Evaluation of Fine Tuning and Feature Extraction methods in Biometric Periocular Recognition

William Barcellos
University of São Paulo
São Carlos, Brazil
william.barcellos@gmail.com

Nicolas Hiroaki Shitara
University of São Paulo
São Carlos, Brazil
hiroshitara@gmail.com

Carolina Toledo Ferraz
Unifaccamp
Campo Limpo Paulista, Brazil
caroltoledoferraz@gmail.com

Raissa Tavares Vieira Queiroga
Federal University of Rio Grande do Norte
Natal, Brazil
raissa.tavares@gmail.com

Jose Hiroki Saito
Unifaccamp
Campo Limpo Paulista, Brazil
saitojosehiroki@gmail.com

Adilson Gonzaga
University of São Paulo
São Carlos, Brazil
agonzaga@sc.usp.br

***Abstract*— The aim of this paper is to evaluate the performance of Transfer Learning techniques applied in Convolucional Neural Networks for biometric periocular classification. Two aspects of Transfer Learning were evaluated: the technique known as Fine Tuning and the technique known as Feature Extraction. Two CNN architectures were evaluated, the AlexNet and the VGG-16, and two image databases were used. These two databases have different characteristics regarding the method of acquisition, the amount of classes, the class balancing, and the number of elements in each class. Three experiments were conducted to evaluate the performance of the CNNs. In the first experiment we measured the Feature Extraction accuracy, and in the second one we evaluated the Fine Tuning performance. In the third experiment, we used the AlexNet for Fine Tuning in one database, and then, the FC7 layer of this trained CNN was used for Feature Extraction in the other database. We concluded that the data quality (the presence or not of class samples in the training set), class imbalance (different number of elements in each class) and the selection method of the training and testing, directly influence the CNN accuracy. The Feature Extraction method, by being more simple and does not require network training, has lower accuracy than Fine Tuning. Furthermore, Fine Tuning a CNN with periocular's images from one database, doesn't increase the accuracy of this CNN in Feature Extraction mode for another periocular's database. The accuracy is quite similar to that obtained by the original pre-trained network.**

Keywords—CNN, transfer learning, fine tuning, feature extraction biometric periocular recognition

I. INTRODUÇÃO

Um dos problemas de um sistema biométrico está relacionado à sua aceitação devido a equipamentos invasivos ou que necessitem da colaboração do usuário. Além disso, a aquisição de imagens em ambientes reais não controlados onde as imagens são geralmente de baixa qualidade, baixa resolução e ruidosas, pode comprometer a acurácia do sistema. Por isso, os sistemas biométricos aplicados por exemplo na área de vigilância, devem ser o mais discretos possíveis, evitando a interação direta com o usuário e resolvendo os problemas de imagens adquiridas em ambientes sem restrições. Diz-se que um sistema biométrico nessas condições é não-cooperativo e não controlado [1][2]. Dentre os traços biométricos possíveis de serem capturados em ambientes sem restrições, a face e a íris ocupam lugar de destaque. No entanto, a face é facilmente disfarçável por adereços, maquiagens e características voláteis como barbas e bigodes. A região ocular pode ser obstruída normalmente por óculos, mas é um traço biométrico que pode ser usado mais facilmente em ambientes não controlados, devido à facilidade de aquisição com um único sensor (câmera) das peculiaridades biométricas oriundas de diferentes partes da face humana.

A peculiaridade biométrica ocular fornece diferentes fontes de informação para o reconhecimento, tais como, a retina, a íris, a conjuntiva e a região periocular. O reconhecimento biométrico ocular tem feito progressos nos últimos anos principalmente devido a avanços significativos realizados no reconhecimento de íris. Recentemente o uso da região periocular no reconhecimento biométrico tem ganhado popularidade. Esta peculiaridade biométrica permite utilizar não somente as informações de textura da íris, mas também as texturas da esclera e da pele perto do olho, bem como a forma da pálpebra, da sobrancelha e dos cílios. Esta região pode ser adquirida em condições não controladas, em contraposição às condições mais rígidas necessárias para o reconhecimento facial ou da íris, tornando-a mais adequada para aplicações em reconhecimento não cooperativo.

O reconhecimento periocular baseia-se na intrínseca capacidade humana de "reconhecer alguém simplesmente por olhar para seus olhos", o que fornece quantidades substanciais de informação discriminante[1][2][3] permanecendo relativamente estável durante longos períodos de tempo. Os elementos típicos da região periocular são mostrados na Fig. 1.

