A Reminiscence of ”Mastermind”: Iris/Periocular Biometrics by ”In-Set” CNN Iterative Analysis

Hugo Proença, Senior Member, IEEE and João C. Neves, Member, IEEE

Abstract—Convolutional neural networks (CNNs) have emerged as the most popular classification models in biometrics research. Under the discriminative paradigm of pattern recognition, CNNs are used typically in one of two ways: 1) verification mode ("are samples from the same person?"); where pairs of images are provided to the network to distinguish between genuine and impostor instances; and 2) identification mode ("whom is this sample from?"), where appropriate feature representations that map images to identities are found. This paper postulates a novel mode for using CNNs in biometric identification, by learning models that answer to the question "is the query’s identity among this set?". The insight is a reminiscence of the classical Mastermind game: by iteratively analysing the network responses when multiple random samples of \( k \) gallery elements are compared to the query, we obtain weakly correlated matching scores that - altogether - provide solid cues to infer the most likely identity. In this setting, identification is regarded as a variable selection and regularization problem, with sparse linear regression techniques being used to infer the matching probability with respect to each gallery identity. As main strength, this strategy is highly robust to outlier matching scores, which are known to be a primary error source in biometric recognition. Our experiments were carried out in full versions of two well known irises near-infrared (CASIA-IrisV4-Thousand) and periocular visible wavelength (UBIRIS.v2) datasets, and confirm that recognition performance can be solidly boosted-up by the proposed algorithm, when compared to the traditional working modes of CNNs in biometrics.

Index Terms—Iris Recognition, Periocular Biometrics, Convolutional Neural Networks.

I. INTRODUCTION

IRIS biometrics is one of the most reliable human recognition technologies. Since the pioneer algorithm [6], a long road has been travelled in this domain, leading to successful applications such as borders control and ID cards. Recently, the periocular region [22] was advocated as a possibility to overcome the limitation of the iris to be used in unconstrained data acquisition conditions, being more robust to expressions than the whole face, while keeping remarkable discriminating power between humans.

CNNs have turned extremely popular in tasks such as image segmentation [16], object detection [39] and classification [15]. The property of shift invariance gives them the biological inspiration and keeps the number of parameters relatively small, making learning a feasible task. As in many other computer vision problems, various CNN-based iris/periocular recognition methods were reported in the literature (e.g., [25] and [26]).

A. Motivation

The menagerie effect [45] is well known to biometric researchers and practitioners: in most recognition systems, there are groups of subjects whose genuine/impostor score distributions are evidently different from the distributions of the general population. In practical terms, matching data from these subjects produces outlier scores that might compromise the effectiveness of the whole system. The problem is particularly evident in cases where degraded data are extracted from subjects and matched only once during the recognition process (i.e., not subjected to outlier correction), which happens in the two most typical working modes of CNNs in biometrics, illustrated in Fig. 1: A) "are samples from the same person?" (1:1 mode); or B) "whom is this sample from?" (1:N mode). In the 1:1 mode, recognition is regarded as a binary classification problem, with pairs of query/gallery samples being shown to the networks, to discriminate between genuine and impostors comparisons. In the 1:N mode, samples are presented individually to the networks, to infer either the likelihood of matching the query to identities (closed-world assumption) or to obtain a compact description of the query that is used subsequently by another classifier (open-world assumption).

B. Contributions

This paper describes a novel working mode of CNNs in biometrics that contributes to attenuate the biometric menagerie effect. The idea consists in inferring (during learning time) one CNN able to answer to the in-set question: "is the query’s identity in this set?", when looking repeatedly to the query plus samples of \( k \) (> 1) gallery elements. Taking profit of the remarkable ability of CNNs to model complex feature spaces, it is possible to obtain multiple weakly correlated responses that - altogether - provide solid cues about the most likely matching identity, resembling the rationale followed to play the classical Mastermind game1, while attenuating the effects of outlier observations.

Next, in test/identification time, the method is composed of two phases: 1) the gallery identities are iteratively sampled and provided to the CNN, together with the query. The CNN responses feed a Bayesian framework, where the most unlikely matching identities are rejected, so that at the end only the most probable identities remain; and 2) the CNN

1https://en.wikipedia.org/wiki/Mastermind_(board_game)
C. Advantages and Weaknesses

Creating CNN instances composed of \( k+1 \) samples not only increases the potential number of learning instances (having \( n \) learning samples, \( \binom{n}{k+1} \gg \binom{n}{k}, \forall k \geq 2 \)), but such inputs also offer different points-of-view to the network. In this setting, the underlying hypothesis is that looking repeatedly to multiple objects (subjects) of different kinds facilitates to recognize one particular class of object (subject). Also, as the CNN sees each query/gallery sample more than one time, and at each iteration integrated in different inputs, repeated outlier scores will be unlikely, which potentially reduces the menagerie effect.

