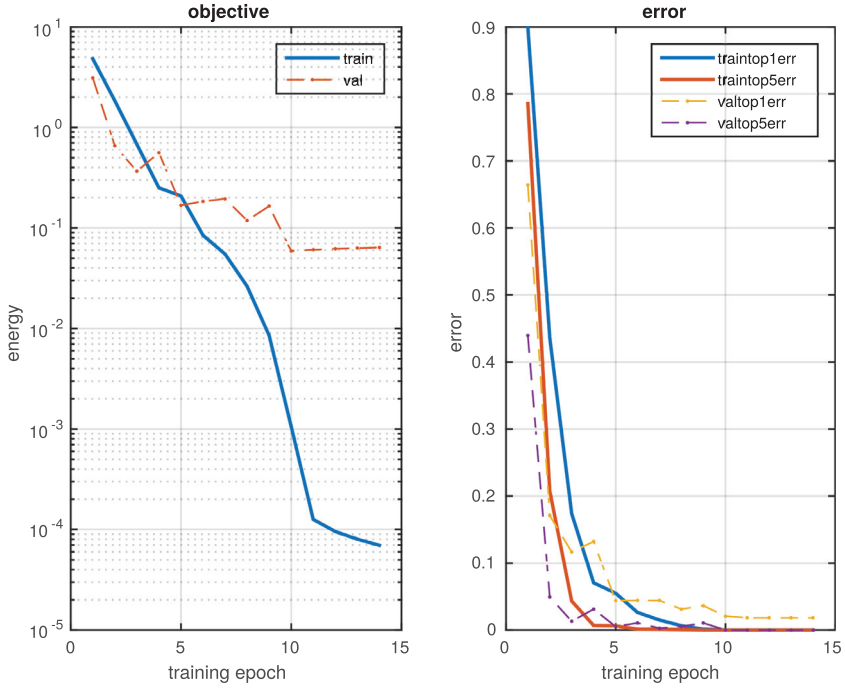


Fig. 6. Fine tuning training on periocular UBIRIS.v2 data for 14 epochs.

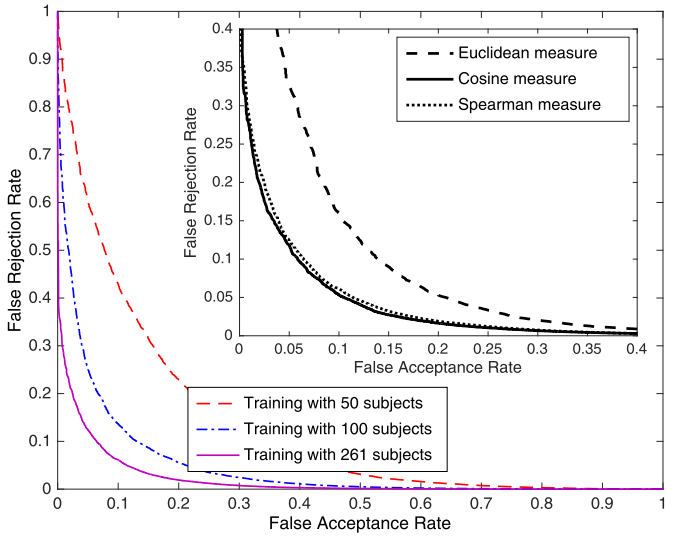


Fig. 7. DET curve for different training sizes and for different distance metrics on NICE evaluation set.

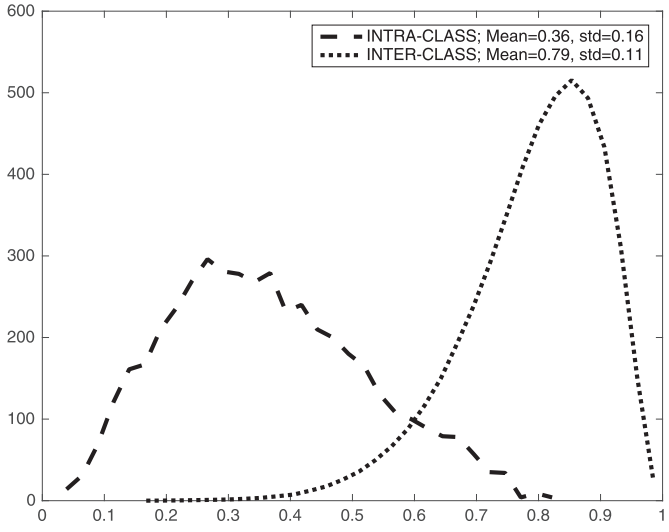


Fig. 8. Intra-class and Inter-class histogram distribution on NICE evaluation set.