Recentemente, as aplicações de Aprendizagem Profunda, especificamente as Redes Neurais Convolucionais (*CNN*) tem alcançado resultados significativos na área de Visão Computacional. Redes profundas aplicadas ao reconhecimento biométrico da região periocular tem sido também avaliadas. Existem diversas maneiras para se utilizar as *CCNs* em aplicações de Visão Computacional. Quando a base de imagens de treinamento não é grande o suficiente para o ajuste conveniente dos pesos da rede, faz a opção pela

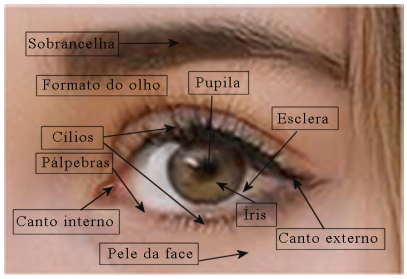

Fig. 1. Região periocular

técnica de *Transfer Learning (TL)* ao invés de se treinar uma rede do zero (*scratch*). A técnica de *TL* pode ser usada de diversas maneiras, mas duas delas se destacam na literatura: *Fine Tuning(FT) e Feature Extraction(FE)* [4].

O objetivo deste trabalho é avaliar a acuráca obtida por redes pré-treinadas no reconhecimento biométrico da região periocular comparando-se as duas técnicas de *FT e FE*.

Este trabalho está dividido em cinco seções. A Seção I introduziu o assunto mostrando os objetivos a serem alcançados. A Seção II apresenta os trabalhos correlatos em Visão Computacional tradicional (*handcraft*) e em Aprendizagem Profunda para o reconhecimento da região periocular. Na Seção III é apresentado o método proposto e as bases de imagem utilizadas. Os resultados obtidos são discutidos na Seção IV finalizando com as conclusões na Seção V.

II. Trabalhos correlatos em Reconhecimento Periocular

Existem alguns benefícios em se usar o traço biométrico periocular. Em imagens onde a íris não pode ser confiavelmente obtida, a região circundante ao olho pode ser usada para confirmar ou refutar uma identidade. Quando toda a face é adquirida a distância, as informações de íris são normalmente de baixa resolução, por outro lado, quando a íris é capturada de perto, o rosto inteiro pode não ficar disponível, forçando o sistema de reconhecimento a confiar apenas na íris. A região periocular pode ser útil em uma ampla gama de distâncias. Quando partes do rosto relativos à boca e nariz são ocluídas, a região periocular pode ser usada para determinar a identidade. Além disso, não existe a necessidade de sensores diferentes para íris e região periocular que podem ser obtidas usando um único sensor.

Os algoritmos tradicionais de Visão Computacional para o reconhecimento biométrico da região periocular buscam extrair características baseadas no formato do olho, formato da sobrancelha, formato das pálpebras, formato dos cantos dos olhos, distribuição dos cílios, textura da íris, textura da esclera e textura da pele da face, realizando ou não a fusão de dados para melhorar o desempenho. Vários métodos foram propostos recentemente, como o proposto por Park et al. [1], que caracteriza a textura periocular usando padrões binários locais (*LBP*), histograma de gradientes orientados (*HOG*) e *Scale Invariant Feature Transform* (*SIFT*). Os autores também descrevem fatores que afetam o desempenho do reconhecimento periocular, incluindo imprecisões de segmentação, oclusões e pose [2].

Woodard et al. [5] utilizaram *LCH* (Histogramas de cor RG), relatando melhor precisão. Tan e Kumar [6] propuseram filtros de Leung-Mallik (*LMF*) como descritores de textura para o banco de dados *CASIA v4*. Karahan et al. [7] avaliaram os descritores *LBP, SIFT* e outros descritores locais incluindo *SURF, BRISK e ORB* sobre o banco de dados *FERET*.

Duas revisões bastante detalhadas sobre métodos tradicionais de reconhecimento biométrico da região periocular foram publicadas por Alonso-Fernadez e Bigun [8] e por Rattani e Derakhshani [9]. Apesar da data de publicação dos dois artigos (2016 e 2017) ser recente, não incluem nenhum trabalho utilizando *CNN* para o reconhecimento periocular.

O *Deep-PRWIS* proposto por Proença e Neves [10] utiliza imagens obtidas no espectro da luz visível, descartando os elementos dentro do globo ocular (íris e esclera). O método é baseado em *CNNs* que definen as regiões de interesse nos dados de entrada que devem ser privilegiadas de forma implícita, ou seja, sem mascarar quaisquer áreas nas amostras de aprendizagem/teste. Essas amostras são usadas para fins de aumento de dados e para alimentar a etapa de treinamento da *CNN*. Durante a fase de teste, as amostras são fornecidas sem qualquer máscara de segmentação e a rede naturalmente desconsidera os componentes oculares, o que contribui para melhorias no desempenho.

Zhao e Kumar [11] usam uma *CNN* para reconhecimento periocular considerando informações de semântica explícita para extrair características mais abrangentes da região periocular, ajudando a *CNN* a melhorar o desempenho. Em um outro artigo [12] os autores propõem uma nova arquitetura de aprendizagem profunda para um reconhecimento periocular mais robusto e mais preciso que incorpora um modelo de atenção para enfatizar importante regiões nas imagens perioculares. Ao focar nessas regiões, a rede convolucional profunda é capaz de aprender características discriminativas adicionais, que por sua vez, melhora a capacidade de reconhecimento de todo o modelo.