As main weakness, the in-set analysis is expected to fail (i.e., produce a false identification) when the mean score provided by the CNN for the instances that contain the identity corresponding to the query (genuine queries) is lower than the mean score for the subset of instances containing one specific impostor identity. Formally, let \( X_1, \ldots, X_n \) be \( n \) independent and identically distributed (i.i.d) random variables describing the scores generated for genuine queries. Also, let \( Y_1, \ldots, Y_m \) be \( m \) i.i.d. random variables for a subset of impostor queries that share one specific impostor identity. Let \( S_X(n) = \sum_i X_i, S_Y(m) = \sum_i Y_i \) and the corresponding means \( \bar{X}(n) = \frac{S_X(n)}{n}, \bar{Y}(m) = \frac{S_Y(m)}{m} \). Let \( E(X_i) = \mu_X, \sigma_X, E(Y_i) = \mu_Y, \sigma_Y \) and \( \text{Var}(X_i) = \sigma_X^2 \) and \( \text{Var}(Y_i) = \sigma_Y^2 \). Then, we have \( \text{Var}(S_X(n)) = n\sigma_X^2 \) and \( E(S_X(n)) = n\mu_X \). Similarly, Gangwar and Joshi [9] described two architectures for CNNs that receive pairs of normalized iris samples and report a binary (genuine/impostor) decision, concluding about the advantages of these models with respect to hand-crafted feature-based approaches. Zhang et al. [48] fused (at score level) two algorithms for iris recognition: 1) based in hand-crafted ordinal measures (multi-lobe differential filters); and 2) based in a CNN that receives pairs of normalized images and performs binary discrimination. Authors argue that scores from both algorithms are complementary, which maximises the benefits of fusion with respect to the best standalone classifier. Nguyen et al. [20] used the responses of the CNN’s fully connected layers as feature descriptors. Five well known models (AlexNet, VGG, Inception, ResNet and DenseNet) were fine-tuned and fed a SVM used for multi-class discrimination (one-against-all mode), having authors reported state-of-the-art performance. Zhao and Kumar [50] used fully CNNs to obtain spatially meaningful iris features, using an adapted loss function accounting for bit shifting and non-iris masking. In all these works, the most discriminative features were automatically inferred by the deep learning frameworks, in opposition to the former generation of methods that explicitly introduced several types of texture, spectral and geometrical features claimed to be good choices for the iris recognition task (e.g. [18]).

2) VW Iris recognition Arsalan et al. [3] proposed a two-stage iris segmentation scheme based on CNNs that run after a coarse estimation of the iris boundaries, based on

http://www.jonathanjordan.staff.shef.ac.uk/IntroPS/part5.pdf

II. RELATED WORK

There is a large number of deep learning-based methods for biometric recognition, using traits such as the face (e.g., [42], [37], [28] and [12]), the gait (e.g., [13]) or the body silhouette (e.g. [14]). There have been several attempts to use deep learning-based models to learn mappings between biometric samples distance and their visual similarity. Schorff et al. [32] described a CNN model based on triplets that attempt to minimize distance between a sample and a genuine (same class) gallery element, while maximising the distance to an impostor sample. This learning scheme directly produces a mapping from facial samples to a compact Euclidean space where distances directly correspond to face similarity. A similar work was due to Wu et al. [44], which introduced a light framework to learn a compact embedding from large-scale facial data inaccurately labelled.

In the specific case of iris/periocular recognition, we divide the existing methods into four groups: 1) working on near-infrared (NIR) iris data acquired under constrained acquisition setups; 2) using visible-wavelength (VW) data to perform iris recognition; 3) working in unconstrained environments and using periocular VW data; and 4) aiming at soft labels estimation.

1) NIR Iris recognition Minaee et al. [19] studied the effectiveness of features resulting from deep learning architectures, that feed support vector machines (SVMs) working in the multi-class one-against-all mode. Authors observe that even this classical processing chain outperforms the former generation of hand-crafted feature based approaches. Similarly, Gangwar and Joshi [9] described two architectures for CNNs that receive pairs of normalized iris samples and report a binary (genuine/impostor) decision, concluding about the advantages of these models with respect to hand-crafted feature-based approaches. Zhang et al. [48] fused (at score level) two algorithms for iris recognition: 1) based in hand-crafted ordinal measures (multi-lobe differential filters); and 2) based in a CNN that receives pairs of normalized images and performs binary discrimination. Authors argue that scores from both algorithms are complementary, which maximises the benefits of fusion with respect to the best standalone classifier. Nguyen et al. [20] used the responses of the CNN’s fully connected layers as feature descriptors. Five well known models (AlexNet, VGG, Inception, ResNet and DenseNet) were fine-tuned and fed a SVM used for multi-class discrimination (one-against-all mode), having authors reported state-of-the-art performance. Zhao and Kumar [50] used fully CNNs to obtain spatially meaningful iris features, using an adapted loss function accounting for bit shifting and non-iris masking. In all these works, the most discriminative features were automatically inferred by the deep learning frameworks, in opposition to the former generation of methods that explicitly introduced several types of texture, spectral and geometrical features claimed to be good choices for the iris recognition task (e.g. [18]).
preprocessing and edge detection steps. Similarly, Bazrafkan et al. [4] described a Fully Convolutional Deep Neural Network model (FCDNN) to segment VW iris images of poor quality. Menon and Mukherjee [17] assessed the applicability of CNN-based frameworks to VW iris biometrics, using fine-tuned frameworks based on deep residual networks.

