Periocular authentication in smartphones applying uLBP descriptor on CNN Feature Maps

William Barcellos
University of São Paulo
São Carlos, Brazil
william.barcellos@usp.br

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

***Abstract*— The outputs of CNN layers, called Activations, are composed of Feature Maps, which show textural information that can be extracted by a texture descriptor. Standard CNN feature extraction use Activations as feature vectors for object recognition. The goal of this work is to evaluate a new methodology of CNN feature extraction. In this paper, instead of using the Activations as a feature vector, we use a CNN as a feature extractor, and then we apply a texture descriptor directly on the Feature Maps. Thus, we use the extracted features obtained by the texture descriptor as a feature vector for authentication. To evaluate our proposed method, we use the AlexNet CNN previously trained on the ImageNet database as a feature extractor; then we apply the uniform LBP (uLBP) descriptor on the Feature Maps for texture extraction. We tested our proposed method on the VISOB dataset composed of periocular images taken from 3 different smartphones under 3 different lighting conditions. Our results show that the use of a texture descriptor on CNN Feature Maps achieves better performance than computer vision handcrafted methods or even by standard CNN feature extraction.**

Keywords— CNN, Feature Maps, smartphone, periocular authentication, AlexNet, VISOB

I. INTRODUÇÃO

Os *smartphones* estão cada vez mais presentes na vida das pessoas. Com a evolução dos processadores e do aumento da capacidade de memória, os *smartphones* têm substituido os computadores pessoais nas mais diversas atividades como envio de e-mails, acesso à conta bancária, transferência de dinheiro, acesso a redes sociais, compras *online*, comunicação por mensagens, entre outras. Devido às diversas utilizações, os *smartphones* tendem a armazenar fotos e documentos pessoais, além de informações sensíveis, sendo então necessário protege-lo contra a sua utilização por terceiros. Os primeiros métodos de autenticação utilizados em *smartphones* foram o PIN, padrão de desenho e senha alfanumérica, e mais recentemente tem se popularizado a utilização de impressão digital e reconhecimento facial [1].

Embora o reconhecimento pessoal por meio de características biométricas da face seja estudado há muito tempo, e já seja utilizado em diversas situações, a eficiência dos sistemas que utilizam reconhecimento facial é afetada por fatores como o envelhecimento, variação de pose, efeitos de iluminação, expressão facial e oclusão parcial do rosto [2]. Para contornar estes problemas, uma alternativa é a utilização da região periocular como uma peculiaridade biométrica.

A região periocular pode ser definida como o local da face nas imediações do olho, normalmente englobando as pálpebras, cílios, sobrancelhas e a área de pele vizinha. Algumas caracterísicas biométricas são comummente extraídas, tais como, o formato das pálpebras, o formato do olho, o formato da sobrancelha, a distribuição dos cílios, a textura e a cor da pele. Usualmente na mesma imagem em que se encontra a região periocular, está presente o olho, composto pela pupila, íris e esclera, que também podem ser utilizadas como peculiaridades para extração de características biométricas. A Figura 1 detalha as principais peculiaridades e características biométricas que podem ser extraídas desta região da face humana.

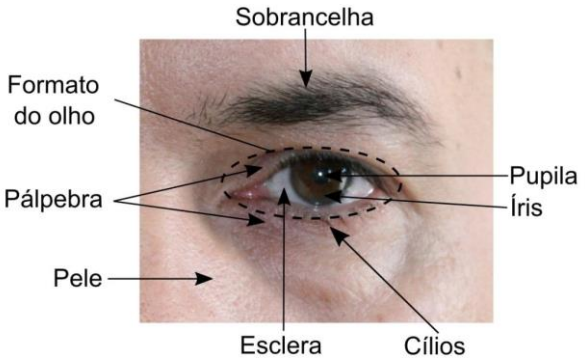

Fig. 1. Peculiaridades e características biométricas da região periocular e do olho.

O reconhecimento facial pode ser realizado utilizando apenas a região periocular, mas também pode utilizar a fusão de características biométricas perioculares com características biométricas globais da face [3].

Além de ser um alternativa tanto de substituição como de melhoria para o reconhecimento da face, a região periocular também têm se mostrado mais robusta com relação à variação de expressão, transformação de gênero e envelhecimento [4] [5] [6].