III. Material e Método

Para o desenvolvimento deste trabalho foram avaliadas duas redes convolucionais pré-treinadas e por meio de duas diferentes técnicas de "aprendizagem por transferência" (*Transfer Learning*), estas redes foram submetidas a duas bases de imagens da região ocular comumente utilizadas para o reconhecimento de íris: a base *LAVI DB2* e a base *ND-IRIS-0405 Iris Image Dataset (LG2200)*. Estas bases apesar de não serem as mais adequadas ao reconhecimento da região periocular, apresentam algumas partes desta, tais como íris, pálpebras, esclera, pupila, cílios, textura da pele e canto dos olhos, e além disso, a investigação proposta centra-se na análise de desempenho das duas técnicas de *Transfer Learning*.

A. Redes Neurais convolucionais pré-treinadas

A primeira aplicação bem conhecida e bem sucedida de redes neurais convolucionais foi a *LeNet-5,* descrita por Yann LeCun, et al. [13]. A partir de então, diversas outras arquiteturas de *CNNs* tem surgido na literatura determinando evoluções importantes na classificação de imagens, tais como a *AlexNet, VGG, Inception e GoogLeNet, ResNet* e outras.

AlexNet [14] é uma rede neural convolucional treinada em mais de um milhão de imagens do banco de dados *ImageNet*. A rede tem 8 camadas de profundidade e pode classificar imagens em 1000 categorias de objetos, como teclado, mouse, lápis e muitos animais. Como resultado, a rede aprendeu representações de uma ampla gama de imagens. A imagem de entrada da *AlexNet* é de 227 x 227 pixels no formato *RGB*.

Um importante trabalho que buscou padronizar o projeto de arquitetura para redes neurais convolucionais profundas e desenvolveu modelos mais profundos e de melhor desempenho foi o artigo de Simonyan e Zisserman [15]. Sua arquitetura é geralmente referida como *VGG* devido o nome de seu laboratório, o *Visual Geometry Group* em Oxford. A primeira diferença importante que se tornou um padrão de fato é o uso de um grande número de pequenos filtros. Especificamente, filtros com o tamanho 3×3 e 1×1 com o *stride* de um, diferentes dos filtros maiores

na *LeNet-5* e os filtros menores, mas ainda relativamente grandes e *stride* de quatro da *AlexNet*. Diversas variantes da arquitetura foram desenvolvidas e avaliadas, embora duas sejam mais citadas, dado seu desempenho e profundidade. Elas são nomeadas pelo número de camadas: são a *VGG-16* e a *VGG-19* para 16 e 19 camadas, respectivamente.

B. *Transfer Learning*

Modelos de *CNNs* podem levar dias ou mesmo semanas para treinar, a partir do zero (*scratch*) em conjuntos de dados muito grandes. Uma maneira de reduzir esse tempo é reutilizar as ponderações de modelos pré-treinados, que foram desenvolvidos para conjuntos de dados padrão, como as tarefas de reconhecimento na base *ImageNet.* Modelos de alto desempenho podem ser baixados e usados diretamente, ou integrados em um novo modelo para outros problemas de visão computacional.

A aprendizagem por transferência geralmente se refere a um processo em que um modelo treinado em um problema é usado de alguma forma em um segundo problema relacionado. Na aprendizagem profunda, a aprendizagem por transferência é uma técnica na qual um modelo de rede neural é treinado primeiro em um problema com uma grande quantidade de dados [4]. Uma ou mais camadas do modelo pré-treinado são então usadas em um novo modelo, treinado sobre o novo problema de interesse. Ou seja, ao se usar um modelo de *CNN* pré-treinada, o número de imagens necessárias para treinamento e teste pode ser significativamente reduzido. Dentre as maneiras possíveis de se utilizar o conhecimento ou aprendizado adquirido por um modelo pré-treinado para outra aplicação, detacam-se as técnicas de *Fine Tuning (FT)* e de *Feature Extraction (FE)* [4].

C. *Fine Tuning (FT)*

O *FT* modifica os parâmetros de uma *CNN* existente para treinar uma nova tarefa. A camada de saída é estendida com pesos aleatoriamente iniciados para a nova tarefa e uma pequena taxa de aprendizado é usada para ajustar os parâmetros a partir de seus valores originais para minimizar a