3) VW Periocular biometrics Ahuja et al. [1] (extended in [2]) compared the effectiveness of unsupervised/supervised CNNs for periocular recognition in the visible spectrum (VW), observing optimal performance when CNNs were used exclusively to extract 512-dimensional feature vectors, latter matched by the cosine similarity. Zhao and Kumar [51] fused scores from multiple CNNs, one of them tuned according to identity and the remaining incorporating explicit semantic information, such as gender, ethnicity and age. The fused model was claimed to recover comprehensive image features and achieve superior performance, when compared to the traditional way to use CNNs. Proença and Neves [26] argue that the iris region should be disregarded in the case of VW periocular biometrics, due to corneal reflections, gaze and frequent occlusions. Using a segmentation algorithm, the iris was separated from the periocular region, producing multi-class samples used in CNN learning that implicitly force the CNN to disregard the iris region from recognition. Wang et al. [43] described a convolutional and residual framework for the periocular recognition, both for near-infrared and VW data, claiming that such architecture learns in a relatively fast way and avoids feature saturation. Raghavendra and Busch [27] extracted texture information (using maximum response filters) from periocular data and learned the corresponding representations by coupling four layers of regularized auto-encoders. Rattani and Derakhshani [30] assessed the effectiveness of CNN models (VGG-16, VGG-19, InceptionNet and ResNet), fine-tuned for periocular recognition in handheld devices, claiming that “fine-tuning” attains performance comparable to “learning-from-scratch”, while demanding less quantities of learning data. A novel concept of multi-glance was due to Zhao and Kumar [52], in which part of the CNN intermediate components are configured to incorporate emphasis on regions considered semantically important (e.g., the eyebrow and the eye globe).

4) Soft biometrics Rattani et al. [29] used shallow CNN (with six hidden layers) to estimate gender and age in periocular samples acquired from handheld devices. They concluded that such frameworks still have enough discriminating power, even in case of poor-quality samples. Similarly, Samangouei and Chellapa [31] used shallow CNN models to estimate soft labels, comparing the effectiveness attained when using the whole face or exclusively the periocular band. Singh et al. [34] presented an auto-encoder that learns discriminative representations for gender and ethnicity information, based on near infrared periocular data. A set of baseline results for soft labels estimation in degraded data was announced by Gonzalez-Sosa et al. [35].

Recently, there were several CNN-based works concerned about the fusion of both the iris and periocular traits, as an attempt to augment the recognition robustness to hand-held devices. Zhang et al. [49] first applied max-out units into the CNNs to generate compact representations for both the iris and periocular traits, fusing the discriminative features of both modalities through weighted concatenation.

III. PROPOSED METHOD: In-Set Iterative Analysis

A. Learning Phase

We adopt the notation suggested by Bolle et al. [5]. Let \( x \in \mathbb{N}^d \) be an iris/periocular image (query) represented as
a column vector, with \( \mathcal{I}(x) \) expressing the corresponding identity. Let \( \{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\} \) be a set of \( k \) samples taken from \( g \) gallery identities. There are two disjoint hypotheses:

\[
H_0: \quad \exists \ i \in \{1, \ldots, k\} : \mathcal{I}(x^{(i)}) = \mathcal{I}(x); \\
H_a: \quad \forall \ i \in \{1, \ldots, k\} : \mathcal{I}(x^{(i)}) \neq \mathcal{I}(x).
\]

Let \( f : \mathbb{R}^{d} \to [0, 1] \) be the function performing the in-set analysis, i.e., \( f([x, x^{(1)}, x^{(2)}, \ldots, x^{(k)}]) = s \), with \( s \) being the matching score. The learning phase comprises the inference of one binary discrimination model to distinguish being the matching score. The learning phase comprises the inference of one binary discrimination model to distinguish the instance is \( k \) samples in the training set, we create \((n)\) hypotheses. In practical terms, we approximate \( f(.) \) by a CNN. During the leaning phase, having \( n \) samples in the training set, we create \( \binom{n}{k} \) combinations of \( k \) gallery elements, each one (plus the query) forming one learning instance. As illustrated in Fig. 2, in any case where the query has the same identity of a gallery element, the instance is considered genuine (i.e., label "1"). Otherwise, if all identities of the \( k+1 \) elements are different, we consider the instance as impostor (label "0").

![Fig. 2. How the learning data is labelled: cases where two elements among the \( k+1 \) (4 in the example) have the same identity are considered positive instances (i.e., label "1"). Otherwise, negative instances have all \( k+1 \) elements with different identities associated (label "0").](image)

### B. Recognition I: Iterative Selection of Gallery Samples

During runtime, the in-set workflow is divided into two parts: 1) iterative selection of the \( k \) gallery elements that - along with the query - form the CNN input; and 2) fusion of the responses provided by the CNN to infer the probability of matching the query to the gallery identities.