which is specially explored in Section 5.2.2. Histogram distribution for PRR after fine-tune on the cosine distance metric is presented in Fig. 8. From this histogram, the decidability can be calculated.

5.1.2. Mobbio database

For this database, we follow the same protocol proposed in the MobBio multimodal recognition competition (2013), where 800 images are made available for training and 406 are reserved for official evaluation. Fine tuning is accomplished with all 800 images dedicated to training. During the evaluation, each one of the images from the test set is used as a probe against all the others in order to generate the intra-class and inter-class distributions, which resulted in 1365 intra-class (genuine) pairs and 76,845 inter-class (impostors) pairs. No segmentation is performed.

5.2. PRR analysis

In this subsection, we analyze aspects of our proposed PRR model. Initially, we evaluate the impact of adding an extra layer to the architecture, allowing feature size control. Subsequently, we investigate three distance metrics, following the protocols of NICE.II and MobBio competitions, and further analysis the computational cost in both time and space.

5.2.1. PRR Feature vector size

The impact of adding an extra layer to control feature vector size is also investigated. The new layer (layer 36) is initialized with zero mean and standard deviation of 10^{-2}, included in the architecture and then the model is re-trained/fine-tunned. A grid search is performed by varying the number of neurons in the last fully connected layer.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

Table 2
Grid search on different number of neurons in the new layer for NICE.II. (* same as VGG architecture (without layer 36)).

Feature Vector Size	Model Size (MB)	Distance metric (decidability)		
		Euclidean	Cosine	Spearman
4096*	964	2.21	3.16	2.98
512	977	2.88	3.00	3.43
256	969	2.99	2.97	3.51
128	965	2.95	2.88	3.41
64	963	2.83	2.82	3.28

Table 3
Memory requirement to store a thousand periocular image representation and time cost for the NICE test phase.

Feature Vector Size	Memory Size (MB)	Time (sec.)
4096*	15.62	3.45
512	1.95	0.42
256	0.97	0.23
128	0.48	0.14
64	0.24	0.10

Regarding the model size, by observing the data in Table 2, we can conclude that the inclusion of one fully connected layer (layer 36 in Table 1) could lead to slightly more connections between neurons and consequently a marginal increase in the final size of the model.

The first row of this table represents the original VGG, with $NF = 4096$ and the final layer for classification with NC neurons, yielding $NF \times NC$ connections. Note that there is no extra layer connected between the original feature layer (4096) and the classification layer. The remainder rows represent models with the inclusion of an extra layer (layer 36 in Table 1). In the worst case, with the inclusion of one layer with $NF = 512$, we have $4096 \times NF$ plus $NF \times NC$ connections and the increase in the model size is 1.5% in terms of MB.

Contrasting to that, the addition of one layer with $NF = 64$ reduces the number of connections.

In general, inclusion of an extra layer (layer 36 in Table 1) results in expressive gains. Besides significant performance improvement on test for the Spearman distance metric with smaller NF values, it provides reduction in computational cost during verification since the feature vector (output of layer 36 in Table 1) is also smaller (see Table 3). Notice that for the cosine distance metric, high dimensional vectors resulted in better performance.

5.2.2. Distance metric impact

One can see in Table 2 that the distance metric used has a great impact on results, especially when the representation vector has different dimensions. The Euclidean distance metric is the most popular to compute similarity although it has some limitations. Since it is calculated as the sum of squares of the differences of each vectors dimension, the magnitude of a particular dimension may deeply affected this dissimilarity [55]. The Euclidean distance metric is represented by

$$d_E(A, B) = \sqrt{\sum_{j=1}^{N} |A_j - B_j|^2} \tag{2}$$

where A and B are features vectors.