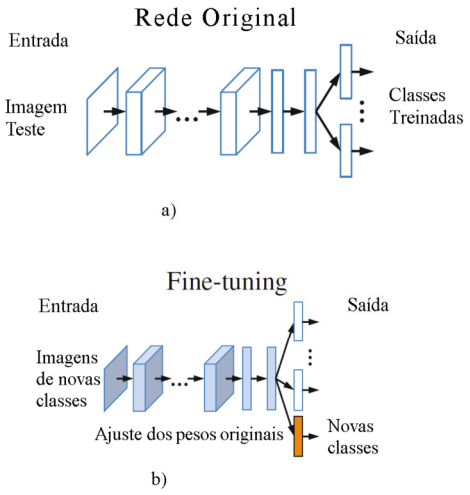

Fig. 2. *Transfer Learning* por *Fine Tuning*. a) *CNN* original pré-treinada. b) Ajustes dos pesos por FT. Adaptado de [4]

perda [16]. O *FT* adapta os parâmetros compartilhados para torná-los mais discriminativos para a nova tarefa, e a baixa taxa de aprendizagem é um mecanismo indireto para preservar algumas das estruturas representacionais aprendidas nas tarefas originais. A Figura 2 mostra um esquema utilizando o *FT* a partir de uma *CNN* original pré-treinada. Ou seja, a rede pré-treinada é submetida a novas imagens e re-treinada para as novas classes na saída. O classificador da saída é trocado visando o número de novas classes pretendidas pela nova tarefa.

D. *Feature Extraction (FE)*

O *FE* usa uma *CNN* pré-treinada para calcular características para uma imagem (Donahue et al, 2014) (Razavian et al, 2014). As características extraídas são as ativações de uma camada (normalmente a última camada oculta) ou de várias camadas, dada a imagem. Classificadores treinados com estas características podem alcançar resultados competitivos, muitas vezes superando o desempenho das características extraídas por métodos convencionais [17]. O *FE* não modifica a rede original e permite que novas tarefas sejam beneficiadas por características complexas aprendidas em tarefas anteriores. No entanto, essas características não são especializadas para a nova tarefa e podem ser melhoradas com *FT*. A Figura 3 mostra um esquema utilizando o *FE* a partir de uma *CNN* original pré-treinada. Neste caso, a *CNN* é usada apenas como um extrator de características das novas imagens, utilizando todos os filtros e parâmetros aprendidos durante o pré-treinamento da rede original. Diferentes classificadores podem ser treinados para o conjunto de novos vetores de características, gerados com a *CNN* pré-treinada, para as novas imagens de entrada.

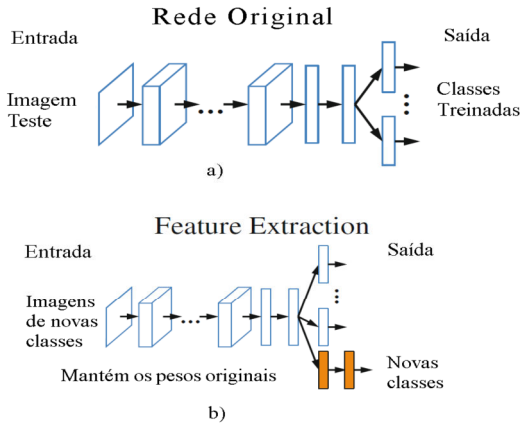

Fig. 3. *Transfer Learning* por *Feature Extraction.* a) *CNN* original pré-treinada. b) Somente o classificador é treinado no *FE.* Adaptado de [4].

E. *Bases de imagens utilizadas*

Duas bases de imagens geradas originalmente para a tarefa de reconhecimento de íris são utilizadas neste trabalho, devido ao fato de possuir nas imagens partes da região periocular e diferentes métodos de amostragem. Uma delas com variação no tamanho da pupila, poucas classes e muitos elementos de classe (quadros de vídeo) e outra adquirida com o equipamento profissional *LG2200*, com mais classes e mesmo elementos de classe.

A base *LAVI DB2* [18] contém 167.965 imagens de íris com parte da periocular apenas do olho direito, de 53 voluntários, adquiridos sob iluminação *NIR* (*Near Infra Red*). As imagens foram capturadas com uma câmera JAI-AD-080GE em diferentes sessões.

Em cada sessão, foi adquirido um vídeo (*.avi*) do olho direito de cada voluntário. As classes possuem de uma a quatro amostras de vídeo. Os quadros de vídeo foram separados individualmente e apresentam variação do tamanho da pupila dentro de cada sequência de vídeo (S1,S2,S3,S4). Todas as imagens são arquivos do tipo *TIFF* em nível de cinza de 8 bits e a resolução da imagem é de 1024x768 pixels. Cada amostra de vídeo possui 1.000 quadros, logo, um indivíduo (classe) com quatro vídeos possui 4.000 imagens do olho direito. Para os experimentos, os conjuntos de treinamento e teste foram selecionados de quatro maneiras diferentes:

1) Modo 1: Os quadros do vídeo da amostra S1 de cada classe formam o conjunto de teste e o restante dos quadros das sequências S2, S3 e S4 formam o conjunto de treinamento. No caso de uma classe possuir apenas a amostra S1, 85% das imagens da amostra vai para o conjunto de treinamento e 15% para o conjunto de teste. Neste modo, o conjunto de teste é formado por um vídeo inteiro, 1000 quadros que nunca foram vistos pela rede durante o treinamento.