In this section we consider that the scores \( s \) assume maximal values under the null hypothesis \((H_0)\), i.e., when the query’s identity is equal to one of the \( x^{(i)} \) elements. To choose the \( k \) gallery elements that form the CNN input at one iteration we use the posteriors for the query’s identity being equal to gallery identities. According to the Bayes rule, such posteriors are given by:

\[
p(I(x) = i | s) = \frac{p(s | I(x) = i) p(I(x) = i)}{p(s)},
\]

where \( p(s | I(x) = i) \) corresponds to the likelihood of score \( s \) in the \( i^{th} \) identity distribution, \( p(I(x) = i) \) is the identity prior and \( p(s) \) is the probability for observing the score \( s \). However, performing Bayesian inference according to (1) requires to provide the likelihood distributions per identity, which have to be learned from labeled training data and would be clearly infeasible, due to large values of \( g \).

By relaxation, \( p(s | I(x) = i) \) can be approximated by \( p(s | H_0) \). This way, the probability that \( x \) corresponds to the \( i^{th} \) identity is given by:

\[
p(I(x) = i | s) = \frac{p(s | H_0) p(H_0)}{p(s)}
\]

with \( p(H_0) = \frac{k}{g} \), and \( \forall i \in \{1, \ldots, k\} \). Assuming that \( p(s | H_0) \) and \( p(s | H_a) \) are given from a training set, a Bayesian framework allows to recursively update the \( p(I(x) = i | s) \) values \( \forall i \in \{1, \ldots, g\} \).

Let \( s^{(t)} = [s_1, \ldots, s_t] \) denote the \( t \) scores given by the CNN after \( t \) iterations. Under a naïve-Bayes formulation, i.e., considering that scores \( s_i \) are conditionally independent, we obtain:

\[
p(s^{(t)} | I(x) = i) = \prod_{j=1}^{t} p(s_j | I(x) = i),
\]

which enables to recursively update the posteriors for each identity:

\[
p(I(x) = i | s^{(t)}) = \frac{p(s_i | I(x) = i) p(I(x) = i | s^{(t-1)})}{p(s_i | s^{(t-1)})}
\]

for \( t > 1 \), with \( p(I(x) = i | s^{(0)}) = \frac{k}{g} \). Additional details are given in [11], particularly how this recursion can be formulated in computationally efficient matrix-vector form.

The probabilities that the query does not correspond to the gallery identities \( p(I(x) \neq i | s^{(t)}) \), \( \forall i \in \{1, \ldots, g\} \) are obtained by probability complement and used to select the gallery elements that form the CNN input at each iteration. Here, the strategy is to privilege the identities that most unlikely match the query, up to a moment when such identities are definitely rejected, when \( p(I(x) \neq i | s^{(t)}) \geq \tau_p \) (\( \tau_p \) close to 1). This way, a decreasing number of plausibly (and choosable) identities remain for the subsequent iterations. Formally, the likelihood for selecting samples from the \( i^{th} \) gallery identity is given by the sigmoid function:

\[
l(i) = \begin{cases} 
\frac{1}{1 + e^{-\tau_m} - \tau_c} s\left(I(x) \neq i | s^{(t)}\right) - \tau_c, & \text{if } p(I(x) \neq i | s^{(t)}) < \tau_p \\
0, & \text{otherwise},
\end{cases}
\]

where \( (\tau_m, \tau_c) \) are the parameters that control the smoothness and center of the sigmoid. Figure 3 illustrates the parameterization of the transfer function used in all our experiments.

At each iteration, the \( l(i) \) values determine the chances for selecting each identity \( p(i) = \frac{l(i)}{\sum_{j=1}^{l(i)}} \). Let \( \Gamma_t \) be the set of identities of the \( k \) gallery elements used in the CNN input at iteration \( t \), i.e., \( \Gamma_t = \{I(x^{(1)}), \ldots, I(x^{(k)})\} \). As the network is supposed to fire when the query identity is equal to one of the \( k \) gallery elements, a score \( s \) has one of two meanings:
1) $s \approx 1$, in case of the null hypothesis, i.e., one element of the input matches the query identity; or 2) $s \approx 0$, when no identities in the input set are repeated.

C. Recognition II: Identification

Identification starts by estimating the probabilities that the query identity doesn’t correspond to each gallery element, to progressively reject some of the identities. Once a sufficient number of iterations is reached, the problem remaining is how to fuse the scores for the uncertain identities, i.e., those having a non-residual probability of corresponding to the query. This part is regarded as a variable selection and regularization problem, and we define an indicator (characteristic) function:

$$I(\Gamma_t, I(x^{(i)})) = \begin{cases} 1 & \text{if } I(x^{(i)}) \in \Gamma_t' \\ 0 & \text{otherwise.} \end{cases}$$  \hspace{1cm} (6)