The cosine metric is the cosine of the angle between two Euclidean vectors and it is a standard metric used in information retrieval due to scalar transformation invariance [56]. The cosine distance metric, a.k.a. angular metric, calculates the normalized inner product and measures the angle between two vectors. The cosine distance metric is represented by

$$d_c(A, B) = 1 - \frac{\sum_{j=1}^{N} A_j B_j}{\sqrt{\sum_{j=1}^{N} A_j^2}\sqrt{\sum_{j=1}^{N} B_j^2}} \tag{3}$$

where A and B stand for features vectors.

The Spearman distance metric is based on Pearson coefficient and therefore is immune to linear transformations [56]. The Spearman distance metric can be defined as

$$d_S(A, B) = 1 - \frac{6 \sum_{i=1}^{n} (r_{A_i} - r_{B_i})^2}{n(n^2 - 1)} \tag{4}$$

where r_{A_i}, r_{B_i} are the rank of A_i and B_i, respectively.

From the figures in Table 2, we argue that: 1) As the cosine distance metric only measures the angle between two vectors and the vector magnitudes don't matter since it uses the normalized inner product between vectors, the cosine distance metric has shown to be more robust for **high dimensional vectors** reporting better results than the Euclidean and Spearman distance metrics. 2) For **low dimensional vectors**, the Spearman distance overcome the Euclidean and cosine distances, and the results yielded by the latter distances are similar to some extend. The Spearman distance metric is both not sensitive to linear and non-linear transformations and these properties could add robustness against object scale, small differences in shape and noise, which could explain better results with the Spearman metric over Euclidean one on low-dimensional data. 3) The best results where achieved by the Spearman metric using a 256 feature vector.

5.2.3. Testing phase

The addition of an extra layer to control the size of feature vector provides significant improvement while keeping computational cost low in terms of time and space during training (see Table 2). As in this work, we follow the protocols of the competitions NICE and MobBIO, our system resembles an open set/gallery problem. In this sense, each image is used as a probe and compared against all others. In an open gallery problem, classification is done by comparing a distance metric against a pre-defined threshold. Therefore, the computational time is directly related to the size of the feature vector and number of comparisons, which is dependent on the number of images in the gallery (1000 images for the NICE competition). The space on disk required to store data is also proportional to the size of the feature vector multiplied by the number of images. Note that in a real application, the number of images tends to increase, as more individuals can be added to the gallery. Thus, the most relevant computational cost is related to the evaluation phase. Furthermore, advantages provided by feature vector reduction can leverage the usage of the proposed method in embedded systems and mobile equipment due to less memory requirement to store an image representation and less time during the test as can be seen in Table 3.

5.3. Training VGG from scratch

In [5], the VGG was constructed for a classification problem of 2622 classes with a softmax loss probabilistic layer, and for this, it generates a feature vector of size 4096 for each instance. Before training, all images were resized to 224×224 pixels and face images were centered.

According to [5], the weight of filters was randomly initialized with zero mean Gaussian distribution and standard deviation of 10^{-2}. For weights optimization stochastic gradient descent was used on mini-batches of size 64 and momentum coefficient of 0.9. In [5] authors also used 50% dropout after two fully connected layers and weight decay with coefficients value of 5×10^{-4}. Initial learning rate was 10^{-2}, decreasing by a factor of 10 when accuracy

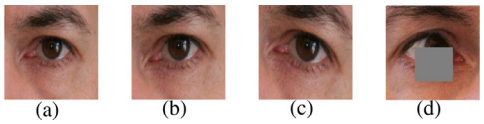


Fig. 9. Data augmentation examples for training VGG from scratch. a) Original image; b) Rotate−5 degree; c) Rotate +10 degree; d) image from other eye with random noise.

stagnates on validation. In fact, there were three learning rates according to the authors. For the triplet-loss, in [5] the learning rate of entire network was frozen except for the last fully connected layer. The triplet-loss process was carried out in 10 epochs at a learning rate of 0.25.