2) Modo 2: 85% dos quadros formam o conjunto de treinamento e 15% formam o conjunto de teste, separados aleatoriamente entre todas as amostras. Os quadros próximos entre sí na mesma sequência de vídeo podem apresentar grande similaridade, nas fases de pupila contraída e dilatada, indo formar alguns deles o conjunto de treinamento e outros o conjunto de teste.

3) Modo 3: Inverso do *Modo 2*, 15% dos quadros formam o conjunto de treinamento e 85% formam o conjunto de teste, separados aleatoriamente entre todas as amostras. Visa a redução do número de amostras do conjunto de treinamento.

4) Modo 4: Inverso do *Modo 1*, ou seja, a sequência S1 forma o conjunto de treinamento e as sequências S2, S3 e S4 formam o conjunto de teste. Visa a redução do número de amostras do conjunto de treinamento.

A Figura 4 mostra um exemplo de duas imagens da base *LAVI DB2* de um mesmo indivíduo em duas posições diferentes dentro da amostra de vídeo, uma com a pupila dilatada e outra com a pupila contraída.

A base da Universidade de Notredame (*LG2200*), subconjunto da base *ND-IRIS-0405 Iris Image Dataset*, contém 117.481 imagens de íris dos olho esquerdo e direito, obtidas de 676 indivíduos [19]. As imagens, em nível de cinza, foram capturadas no espectro do infra-vermelho próximo (*NIR*), em ambiente interno com um sensor biométrico para íris *LG2200*. As imagens são de 480x640 pixels de resolução e são armazenadas com 8 bits de intensidade. Como as imagens desta base foram adquiridas por fotos tiradas em épocas diferentes, não existe a mesma

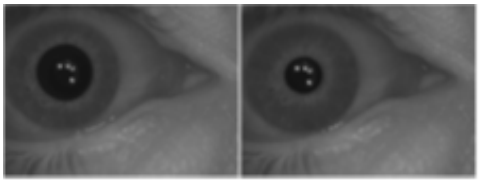

Fig. 4. Exemplos de dois quadros da mesma amostra de periocular da base *LAVI DB2*.

característica da base *LAVI DB2* cujas imagens da mesma classe pertencem a uma mesma sequência de vídeo. No entanto, cada classe em cada época de aquisição, apresenta seis imagens do olho direito seguidas de seis imagens do olho esquerdo. A Figura 5 mostra um exemplo de duas imagens da periocular do mesmo indivíduo da base *LG2200* obtidas em épocas diferentes.

Para os experimentos, os conjuntos de treinamento e teste foram selecionados de três maneiras diferentes:

5) Modo 5: Foram separadas aleatoriamente 85% das imagens para o conjunto de treinamento e 15% das imagens para o conjunto de teste.

6) Modo 6: O conjunto de treinamento foi gerado com as primeiras imagens (85%) sequencialmente, e o conjunto de teste com as 15% restante de cada classe. Isso equivale a

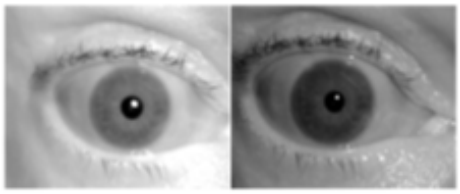

Fig. 5. Exemplo de duas imagens da periocular do mesmo indivíduo da base *ND-IRIS-0402 (LG2200)*

dizer que, para uma classe digitalizada em uma determinada época com 12 imagens, as 10 primeiras imagens (6 do olho direito mais 04 do olho esquerdo) vão para o conjunto de treinamento, e as 02 últimas imagens do olho esquerdo formam o conjunto de teste. Ou seja, o conjunto de teste conterá praticamente apenas imagens do olho esquerdo.

7) Modo 7: Inverso do *Modo 5*, ou seja, 15% das imagens formam o conjunto de treinamento e 85% formam o conjunto de teste, escolhidas aleatoriamente.

F. Método proposto para avaliação da acurácia em cada uma das técnicas

Para avaliação das acurácias das duas técnicas de *Transfer Learning* no reconhecimento biométrico periocular, foram selecionadas as duas *CNNs* pré-treinadas: *AlexNet e VGG-16.* A Figura 6 mostra um diagrama em blocos para o método proposto. Como as imagens das bases foram adquiridas em iluminação *NIR,* as mesmas foram geradas em nível de cinza e 8 bits de resolução. Assim, é apresentada, a cada rede avaliada, a mesma imagem em cada canal de cor *RGB*.