Em aplicações no reconhecimento biométrico as redes neurais convolucionais (CNN) tem gerado soluções com alta taxa de acerto. As CNNs foram introduzidas na década de 90 com o desenvolvimento da LeNet [7], capaz de reconhecer digitos escritos à mão. Porém, por muito tempo as CNNs não foram utilizadas porque a limitada capacidade computacional e de memória dificultavam a implementação dos algoritmos.

Com a vitória da CNN AlexNet [8] na *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) [9] de 2012, atingindo taxa de erro de 16% (contra 26% do vencedor do ano anterior) e o uso de GPUs (*Graphical Processing Units*) as CNNs começaram a ganhar popularidade. Atualmente as CNNs são consideradas o estado da arte em diversas áreas como classificação de imagens [10], detecção de objetos [11], segmentação [12] e reconhecimento de cena [13].

Apesar de serem extremamente eficientes, as CNNs dependem de base de dados com grande número de amostras por classe para serem treinadas, como a ImageNet, que é composta de 1000 classes contendo em média 1000 imagens por classe. No entanto, em grande parte dos problemas em visão computacional, o número de amostras disponíveis, para uma determinada aplicação, não é suficiente para o treinamento de uma CNN a partir do zero. Nestes casos,

redes previamente treinadas em bases grandes e multiclasses, chamadas *off-the-shelf*, podem ser utilizadas como extratores de características [14].

Aplicando uma CNN *off-the-shelf* como extrator de características, as Ativações são diretamente utilizadas como vetores de características.

As Ativações das últimas camadas de uma CNN são vetores, porém, nas camadas intermediárias, as Ativações são volumes 3D, sendo necessário realizar algum tipo de operação para transformá-las em vetores.

As Ativações das camadas intermediárias são compostas de Mapas de Características (*Feature Maps*) de duas dimensões, que podem ser interpretadas como imagens derivadas da imagem de entrada da rede, que foi submetida aos filtros da CNN. Estas imagens dos Mapas de Características apresentam uma estrutura de *pixels* que pode ser processada diretamente por descritores de textura.

Neste trabalho propomos extrair características dos *Feature Maps* das camadas de uma CNN *off-the-shelf* utilizando o descritor de texturas uLBP (*uniform Local Binary Patterns*). A aplicação principal é na autenticação pessoal com celulares *smartphones* adquirindo imagens da região periocular. Para avaliar nosso método, foi utilizada a base de dados VISOB [19], composta de imagens da região periocular adquiridas pela câmera frontal de três *smartphones* diferentes, em três tipos de iluminação. As imagens são processadas pela CNN AlexNet pré-treinada na ImageNet; o descritor de texturas uLBP é aplicado nos Mapas de Características das camadas intermediárias da CNN para geração dos vetores de características utilizados no processo de autenticação.

O método proposto foi comparado com métodos tradicionais (*handcrafted*) de Visão Computacional tais como LBP (*Local Binary Pattern*), HOG (*Histogram of Oriented Gradients*), SURF (S*peeded Up Robust Features*) e SIFT (*Scale Invariant Feature Transform*). Também foi comparada com a utilização das Ativações de uma CNN *off-the-shelf* diretamente como vetor de características, que é a maneira usual de se utilizar uma CNN como extrator de características.

II. TRABALHOS CORRELATOS

Com a evolução do poder de processamento e da qualidade das câmeras dos *smartphones*, imagens do rosto tiradas por celulares, as chamadas *selfies*, começaram a ser estudadas como uma alternativa para autenticação. Neste contexto, a região periocular é sempre considerada por fazer parte deste tipo de imagem.

Raja et al. [15] propuseram um sistema de autenticação para *smartphones Android* utilizando a região periocular. As características foram extraídas utilizando SIFT, SURF e BSIF (*Binarized Statistical Image Features*). Os *scores* foram calculados utilizando FLANN (*Fast Library for Approximate Nearest Neighbors*) quando da extração de características por SIFT e SURF, e utilizando distância Bhattacharya quando da extração de características por BSIF.

Em [16] Raja et al. utilizaram a técnica de *Deep Sparse Filtering* para aprender 256 filtros de 16x16. A extração de características utilizou convolução das imagens com os filtros e *Collaborative Representation* para realizar a comparação entre os *scores*. O método foi avaliado na base de dados VISOB, obtendo melhores resultados quando comparado com as técnicas BSIF, HOG e LPQ *(Local Phase Quantization*).

Raghavendra e Busch [17] utilizaram filtros MR (*Maximum Response filters*) para extração de características, e uma rede neural baseada em *deeply coupled autoencoders* para classificação. O método foi avaliado na base de dados VISOB, obtendo melhores resultados quando comparado com as técnicas BSIF, HOG e LPQ.