For comparison, the same VGG architecture is trained from scratch, with randomly initialized weights (zero mean and standard deviation 10^{-2}). Biases were initialized to zero. Since all attempts to train the full network fails, an approach similar to that proposed in [5] are followed, in which the VGG are trained in parts. First, a small network comprising the operations from layer 1 to layer 10 of Table 1 is used for a simple classification problem, with an addition of one FC layer and learning rates of [0.01, 0.001] by 30 epochs. Subsequently, the last FC layer was removed, and operations related to layer 11 to layer 17 are appended to the architecture along with a new FC layer. The new network is trained with learning rates of [0.001, 0.0001] for 20 epochs. The same process was done to include the layers from 18 to 24 and 25 to 38. Therefore, the final VGG model is trained in four steps. This process is necessary due to the instability of the gradient in deep networks when the weights are randomly initialized [51]. The model is regularized using a dropout of 50% and a weight decay of $5 * 10^{-4}$. However, it was not possible to effectively train the network without data augmentation.

5.3.1. Data augmentation

For the data augmentation process, new images are created by translating ([+10%, −10%] in pixels), rotating and cropping ([+5, −5, +10, −10] in degrees) images. Also, random noise is inserted as illustrated in Fig. 9. Three data augmentation techniques are applied randomly for each eye image. Although this process improved the training process, another data augmentation technique is evaluated.

5.3.2. Generative adversarial networks (GAN) based data augmentation

The GAN based data augmentation process was of paramount importance for successfully training the network from scratch (see Fig. 10). The GANs comprises of two networks: a generator and a discriminator [57]. The discriminator classifies whether a sample is fake or real, while the generator produces samples to cheat the discriminator. GANs have show potential to generate synthetic images on different domains [58–60]. To create synthetic eye images, the Deep Convolutional Generative Adversarial Networks (DC-GAN) are chosen due to more stability during training [58].

For the generator architecture, we follow the configuration proposed in [59], with 100 dimension vector as input transformed to 4x4x16 tensor by a linear function. Five deconvolutional layers with kernel size of 5x5 and stride 2 are applied to the input tensor, followed by a deconvolution layer with size of 5x5 and stride 1. Rectifier operation and batch normalization are used after every deconvolution operation. An image of size 128x128x3 is generated as output.

Generator outputs (fake images) and real images are then used as input to discriminator network. The discriminator network comprises of 5 convolutional layers with a 5x5 filter size and stride 2.

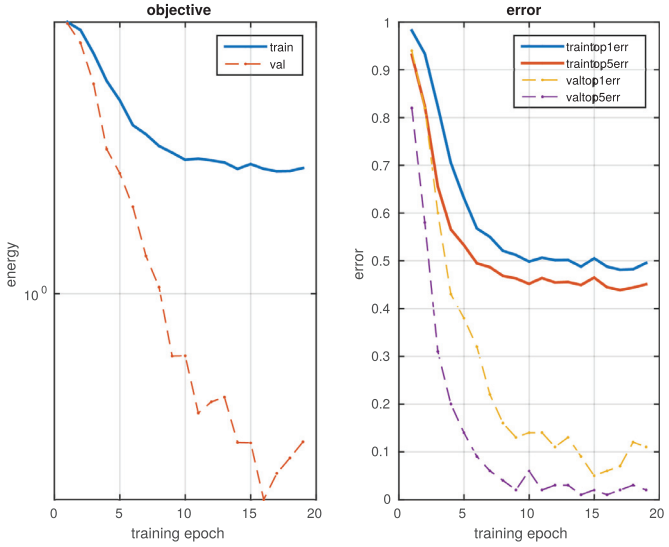


Fig. 10. Training VGG from scratch on periocular UBIRIS.v2 data - 64 new imgs/class generated with DCGAN.

A fully-connected layer is used for binary classification (real image/fake image).

Since a large number images from different classes are needed in order to successfully train a GAN model, first we have trained a generic DC-GAN on 18,052 images from 5 datasets (MICHE [61], UBIRIS.v2, MobBio, CSIP [62], CrossSpectrumDB [63]) for 130 epochs (see Fig. 11-a). Finally, the generic GAN are fine tuned during 50 epochs for each subject on UBIRIS.V2 and MobBIO. Then, the fine tunned GAN are used to create up to 128 new images for each subject of the databases (see Fig. 11-b and c). Results on different data augmentation techniques are presented in Table 4. Due to stochastic nature of the CNNs the experiments in Table 4 are performed 15 times and the values reported stand for the mean values of the metrics obtained.