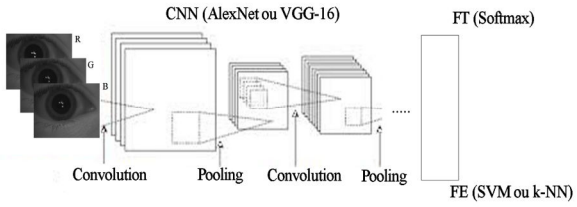

Fig. 6. Diagrama em blocos do método de avaliação proposto

Para a técnica de *FE*, foram escolhidos os classificadores *SVM/ECOC* (*Suport Vector Machine/Error- Correcting Output Codes*) e *k-NN* (*k-Nearest Neighbor)*. Para avaliação das acurácias, os conjuntos de treinamento e teste foram utilizados para ajuste e teste do classificador *SVM*. Para o classificador *k-NN* as imagens do conjunto de teste foram comparadas com as do conjunto de treinamento. A camada da *CNN* determinada para operar como vetor de características foi a *FC7* (penúltima camada completamente conectada) para ambas as redes. Para a técnica de *FT*, manteve-se o classificador *Softmax* original das redes,

alterando-se apenas o número de classes de saída e retreinando-o junto aos parâmetros da rede.

Foram realizados três experimentos.

1) Experimento_1.

Neste experimento foram avaliadas as acurácias, obtidas pelas redes *AlexNet e VGG-16* usando-se *Feature Extraction* na camada *FC7* de cada uma delas, nas duas bases de *imagens* e com os classificadores *SVM* e *k-NN.*

2) Experimento_2.

Neste experimento as duas redes, *AlexNet e VGG-16*, foram treinadas por *Fine Tuning* com as duas bases de imagens da periocular, sendo obtidas suas acurácias.

3) Experimento_3.

Neste experimento foi realizado o mesmo *Fine Tuning* do *Experimento_2* na rede *AlexNet*, na base *LG2200.* Em seguida a camada *FC7* da rede com os pesos ajustados para esta base foi utilizada como *Feature Extraction* para a base *LAVI DB2*, com *SVM* e *k-NN.*

IV. RESULTADOS

Ao se aplicar a técnica de *FE,* a *CNN* opera como um descritor de características. As imagens do conjunto de treinamento geram os vetores de características formados pela camada *FC7* da *CNN*. A acurácia do conjunto de teste em cada modo e classificador é obtida pela média das acurácias por classe dada pela diagonal da Matriz de Confusão. Os resultados de acurácia obtidos para a técnica de *Feature Extraction* nas bases, *LAVI DB2* e *LG2200* podem ser vistos na Tabela I. Como os filtros da rede não foram treinados com imagens da região periocular, é de se esperar que as acurácias não sejam tão elevadas. No entanto, as acurácias obtidas em *FE nos Modos 2 e 3* foram sensivelmente altas. Nestes dois modos, as imagens do treinamento e teste foram escolhidas aleatoriamente, logo, não existe nenhuma sequência de vídeo de teste com imagens similares entre sí e diferentes do treinamento usadas para ajustar o classificador. Além disso, a redução de imagens de treinamento no *Modo 3* não alterou a acurácia do sistema (*CNN* + Classificador). Já a redução de amostras de treinamento no *Modo 4* diminuiu drasticamente a acurácia.

TABLE I. ACURÁCIAS OBTIDAS PELA TÉCNICA DE *FEATURE EXTRACTION (FE)* PARA O EXPERIMENTO 1

FE		*AlexNet*	*VGG-16*
LAVI DB2 Modo 1(%)	*SVM*	*75,14*	*79,81*
	K-NN	*79,98*	*90,24*
LAVI DB2 Modo 2(%)	*SVM*	*99,92*	*99,94*
	K-NN	*89,67*	*99,85*
LAVI DB2 Modo 3(%)	*SVM*	*99,70*	*99,55*
	K-NN	*99,47*	*99,39*
LAVI DB2 Modo 4(%)	*SVM*	*59,49*	*70,67*
	K-NN	*62,75*	*80,50*
(LG2200) Modo 5(%)	*K-NN*	*89,67*	*92,91*
(LG2200) Modo 6(%)	*K-NN*	*72,26*	*78,08*
LG2200 Modo 7(%)	*K-NN*	*64,49*	*69,27*

A Figura 7 mostra a ativação do canal 43 nas camadas *Conv-1 e Conv-5* da rede *AlexNet* para uma imagem da base *LAVI DB2 no Modo 1* quando aplicada a técnica de *FE*. Os pixels brancos representam fortes ativações positivas e os pixels pretos representam fortes ativações negativas. A posição de um pixel na ativação de um canal corresponde à mesma posição na imagem original. Um pixel branco em algum local em um canal indica que o canal está fortemente ativado nessa posição.

Em uma *CNN,* os canais nas camadas anteriores aprendem recursos simples como cor e bordas, enquanto os canais nas camadas mais profundas aprendem recursos complexos. O que se pode observar é que, como a *AlexNet* em *FE* não foi treinada originalmente com imagens da região periocular, as ativações de uma camada mais rasa (Figura 7b) são maiores dentro da pupila, onde existe a reflexão da iluminação *NIR* gerada pelo sistema de aquisição da imagem,

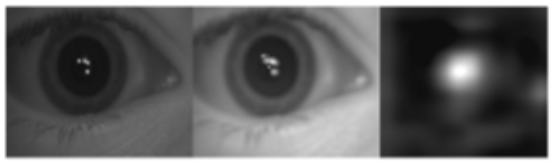

Fig. 7. Ativação da rede *AlexNet* no canal 43 em *FE*. a) Imagem de entrada. b) Camada *Conv-1*. c) Camada *Conv-5*

mantendo esta tendência na camada mais profunda (Figura 7b) responsável por detectar características mais complexas da imagem.