Let $c_{ji} = I(\Gamma_j, I(x^{(i)}))$ express the value of the indicator function (6) for the $j^{th}$ iteration and the $i^{th}$ identity ($c_{ji} = 1$ denotes that the $i^{th}$ identity was part of the CNN input in the $j^{th}$ iteration). After $t$ iterations, the following matrix is obtained:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1g} \\ c_{21} & c_{22} & \cdots & c_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ c_{t1} & c_{t2} & \cdots & c_{tg} \end{bmatrix}_{t \times g}, \; s = \begin{bmatrix} \hat{s}_1 \\ \hat{s}_2 \\ \vdots \\ \hat{s}_t \end{bmatrix}$$  \hspace{1cm} (7)

with $s$ representing the CNN scores after $t$ iterations. Keeping in mind that:

$$\sum_{i=1}^{g} c_{ji} = k, \; \forall j \in \{1, \ldots, t\},$$  \hspace{1cm} (8)

and that $k \ll g$, $C$ is a sparse matrix that determines $s$. Looking to this relationship from the perspective of the correlation between the inputs measurements $C$ and the outcomes $s$, yields a regression problem with variable selection and regularization, in which finding the query’s identity is equivalent to determine the most important measurement (column of $C$) for obtaining $s$. This is a LASSO problem, solved as described in [40]:

$$\hat{\alpha} = \arg \min_{\alpha} ||C\alpha - s||^2 \text{ s.t. } ||\alpha||_1 = 1.$$  \hspace{1cm} (9)

The resulting vector $\hat{\alpha}$ has $g$ coefficients, and expresses the likelihood of the query having equal identity to each of the gallery elements. In the noiseless case: $\exists i \in \{1, \ldots, g\} : \hat{\alpha}_i = 1 \land \forall j \neq i : \hat{\alpha}_j = 0$, but in practical terms: $\exists i \in \{1, \ldots, g\} : \hat{\alpha}_i \approx 1 \land \forall j \neq i : \hat{\alpha}_j \approx 0$, or at least $\exists i \in \{1, \ldots, g\} : \hat{\alpha}_i \gg \hat{\alpha}_j, \forall j \neq i$.

D. Computational Complexity

In terms of the learning phase, the computational (time) complexity of the in-set analysis is roughly the same of the baselines 1:1 and 1:N. The unique exception is the depth of the input data and the corresponding depth of the filters used in the first convolution layer of the CNN. A much more important factor is the classification time cost, since this is the phase that should run at real-time. It is known that this cost depends of many parameters of the CNN architectures, such as the number of convolution layers, the size each layer receptive field and the input dimension (the architecture we used - VGG-16 - has about 124 million weights).

Independently of the CNN architecture, the key point is the relative complexity of the in-set analysis, when compared to the baselines 1:1 and 1:N working modes. During classification, we show ($t$ times) groups of $k + 1$ elements to the network, to reject some of the known identities (determined by the value of $\tau_p$). However, in all our experiments, the number of iterations was always far below ($t \ll g - 1$) the number of identities, being $g - 1$ the number of times the CNN’s forward step runs for one query in the 1:1 mode. On the other way, a 1:N query requires only one forward propagation of the CNN, but it is far more demanding in terms of the amount of data used in the learning phase to learn appropriate feature representations.

Finally, the identification step uses the Lasso optimization algorithm, which is done in $O(g^3 + g^2t)$, being $g$ the number of columns (number of identities) and $t$ the number of observations (CNN queries) [8]. Empirically, the average time taken by the in-set analysis to perform one identification query was $0.716 \pm 0.170$ ms. (using $\tau_p = 0.9995$, $k = 5$ and $g = 1000$), which was slightly lower than the value observed for the 1:1 analysis (0.950 ± 0.014 ms.), but higher than the 1:N mode (0.112 ± 0.009 ms.). These values were obtained using the hardware infrastructure and software framework described in section IV-A, without any performance optimization concerns.

IV. RESULTS AND DISCUSSION

A. Data and Experimental Protocol

Two datasets were used in our experiments: 1) the CASIA-IrisV4-Thousand\(^3\), for evaluating near-infrared iris recognition performance. It contains 20,000 iris images from 2,000 classes (eyes), with the sources of intra-class variations being mostly

\(^3\)https://arxiv.org/pdf/1703.09039.pdf
\(^4\)CASIA iris image database, http://biometrics.idealtest.org
eyeglasses and iris occlusions, due to eyelids and specular reflections; and 2) the UBIRIS.v2 [23], for evaluating periocular recognition performance in VW data acquired in unconstrained conditions. This set contains 11,102 images from 522 eyes, taken from varying distances and subjects’ poses, leading to some severely degraded samples (Fig. 4).

Images from both sets were resized to 150 × 200 pixels. Additionally, in the near-infrared set, irises were segmented according to a coarse-to-fine strategy [33], using form fitting and geodesic active contours algorithms. The pupillary boundaries were described by shapes of 20 degrees-of-freedom (dof) and the scleral boundaries by shapes of 3 dof. Next, images were normalised into the pseudo-polar domain [7], with size of 64 × 256 pixels. The right halves of all images were discarded, corresponding to the upper half of the irises in the original representation, known to have the highest probability of being occluded.