The data augmentation process with GANs has shown to be promising, allowing a considerable reduction of error during training. According to Table 4, it is not possible to determine with statistical significance the amount of synthetic images that yields the best result. However, even with the addition of new synthetic images for each individual (both datasets), the result with VGG from scratch is still worse than the proposed transfer learning approach. Besides, training VGG from scratch, along with the GAN based data augmentation process has a high computational cost (training time is about 30 h per model). Contrasting to that, the transfer learning approach is simple and straightforward, requires little computational resources and few parameter adjustment, which favors the reproducibility of the results as well as the extension of the proposed approach to other biometric modalities (training time is about 4 hours per model).

5.4. Robustness analisys

Images are artificially corrupted to simulate an eye occlusion, to assess model robustness. Fig. 12 illustrates this process for an image from the NICE test set. For this scenario, the detection threshold is 0.625 which means that any score below this value indicates a pair from the same person. In Fig. 12, we observed that the inclusion of noise worsens the scores although not enough to alter the classification result.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

(a) Examples of images generated by the generic DC-GAN.

(b) Original images from class 19 of UBIRIS.V2.

(c) Syntetic images for class 19 of UBIRIS.V2.

Fig. 11. Syntetic images generated with DC-GAN.

Table 4
VGG from scratch trained with data augmentation techniques. Mean values after 15 executions.

Data Augmentation	Results on NICE		Results on MobBio	
Technique	EER	decidability	EER	decidability
rotating, translating, random noise	36.24 ± 0.95	0.64 ± 0.08	**23.27 ± 1.53**	**1.49 ± 0.12**
rotating, translating, random noise + 32 imgs/class generated with GAN	25.16 ± 2.23	1.36 ± 0.13	**19.79 ± 0.89**	**1.64 ± 0.05**
rotating, translating, random noise + 64 imgs/class generated with GAN	23.07 ± 0.69	1.52 ± 0.07	**19.83 ± 0.86**	**1.65 ± 0.05**
rotating, translating, random noise + 128 imgs/class generated with GAN	22.99 ± 1.47	1.52 ± 0.11	**19.93 ± 0.60**	**1.64 ± 0.05**

Table 5
Results summarization. FS = from scratch; (* evaluation on NICE.II training set on 161 subjects – # mean values after 15 executions).

Methods	Database	Modalities	Decidability	EER
Proença [64]	UBIRIS.v2	iris + periocular	2.97	-
Tan et al. [37]	NICE	iris + periocular	2.57	12 %
Wang et al. [39]	NICE	iris	1.82	19 %
Zhao and Kumar. [28]	NICE*	periocular	-	10.98%
Alonso-Fernandez et al. [49]	MobBio	periocular	-	8.73 %
VGG#	NICE	periocular	0.64	36.24 %
	MobBio	periocular	1.49	23.27 %
VGG-FS#	NICE	periocular	1.52	23.07 %
	MobBio	periocular	1.65	19.83 %
PRR#	NICE	periocular	3.15	7.45 %
	MobBio	periocular	3.02	7.48 %
PRR (256)#	NICE	periocular	3.47	5.92 %
	MobBio	periocular	3.53	5.42 %

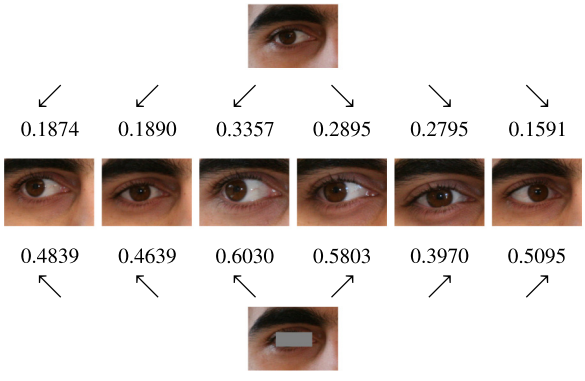


Fig. 12. Robustness analysis for NICE test set, using the PRR 256 network, the cosine distance metric, and EER threshold of 0.625. Scores before noise/after noise.