Para a técnica de *Fine Tuning*, pode-se observar na Tabela II que, como as imagens da base *LG2200* são imagens adquiridas em épocas diferentes, apresentando variações intraclasse que podem ser "aprendidas pela rede" e o método de separação de conjuntos de treinamento e teste foi aleatório, os resultados das acurácias ultrapassam 99%. O mesmo não ocorre com as acurácias obtidas pelas duas redes para a base *LAVI DB2 em Modo 1.* O fato se explica pela maneira como os conjuntos de treinamento e teste foram divididos no *Modo 1,* ou seja, tomando-se para o teste uma sequência inteira de vídeo, as características deste vídeo nunca foram aprendidas pela *CNN*. Além disso, a grande quantidade de amostras similares para uma mesma classe, dadas pelos quadros próximos dentro da sequência, que apresentam poucas diferenças entre si, são aprendidas pela rede em contraposição às amostras do conjunto de teste.

TABLE II. ACURÁCIA DA TÉCNICA DE *FINE TUNING (FT)* PARA O EXPERIMENTO 2

Fine Tuning	*LAVI DB2 em Modo 1 (%)*	*LG2200 (%)*
AlexNet	*78,84*	*99,20*
VGG-16	*87,43*	*99,04*

A Figura 8 mostra a ativação do canal 43 nas camadas *Conv-1 e Conv-5* da rede *AlexNet* para uma imagem da base *LAVI DB2 no Modo 1* quando aplicada a técnica de *FT*.

Fig. 8. Ativação da rede *AlexNet* no canal 43 em *FT*. a) Imagem de entrada. b) Camada *Conv-1*. c) Camada *Conv-5*.

Pode-se observar que devido ao treinamento da rede com imagens da região periocular do conjunto de treinamento, a camada mais rasa (Figura 8b) apresenta ativação positiva de partes da textura da íris, e a pupila iluminada na camada mais profunda (Figura 8c) tem ativação fortemente negativa.

As classes da base *LG2200* são também desbalanceadas tendo o mínimo de 24 e o máximo de 636 elementos por

classe. Este desbalanceamento gera variação nas acurácias obtidas por classe. Em geral, os menores valores de acurácia referem-se às classes com o menor número de elementos. Este resultado está intimamente ligado ao método de separação dos conjuntos de treinamento e teste em classes desbalanceadas o que contribui para redução da curácia total.

O *Experimento_3* utiliza o *Fine Tuning* na rede *AlexNet* sobre a base *LG2200* dada pelo *Experimento_2*. A camada *FC7* da rede foi utilizada como extrator de característica *(FE)* na base *LAVI DB2*. A Tabela 3 mostra as acurácias obtidas. Como pode ser observado, não houve ganho de acurácia, demonstrando que a rede precisa ser treinada com elementos das mesmas classes a serem reconhecidas para que seus filtros ativem positivamente nas características principais.

TABLE III. *FE* NA BASE *LAVI DB2* DEPOIS DE *FT* NA BASE *LG2200* PARA O EXPERIMENTO 3

FE		*AlexNet (FT na LG2200)*	*AlexNet (ImageNet)*
LAVI DB2 Modo 1 (%)	*SVM*	*74,03*	*75,14*
	K-NN	*79,07*	*79,98*
LAVI DB2 Modo 2 (%)	*SVM*	*99,69*	*99,92*
	K-NN	*99,83*	*89,67*
LAVI DB2 Modo 3 (%)	*SVM*	*99,43*	*99,70*
	K-NN	*99,33*	*99,47*
LAVI DB2 Modo 4 (%)	*SVM*	*59,58*	*59,49*
	K-NN	*67,11*	*62,75*