![Fig. 4. Iris/periocular datasets used in the evaluation of our method. The upper part of the figure illustrates degraded (original + augmented) periocular samples from the UBIRIS.v2 set, whereas the bottom rows regard original, segmented and augmented samples of the CASIA-Iris-V4-Thousand set.](image)

In all experiments, disjoint identities were used in the learning/test phases. For the CASIA-IrisV4-Thousand set, only the first 1,000 classes were used in the learning phase of the CNN, while for the UBIRIS.v2 only the first 261 classes were used, as described in Table I. Performance was evaluated according to a bootstrapping-like strategy widely reported in biometric recognition literature (e.g. [10]): having n test images available, the bootstrap randomly selects 0.9n images, with experiments being repeated in each bootstrap set, and the average and standard deviation performance values taken at all operating points. These are the values reported in Table I and correspond to the lines in the ROC and RANK-N plots (with the shadowed regions denoting the standard deviations). The MATLAB® programming language was chosen, and the MatConvNet [41] toolbox used to learn the CNN models, according to the details provided in Table I. A NVIDIA® Titan X GPU with 12GB memory and 3,072 CUDA cores speeded-up the learning processes.

### B. Learning and Parameter Tuning

The VGG-16 [38] was the CNN architecture considered for our experiments, which is one of the most popular deep learning models for image classification. The unique adaptations were related to the size of the input data: 64 × 128 for CASIA-IrisV4-Thousand data and 200 × 150 × 3 for UBIRIS.v2 samples. Also, as the in-set method uses CNN learning instances composed of k + 1 images, the CNN inputs had k + 1 and 3(k + 1) channels respectively for the CASIA-IrisV4-Thousand and UBIRIS.v2 datasets, requiring filters of the same depth in the CNN input layer.

Learning was based in the stochastic gradient descent algorithm, minimizing the multinomial logistic regression (multinomial logit) loss on mini-batches of 128 samples. This choice was due to the intention of using the same loss function for all CNN variants tested (in-set, CNN-Pairwise and CNN-SVM). Following the parameterizations suggested by authors of the VGG-16 model, momentum was set to 0.9, the initial learning rate set to 0.001 and then iteratively decreased one order of magnitude at the end of every epoch without improvements in the validation performance. According to the strategy described in section III-B, we used binary labels, with “0” being the target for any instance where the k input samples regard different identities and “1” the target corresponding to instances where one gallery element has the same class as the query sample. Essentially, the CNN models were asked to learn a binary discrimination problem, i.e., to distinguish between samples of k elements that share some identity (our null hypothesis) or not.

As data augmentation, two label-preserving transformations were used: 1) to simulate the scale and translation samples inconsistency, patches of scale [0.75, 0.90] (values drew uniformly) were randomly cropped; and 2) as a color transformation, the principal components of the RGB/intensity

<table>
<thead>
<tr>
<th>Parameter</th>
<th>UBIRIS.v2</th>
<th>CASIA-IrisV4-Thousand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total images</td>
<td>11,102</td>
<td>20,000</td>
</tr>
<tr>
<td>Total classes</td>
<td>522</td>
<td>2,000</td>
</tr>
<tr>
<td>Learning classes</td>
<td>261 (1-261)</td>
<td>1,000 (1-1,000)</td>
</tr>
<tr>
<td>Data augmentation</td>
<td>18x: 12x scale + translation, 6x color</td>
<td></td>
</tr>
<tr>
<td>CNN learning</td>
<td>batch size: 256, learning rate: 0.001; momentum: 0.9</td>
<td></td>
</tr>
<tr>
<td>Test classes</td>
<td>261 (262-522)</td>
<td>1,000 (1,001-2,000)</td>
</tr>
<tr>
<td>Samples/class</td>
<td>15-30</td>
<td>10</td>
</tr>
<tr>
<td>Gallery samples/class</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

TABLE I DETAILS ABOUT THE DATA SETS AND THE LEARNING/TEST PROTOCOLS USED IN THE EXPERIMENTAL VALIDATION OF THE PROPOSED METHOD.
values in 10,000,000 pixels of the learning data were found, and used to create synthetic images by adding to each pixel multiples of the largest eigenvectors with magnitude equal to their eigenvalues [15]:

\[ \mathbf{x}^{(\text{new})} = \mathbf{x}^{(\text{old})} + [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \left( \alpha \odot [\lambda_1, \lambda_2, \lambda_3]^T \right), \]

with \( \odot \) denoting the element-wise multiplication, \( \mathbf{v} \), and \( \lambda \), denoting the eigenvectors and eigenvalues of the learning data covariance matrix and \( \alpha \in \mathbb{R}^3 \) being randomly drawn from the Gaussian \( \mathcal{N}(0, 0.1) \). As Table 1 describes, we chose a factor of \( 18 \times \) for the amount of augmented data, with respect to the original size of the dataset. This value was obtained by trial-and-error on both datasets, using \( \{0, 1, \ldots\} \) times of augmented data and observing the variations in performance, having stopped when the improvements became residual (even though for the UBIRIS.v2 dataset, a higher amount of learning data could have been used, obtaining slightly higher recognition rates).