5.5. Main results and final discussion

In Table 5, the result of our method is compared against results considered by this work as the state-of-the-art for NICE and MobBio. Our approach outperformed results obtained in [37] (NICE) and [49] (MobBio) by 35% and 38% respectively. Thus, transfer between those domains resulted in state-of-the-art feature extractor for both NICE and MobBio databases. We do believe that a new state-of-the-art result would be achieved for UBIRIS.v2 as well, since NICE database is extracted from it.

We trained each different model (PRR, PRR (256), VGG with and without GAN) using the same seed to control all stochastic processes. With the mean and standard deviation figures, we establish statistical significance analysis and the results indicated that the models learned from scratch with and without samples generated by GANs are not competitive with the ones fined-tuned from VGG network.

The main advantage of the proposed method is robustness (see Fig. 13), since the image does not need any type of segmentation or region of interest location, which is desirable for in-the-wild images and video surveillance environments. Another advantage is that the method does not need very large databases for training deep CNN models, which could make it feasible to use in different modalities beyond the ocular region, such as the ear shape. It is important to emphasize that although the databases investigated in this work are considered non-cooperative and in-the-wild in the literature, their resolution is considered high for conventional surveillance systems.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

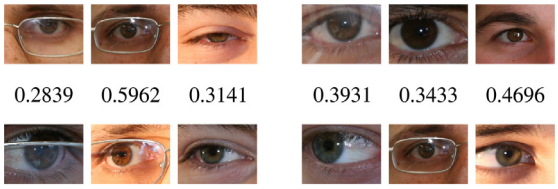


Fig. 13. Very noisy images from the NICE test set. Evaluation is done with the PRR 256 network, the cosine distance metric, and ERR threshold of 0.625. Top images are used as probe and the respective bottom one is the image recovery from the remaining data as the best match (scores in the middle row). Values lower than the established threshold (0.625) are considered as the same subject. a) Images of the same subject, correctly classified; b) Images of different subjects.

6. Conclusion

Biometrics plays an important role in surveillance systems. However, surveillance systems commonly provide low resolution images and conditions for acquisition are diverse, which means all kinds of noise and artifacts. In practice, few methods can adapt to this challenging scenarios.

Multimodality today represents a promising direction for the improvement of biometric systems, bringing robustness and performance improvement. Hence, the face recognition problem is the one that receives the most attention in the literature and one can find large, diverse and in-the-wild face databases, which favors the development of deep learning techniques for this domain.

In this work, we answer a question raised in [25] regarding the feasibility of using deep learning on biometric modality with limited databases. We showed that outstanding results obtained with face recognition and deep learning could be transferred to periocular modality and that periocular recognition can be considered a valid option for surveillance systems. Our proposal achieved new state-of-the-art results for two well-known databases/competitions, i.e., NICE and MobBio. Also, we briefly highlighted that with small modifications we can provide deeper and smaller but still powerful representations for biometrics.

Acknowledgments

The authors would like to thank UFOP, UFPR, FAPEMIG (APQ-#02825-14), CAPES and CNPq (Grant #307010/2014-7 & #428333/2016-8) for the financial support. The authors also would like to thank NVIDIA Corporation for the donation of one GPU Titan Black and two GPU Titan X. An IBM PhD Fellowship generously supports the first author.