V. CONCLUSÕES

Quando o conjunto de imagens a serem classificadas por uma *CNN* não é grande o suficiente para se treinar uma rede a partir do zero, as técnicas de *Transfer Learning* permitem atingir resultados excelentes. Neste trabalho foram investigadas as diferenças nas acurácias de duas destas técnicas chamadas de *Fine Tuning e Feature Extraction* para o reconhecimento biométrico periocular. Duas bases de imagens com características diferentes no modo de aquisição, no número de classes e elementos de classes, foram utilizadas. Duas redes, *AlexNet e VGG-16* foram avaliadas quanto a suas capacidades de reconhecerem a região periocular das imagens destas bases. O que se conclue é que mesmo com a técnica de *FT* que re-treina a *CNN* com imagens novas, mas a partir de pesos "aprendidos" no treinamento original, a separação dos conjuntos de treinamento e teste exerce papel preponderante na acurácia final. O resultado obtido com *FT*, usando-se uma única sequência de vídeo nunca vista pela *CNN* como conjunto de teste apresentou acurácia semelhante à obtida com a técnica de *FE*, na qual a rede é usada apenas como extrator de características sem nunca ter sido treinada com nenhuma das imagens de periocular. Outra conclusão refere-se à qualidade dos dados. Uma divisão aleatória entre treinamento e teste em classes desbalanceadas pode não garantir a presença de representantes de todas as classes nos dois conjuntos. Isso tende a reduzir a acurácia total devido a classes com poucos elementos não estarem representadas nos dois conjuntos. Isso ocorreu com a base *LG2200,* que apesar de ter atingido mais de 99% de acurácia, teve diversas classes com menos de 86%. Além disso, o fato de se treinar com *FT* uma *CNN* com imagens de periocular de uma base diferente da base usada como teste, não melhora a acurácia em *FE* considerando o treinamento original da *CNN*.

AGRADECIMENTOS.

Os autores agradecem à NVIDIA pela doação de uma GPU Titan XP para a realização deste trabalho e à CAPES pelo apoio financeiro.

REFERÊNCIAS.

1. Park, U.; Ross, A.; Jain, A.; Periocular biometrics in the visible spectrum: a feasibility study. In: IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS '09, pp. 1–6.

2. Park, U.; Jillela, R. R.;, Ross, A.; Jain, A. K.; Periocular biometrics in the visible spectrum. IEEE Trans on Information Forensics Security, vol. 6, no. 1, march 2011, pp. 96–106.

3. Ambika, D.; Radhika, K.; Seshachalam, D.; The eye says it all: periocular region methodologies, in: International Conference on Multimedia Computing and Systems (ICMCS), 2012, pp. 180–185 .

4. Li, Z.; Hoiem, D.; Learning Without Forgetting, B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, Springer International Publishing AG 2016, pp. 614–629, 2016. DOI: 10.1007/978-3-319-46493-0 37

5. Woodard, D.L., Pundlik, S.J., Lyle, J.R., Miller, P.E.; Periocular region appearance cues for biometric identification. Proc IEEE Computer Vision and Pattern Recognition Biometrics Workshop, 2010, CVPRW.

6. Tan, C.W., Kumar, A.; Human identification from at-a-distance images by simultaneously exploiting iris and periocular features. Proc Intl Conf Pattern Recognition, ICPR, 2012, pp. 553–556.

7. Karahan, S., Karaoz, A., Ozdemir, O., Gu, A.; Uludag, U.; On identification from periocular region utilizing SIFT and SURF. Proc European Signal Processing Conf, EUSIPCO, 2014, pp. 1392–1396.

8. Alonso-Fernandez, F.; Bigun, J.; A survey on periocular biometrics research," Pattern Recognit. Lett., vol. 82, pp. 92–105, Oct. 2016.

9. Rattani, A; Derakhshani, R.; Ocular biometrics in the visible spectrum: A survey, Image and Vision Computing, Elsevier, Volume 59, March 2017, Pages 1-16.

10. Proença, H.; Neves, J.C.; Deep-PRWIS: Periocular Recognition Without the Iris and Sclera Using Deep Learning Frameworks, IEEE Transactions On Information Forensics And Security, Vol. 13, No. 4, April 2018, pp. 888-896.

11. Zhao, Z.; Kumar, A.; Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network, IEEE Trans. Inf. Forensics Security, vol. 12, no. 5, pp. 1017–1030, May 2016, doi: 10.1109/TIFS.2016.2636093.

12. Zhao, Z.; Kumar, A.; Improving Periocular Recognition by Explicit Attention to Critical Regions in Deep Neural Network, IEEE Transactions On Information Forensics And Security, Vol. 13, N0. 12, December 2018, pp. 2937-2952.

13. Lecun, Y. ; Bottou, L.; Bengio,Y.; Haffner, P.; Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol: 86, Issue: 11, November 1988, pp. 2278 – 2324.

14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E.; ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*. 2012.

15. Simonyan, K.; Zisserman, A.; Very Deep Convolutional Networks For Large-Scale Image Recognition, ICLR 2015, **arXiv:1409.1556 [cs.CV]**

16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J.; Rich feature hierarchies for accurateobject detection and semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

17. Donahue, J.; Jia, Y.; Vinyals, O., Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T.; DeCAF: a deep convolutional activation feature for generic visual recognition. In: Int. Conf. in Machine Learning (ICML) (2014).

18. http://imagem.sel.eesc.usp.br/base/iris/Gallery/index-2.html.

19. Phillips, P.J.; Bowyer, K.W.; Flynn, P.J.; Liu, X.; Scruggs, W.T.; The Iris Challenge Evaluation 2005, 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, 29 Sept.-1 Oct.