The initial concern was the sensitivity of the proposed method with respect to the \( k \) and \( \tau_p \) parameters, i.e., how many images at once should be provided as CNN input (\( k \)) and how early (\( \tau_p \)) the gallery identities can be confidently rejected. Overall, we observed a higher sensitivity to the \( \tau_p \) parameter than to the value of \( k \), even though in this case only moderately low values (up to \( 7 \)) were tested. Regarding the \( \tau_p \) parameter, good values were observed to be larger than 0.99, as smaller values often led to erroneous precocious rejections of the ground identity. Oppositely, values around 1 were observed not to significantly affect performance, yet they increased the number of identities considered in the fusion step (section III-C). Overall, CNN queries composed of 4 gallery elements (plus the query) yielded the optimal recognition performance, leading to conclude that higher values would imply to learn in feature spaces of increasing complexity that would demand too large amounts of learning data.

The most important conclusions were drawn from the results provided in Fig. 5, which summarizes the variations in performance with respect to the \( k \) and \( \tau_p \) parameters. The plot given at the left side illustrates the in-set performance (3D plot) with respect to the \( k \) and \( \tau_p \), using as baseline the performance attained by the pairwise matching CNN mode, represented by white horizontal plane.

C. Results' Insight

Fig. 6 provides the insight for the effectiveness of the in-set analysis: the left plot compares the decision environments obtained for CNNs working in the pairwise (continuous lines) and the in-set modes (dashed lines), with the green color representing genuine scores and the red color representing impostors’ (the distributions were approximated based in 4,096 genuine (in green) plus 4,096 impostor (in red) scores. Not surprisingly, there is a slight degradation in the separability between both distributions in the case of the in-set analysis, due to the higher complexity of the feature space. This is evident in the zoomed-in region (near the vertical axis), with the pairwise impostors distribution showing a much narrower peak than its in-set counterpart. Essentially, this means that the in-set CNN was not as efficient as its pairwise counterpart to return low (\( \approx 0 \)) matching scores for the impostor instances. A similar observation, yet less evident, can be made for the genuine distributions.

However, the most interesting point is to perceive which instances had their scores degraded and by how much, which
can be observed from the pairwise/in-set scores correlation. The plot given at the right part of Fig. 6 correlates the scores for pairwise (horizontal axis) and in-set (vertical axis) analyzes, for impostor and genuine comparisons. For each pairwise comparison, $k-1$ random gallery images were added iteratively to create the corresponding in-set samples. Under this experimental setting, impostor observations above the diagonal dashed line in the 2D plot represent worse performance for the in-set than the pairwise mode, with the opposite occurring for the genuine observations. The key observation is that there are almost no impostor scores in the quadrant "Q1", as practically there are not genuine observations in "Q3", which will be the concerning cases (i.e., low pairwise and high in-set impostor scores, or high pairwise and low in-set genuine scores). However, note the large number of genuine scores that spread along the $y=1$ line, which are cases where the pairwise CNN had difficulties to consider the comparison as genuine, but where the iterative in-set analysis yielded much better results in this task. We draw two main conclusions here:

1) the in-set iterative analysis decreases the probability of (e.g., due to data acquisition settings) observing outlier low genuine matching scores, when comparing to the pairwise matching strategy. This is known to be a major error source of biometric recognition, particularly in case of poor quality samples;

2) globally, the in-set iterative analysis provides slightly higher (worse) impostor scores and slightly lower (worse) genuine scores than pairwise matching. However, such deteriorations play a minor role in the final recognition performance, as the deviations are typical far from the critical uncertain region that separates both classes.

D. Baseline Methods

The main baselines considered were the traditional working modes of CNNs in biometric recognition: using pairwise comparisons (1:1, CNN-Pairwise), or providing samples individually to the networks and using the feature descriptions from the first fully connected layer to feed a SVM for classification (1:N, CNN-SVM). In all cases, the VGG-16 architecture was chosen, with the unique adaptation regarding the depth of the filters in the input layer, that was set equal to the number of channels in the input data ("2" for the CNN-Pairwise and "1" for the CNN-SVM).

As additional iris recognition baselines, we have chosen: 1) Sun and Tan’s method [36], with dillobe and tri-lobe filters, Gaussian kernels $5 \times 5$, $\sigma = 1.7$, inter-lobe distances $\{5,9\}$ and sequential feature selection; 2) Yang et al. [46]’s method (using the O2PT iris-only variant, with block size $w = 2$, $h = 14$, translation vector $[6,3]^T$ and neighbourhoud $8 \times 8$); and 3) the OSIRISv4 framework [21], to represent the processing chain proposed by Daugman [7]. The used version segments the iris based on the Viterbi algorithm and normalizes data according to the Rubber Sheet scheme. Feature extraction is carried out using a set of 2D-Gabor filters and iriscodes are matched by the Hamming distance.