References

[1] J. Neves, F. Narducci, S. Barra, H. Proença, Biometric recognition in surveillance scenarios: a survey, Artif. Intell. Rev. 46 (4) (2016) 515–541.
[2] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[3] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? metric learning approaches for face identification, in: Computer Vision, 2009 IEEE 12th international conference on, IEEE, 2009, pp. 498–505.
[4] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, D.A. Forsyth, Names and faces in the news, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2, IEEE, 2004, p. II.
[5] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015, pp. 1–12.
[6] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 529–534.
[7] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[8] J. Hu, J. Lu, Y.-P. Tan, Discriminative deep metric learning for face verification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1875–1882.
[9] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report, Technical Report 07–49, University of Massachusetts, Amherst, 2007.
[10] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: a joint formulation, in: European Conference on Computer Vision, Springer, 2012, pp. 566–579.
[11] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.
[12] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognit. Lett. 24 (13) (2003) 2115–2125.
[13] K.W. Bowyer, K.I. Chang, P. Yan, P.J. Flynn, E. Hansley, S. Sarkar, Multi-modal biometrics: an overview, in: Workshop on Multi-Modal User Authentication, 2006, pp. 1221–1224.
[14] S. Shekhar, V.M. Patel, N.M. Nasrabadi, R. Chellappa, Joint sparse representation for robust multimodal biometrics recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 113–126.
[15] D. Karmakar, C. Murthy, Generation of new points for training set and feature-level fusion in multimodal biometric identification, Mach. Vis. Appl. 25 (2) (2014) 477–487.
[16] H.M. Sim, H. Asmuni, R. Hassan, R.M. Othman, Multimodal biometrics: weighted score level fusion based on non-ideal iris and face images, Expert Syst. Appl. 41 (11) (2014) 5390–5404.
[17] J. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE TPAMI (1993).
[18] S. Crihalmeanu, A. Ross, Multispectral scleral patterns for ocular biometric recognition, Pattern Recognit. Lett. 33 (14) (2012) 1860–1869.
[19] C.-W. Tan, A. Kumar, Towards online iris and periocular recognition under relaxed imaging constraints, IEEE Trans. Image Process. 22 (10) (2013) 3751–3765.
[20] F. Juefei-Xu, K. Luu, M. Savvides, T.D. Bui, C.Y. Suen, Investigating age invariant face recognition based on periocular biometrics, in: Biometrics (IJCB), 2011 International Joint Conference on, IEEE, 2011, pp. 1–7.
[21] J. Xu, M. Cha, J.L. Heyman, S. Venugopalan, R. Abiantun, M. Savvides, Robust local binary pattern feature sets for periocular biometric identification, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE, 2010, pp. 1–8.
[22] S. Bharadwaj, H.S. Bhatt, M. Vatsa, R. Singh, Periocular biometrics: When iris recognition fails, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE, 2010, pp. 1–6.
[23] K. Hollingsworth, K.W. Bowyer, P.J. Flynn, Identifying useful features for recognition in near-infrared periocular images, in: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, IEEE, 2010, pp. 1–8.
[24] F. Alonso-Fernandez, J. Bigun, Periocular biometrics: databases, algorithms and directions, in: Biometrics and Forensics (IWBF), 2016 4th International Workshop on, IEEE, 2016, pp. 1–6.
[25] S. Ghosh, Challenges in deep learning for multimodal applications, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 611–615.
[26] N. Liu, M. Zhang, H. Li, Z. Sun, T. Tan, Deepiris: learning pairwise filter bank for heterogeneous iris verification, Pattern Recognit. Lett. 82 (2016) 154–161.
[27] A. Gangwar, A. Joshi, Deepirisnet: deep iris representation with applications in iris recognition and cross-sensor iris recognition, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 2301–2305.
[28] Z. Zhao, A. Kumar, Accurate periocular recognition under less constrained environment using semantics-Assisted convolutional neural network, IEEE Trans. Inf. Forensics Secur. 12 (5) (2017) 1017–1030.
[29] H. Proença, L.A. Alexandre, Toward covert iris biometric recognition: experimental results from the nice contests, IEEE Trans. Inf. Forensics Secur. 7 (2) (2012) 798–808.
[30] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? in: Advances in neural information processing systems, 2014, pp. 3320–3328.
[31] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717–1724.
[32] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition., in: Icml, 32, 2014, pp. 647–655.
[33] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, 2016. MIT Press. http://www.deeplearningbook.org
[34] H. Proença, S. Filipe, R. Santos, J. Oliveira, L.A. Alexandre, The ubiris. v2: a database of visible wavelength iris images captured on-the-move and at-a-distance, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1529.
[35] J.G. Daugman, G.O. Williams, A proposed standard for biometric decidability, in: Proc. CardTech/SecureTech Conference, 1996, pp. 223–234.
[36] A.F. Sequeira, J.C. Monteiro, A. Rebelo, H.P. Oliveira, Mobbio: a multimodal database captured with a portable handheld device, in: Computer Vision Theory and Applications (VISAPP), 2014 International Conference on, 3, IEEE, 2014, pp. 133–139.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009