Ahuja et al. [18] propuseram um sistema de verificação em 2 estágios. No primeiro foi utilizado SURF para extração de características e um classificador *Naive Bayes*. No segundo estágio, foi utilizado *Dense* SIFT nos *top-5* do classificador, fornecendo os *scores* para classificação final na base de dados VISOB.

Alguns trabalhos usam fusão de informações da face, íris e região periocular, para aumentar a eficiência. Raja et al. [20] propuseram a utilização da face, iris e região periocular para autenticação em *smartphones*. Utilizaram SIFT, SURF e BSIF para extrair características da face e da região periocular, e para íris foi utilizado *IrisCode*. Para cada traço biométrico, os *scores* foram calculados para cada extrator de características, e então realizada a fusão com pesos (*weighted fusion*). Posteriormente foi feita a fusão ponderada dos três traços biométricos. O método foi avaliado em um banco de imagens próprio, composto de 78 usuários e imagens obtidas de 2 *smartphones*.

Santos et al. [21] propuseram a fusão de características da iris e da região periocular. As características foram extraídas utilizando LBP, HOG, uLBP, SIFT, *Gist-of-the-scene* e *IrisCode*. Os *scores* foram calculados para cada extrator e então foi realizada a fusão dos *scores* utilizando uma rede neural. Utilizaram a base de dados CSIP (*Cross-sensor iris and periocular dataset*), composta por 50 usuários e imagens obtidas de 4 *smartphones*.

Ahmed et al. [22] utilizaram o *IrisCode* para extração de características da íris, e MB-TLBP (*Multi-Block Transitional Local Binary Patterns*) para extração de características da região periocular. A utilização da íris se mostrou superior à região periocular. A fusão de *scores* apresentou uma pequena melhora nos resultados na base MICHE-II [31].

Fernandez et al. [23] avaliaram a fusão de *scores* considerando a comparação entre imagens adquiridas com o mesmo modelo de *smartphone* e entre imagens adquiridas com *smartphones* diferentes. Foram utilizados os extratores de características SAFE (*Symmetry Assessment by Feature Expansion*), *Gabor*, SIFT, LBP e HOG. A técnica foi avaliada na base de dados VSSIRIS, composta de 28 ususários e imagens obtidas de 2 *smartphones*.

Stokkenes et al. [24] extrairam características da região periocular do olho esquerdo, direito e da face, utilizando BSIF realizando fusão das características por concatenação, obtendo melhores resultados. O método foi avaliado em uma base de imagens própria, composta de 78 usuários e imagens obtidas de um *smartphone*.

Aginako et al. [25] fizeram uma avaliação da utilização de vários descritores, em diferentes situações, para reconhecimento utilizando a íris e a região periocular. Os descritores testados foram HOG, LBP, uLBP, LGP (*Local Gradient Patterns*), LTP (*Local Ternary Patterns*), LSP (*Local Similarity Pattern*), WLD (*Weber Local Descriptor*), LPQ, NILBP (*Intensity based Local Binary Patterns*) e LOSIB (*Local Oriented Statistics Information Booster*).

Tanto as imagens de íris como da região periocular foram divididas em grades de 2x2, 3x3 e 4x4, e os descritores aplicados nas células destas grades. Foram utilizados os classificadores *kNN*, *Bagging*, *Random Forest*, *Naive Bayes* e *C4.5*. As diferentes combinações de descritor-grade-classificador foram avaliadas na base de dados MICHE-II. Foi avaliada a fusão das combinações *top-3* na região periocular, obtendo uma pequena melhora de desempenho. As combinações *top-5* também foram avaliadas, mas usando a base de dados VISOB.

Trabalhos mais recentes têm se concentrado no uso de redes convolucionais, visto que estas redes têm se mostrado como o estado da arte nas mais diversas aplicações.

Rattani e Derakhshani [26] re-treinaram (*fine tunning*) 4 CNNs *off-the-shelf* utilizando a base de dados VISOB. Para comparação, uma nova rede foi treinada do zero (*from scratch*). As redes re-treinadas obtiveram melhores resultados.