Additionally, two baselines were chosen for periorcular recognition: 1) Zhao and Kumar [47]’s method, using feature representations from a pair of CNNs, with one network learning semantic information ("right"/"left" eye classes and gender), and the other inferring samples’ identity. The feature vectors from both networks were matched according to a log likelihood ratio, in a verification (1:1) setting; and 2) Proencña [24]’s method, using shape and texture descriptors to parameterize a weak biometric expert (periorcular), and multi-lobe differential filters from the RGB, HSV, XYZ and Opponent-RGB colour spaces to characterise the iris. These methods were selected not only to represent the hand-crafted feature-based recognition strategies, but also the deep learning-based frameworks, while both aiming at increasing the recognition robustness to degraded data.

E. Performance Comparison

Fig. 7 provides the ROC curves for the CASIA-IrisV4-Thousand (at left) and UBIRIS.v2 (at right) sets, expressing the recognition performance in the verification mode. In all cases, the lines represent the average performance in the bootstrap sub-sets and the shadowed regions denote the standard deviation values observed at each performance point. Overall, the in-set strategy solidly outperformed all competitors, not only with respect to the CNN-Pairwise and CNN-SVM strategies, but also the hand-crafted feature-based and deep-learning based methods. Importantly, this happened in practically all regions of the performance space, and in most parts providing disjoint performance confidence intervals with respect to the other methods. Inside each ROC, we provide the corresponding values in logarithmic scale, that turn particularly evident the solid improvements in performance for low false acceptance rates, which is particularly important for large scale applications. The CNN-Pairwise was the runner-up in most regions of the performance space, with exception to the region with the lowest levels of false acceptances, where pairwise matching was affected by outlier genuine matching scores. In all cases, the hand-crafted feature-based methods got far worse performance than the deep-learning based techniques, which also accords the most recent reports comparing the performance attained by both families of methods.

As a complement, Fig. 8 provides the counterpart results for the identification mode, showing the accumulated rank-n curve in linear and logarithmic scales. Overall, results accord the performance levels previously observed for the verification mode, with the proposed in-set strategy obtaining over 0.99 rank-1 average accuracy in the CASIA-IrisV4-Thousand set, and over 0.88 in the challenging UBIRIS.v2 dataset. Regarding the baselines, Yang et al. [46] got the second best performance in iris data, while on the more challenging periorcular environment, the method due to Zhao and Kumar [47] was consistently better than the hand-crafted based Proencn’s approach. Also, in this case, it should be noted that we performed some preliminary experiments using additional semantic features (eye color), that point that the performance Zhao and Kumar [47]’s method might be boosted up in case that additional reliable semantic features are used.
Overall, we observed that OSIRIS matching scores tended to degrade evidently in case of samples inaccurately segmented by the Viterbi algorithm, while Sun and Tan’s approach faced difficulties in case of large occlusions in regions that were almost noise-free in the whole learning set. In these cases, the learned set of filters extracts data from poorly discriminating regions of the irises. Being phase-based, the approach of Yang et al. [46], Sun and Tan [36] and OSIRIS [21] for iris recognition and to Zhao and Kumar [47] and Proença [24] for periorcular recognition are given (the standard deviation in performance observed in 10 bootstrap test subsets is denoted by the shaded region of each line series).

As a summary, Table II includes three performance measures (AUC, Rank-1 and EER) for the methods evaluated, in the CASIA-IrisV4-Thousand and UBIRIS.v2 sets. The average values in the 10 bootstrap test subsets are given, together with the corresponding standard deviation values (denoted by the ± symbol).

V. CONCLUSIONS AND FURTHER WORK

This paper describes a novel way to use CNNs in biometrics. The idea is to obtain an answer to the in-set question: “is the
query’s identity in this set?”, by showing to the network not only the query but also k gallery elements at once. Iteratively, if multiple random gallery samples are used, we concluded that many weakly correlated CNN matching scores can be obtained, which altogether provide solid cues about the most likely matching identity, resembling the rationale followed to play the classical Mastermind game. At the end, by analysing the CNN responses, identification is regarded as a variable selection and regularization problem, solved by sparse linear regression techniques, in which finding the true identity is equivalent to determine the most important measurement, for the set of observed scores.

Using k + 1 samples as CNN input not only augments the potential amount of learning data (having n learning samples, \( \binom{k+1}{n} \gg \binom{k}{n}, \forall k \geq 2 \)), but can also be seen as an attempt to recognize one particular class of object (subject) from different perspectives, i.e., when comparing it to samples of many other object types. Being known as heavily data-driven models, both properties contribute to improve the CNN’s classification performance.

The experimental validation of our method was carried out in two well known iris/periocular data sets (CASIA-IrisV4-Thousand and UBIRIS.v2). In both cases, the proposed in-set method got solidly the best results among all competitors tested, without substantially overloading the temporal complexity of the recognition task. It should be noted that these results were observed for full versions of the data sets, i.e., without disregarding any sample or using any friendly version of the datasets.

Acknowledgements

This work was supported by PEst-OE/EEI/LA0008/2013 research program. Also, we acknowledge the support of NVIDIA Corporation®, by the donation of one Titan X GPU.

References


[51] Z. Zhao and A. Kumar. Accurate Periocular Recognition Under Less Constrained Environment Using Semantics-Assisted Convolutional Neu-