[37] T. Tan, X. Zhang, Z. Sun, H. Zhang, Noisy iris image matching by using multiple cues, Pattern Recognit. Lett. 33 (8) (2012) 970–977.
[38] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[39] Q. Wang, X. Zhang, M. Li, X. Dong, Q. Zhou, Y. Yin, Adaboost and multi-orientation 2d Gabor-based noisy iris recognition, Pattern Recognit. Lett. 33 (8) (2012) 978–983.
[40] J.G. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE Trans. Pattern Anal. Mach. Intell. 15 (11) (1993) 1148–1161.
[41] G. Santos, E. Hoyle, A fusion approach to unconstrained iris recognition, Pattern Recognit. Lett. 33 (8) (2012) 984–990.
[42] C.-W. Tan, A. Kumar, Accurate iris recognition at a distance using stabilized iris encoding and zernike moments phase features, IEEE Trans. Image Process. 23 (9) (2014) 3962–3974.
[43] F. Alonso-Fernandez, J. Bigun, Best regions for periocular recognition with NIR and visible images, in: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 4987–4991.
[44] F. Alonso-Fernandez, J. Bigun, Periocular recognition using retinotopic sampling and Gabor decomposition, in: European Conference on Computer Vision, Springer, 2012, pp. 309–318.
[45] F. Alonso-Fernandez, J. Bigun, Near-infrared and visible-light periocular recognition with Gabor features using frequency-adaptive automatic eye detection, IET Biom. 4 (2) (2015) 74–89.
[46] C. Rathgeb, A. Uhl, Secure iris recognition based on local intensity variations, in: International Conference Image Analysis and Recognition, Springer, 2010, pp. 266–275.
[47] D.M. Monro, S. Rakshit, D. Zhang, Dct-based iris recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007).
[48] J.-G. Ko, Y.-H. Gil, J.-H. Yoo, K.-I. Chung, A novel and efficient feature extraction method for iris recognition, ETRI J. 29 (3) (2007) 399–401.
[49] F. Alonso-Fernandez, A. Mikaelyan, J. Bigun, Comparison and fusion of multiple iris and periocular matchers using near-infrared and visible images, in: Biometrics and Forensics (IWBF), 2015 International Workshop on, IEEE, 2015, pp. 1–6.
[50] A. Mikaelyan, F. Alonso-Fernandez, J. Bigun, Periocular recognition by detection of local symmetry patterns, in: Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on, IEEE, 2014, pp. 584–591.
[51] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[52] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.
[53] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 142–158.
[54] A. Vedaldi, K. Lenc, Matconvnet: convolutional neural networks for matlab, in: Proceedings of the 23rd ACM international conference on Multimedia, ACM, 2015, pp. 689–692.
[55] S.D. Bharkad, M. Kokare, Performance evaluation of distance metrics: application to fingerprint recognition, Int. J. Pattern Recognit. Artif. Intell. 25 (2011) 777–806.
[56] S. Van Dongen, A.J. Enright, Metric distances derived from cosine similarity and pearson and spearman correlations, arXiv preprint arXiv:1208.3145 (2012).
[57] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[58] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, arXiv preprint arXiv:1701.07717 (2017).
[59] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).
[60] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, arXiv preprint arXiv:1610.09585 (2016).
[61] M. De Marsico, M. Nappi, D. Riccio, H. Wechsler, Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols, Pattern Recognit. Lett. 57 (2015) 17–23.
[62] G. Santos, E. Grancho, M.V. Bernardo, P.T. Fiadeiro, Fusing iris and periocular information for cross-sensor recognition, Pattern Recognit. Lett. 57 (2015) 52–59.
[63] A. Sequeira, L. Chen, P. Wild, J. Ferryman, F. Alonso-Fernandez, K.B. Raja, R. Raghavendra, C. Busch, J. Bigun, Cross-eyed-cross-spectral iris/periocular recognition database and competition, in: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE, 2016, pp. 1–5.
[64] H. Proença, Ocular biometrics by score-level fusion of disparate experts, IEEE Trans. Image Process. 23 (12) (2014) 5082–5093.

Please cite this article as: E. Luz et al., Deep periocular representation aiming video surveillance, Pattern Recognition Letters (2017), https://doi.org/10.1016/j.patrec.2017.12.009