Ahuja et al. [27] propuseram dois modelos, um supervisionado e outro não supervisionado. O modelo 1, não supervisionado, utiliza a fusão dos *scores* de três técnicas: *Root*SIFT, utilizada para verificação por meio da íris; *Openface*, utilizada para verificação por meio da face; e VisobNet, uma CNN proposta pelos autores e treinada na base de dados VISOB. O Modelo 2, supervisionado, é uma CNN proposta pelo autor, treinada na base de dados MICHE-II. Ambos os modelos são avaliados na MICHE-II. O modelo 1 é considerado não supervisionado porque não foi treinado no banco de dados alvo, já o modelo 2 é considerado supervisionado porque foi treinado no banco de dados alvo. Os resultados mostram que o modelo supervisionado é mais eficiente.

Reddy et al. [28] fizeram um estudo avaliando a utilização de 9 CNNs, sendo uma delas proposta pelos autores, no processo de autenticação por *smartphone*. As CNNs foram utilizadas como extrator de carcaterísticas, sendo que as características foram extraídas na penúltima camada de cada CNN. Os *scores* foram calculados por similaridade usando a distância cosseno na base de dados VISOB.

Rattani et al. [29] propuseram uma fusão nas camadas intermediárias de uma CNN. Para isso, é proposta uma CNN possuindo duas entradas, sendo duas ramificações que processam as duas entradas em paralelo, e as camadas totalmente conectadas desses dois ramos são concatenadas antes da camada de classificação. Foi utilizada a base de dados VISOB.

Kondapi et al. [30] avaliaram a fusão de características extraídas por descritores *handcrafted* (LBP e HOG) e fusão de características extraídas por diferentes CNNs, além da fusão de características na região do olho esquerdo, com a do olho direito, com a face. Também foi avaliada a comparação entre imagens com diferentes tipos de iluminação na base de dados VISOB.

Diferente das abordagens publicadas e citadas neste item, nossa proposta utiliza um extrator de características (descritor de textura) nos Mapas de Características das camadas intermediárias de uma CNN, ao invés de aplicar um descritor diretamente nas imagens (método *handcrafted*), ou utilizar uma CNN como o único extrator de características para autenticação.

III. Material e Método

A. Base de imagens VISOB

A base de imagens *Visible Light Mobile Ocular Biometric* (VISOB) é composta por imagens de olho de 586 indivíduos, adquiridas usando três *smartphones* diferentes: um *iPhone 5S*, um *Samsung Note 4* e um *Oppo N1*. Os indivíduos se apresentaram em duas sessões, separadas por 10 a 15 minutos, onde foram instruidos a tirar *selfies* (foto do próprio rosto) com a câmera frontal do *smartphone*. Os usuários utilizaram o *smartphone* naturalmente, segurando este a uma distância de 20 a 30 centímetros da face. As fotos foram tiradas em três iluminações diferentes: *Office Light* (luz de escritório), *Dim Light* (luz fraca) e *Day Light* (luz diurna natural).

As imagens da sessão 1 foram consideradas como conjunto de treinamento, e da sessão 2 como conjunto de teste. Existem indivíduos presentes nas duas sessões, presentes apenas na sessão 1 e apenas na sessão 2. Os indivíduos presentes apenas no conjunto de teste (sessão 2) foram considerados impostores, e os demais genuinos. A Figura 2 apresenta exemplos da região periocular de alguns indivíduos da base de imagens VISOB.

Fig. 2. Região periocular de indivíduos da base de imagens VISOB.

A Tabela I apresenta o número de imagens utilizadas no treinamento e teste para cada smartphone em cada condição de iluminação.

TABELA I. Número de imagens de treinamento e teste para cada smartphone em cada condição de iluminação

		Treinamento	Teste
iPhone	Day Light	5270	5103
	Dim Light	3762	3552
	Office Light	5045	4553
Oppo	Day Light	7896	3926
	Dim Light	7497	7482
	Office Light	10553	9897
Samsung	Day Light	3230	3265
	Dim Light	4294	4182
	Office Light	4673	4792

B. CNN AlexNet off-the shelf.

As Redes Neurais Convolucionais (CNN) são compostas de camadas que processam as informações fornecidas na entrada da rede, e fornecem uma classificação na sua saída. As camadas são compostas de filtros, cujos pesos são ajustados por treinamento. As principais camadas de uma CNN são: a camada de entrada, as camadas Convolucionais (ou camadas de convolução), as camadas de *Pooling*, as

camadas totalmente conectadas, e a camada de saída. A arquitetura básica de uma CNN é mostrada na Figura 3.

Ao passar por uma camada, as Ativações da camada anterior são processadas pelos filtros desta camada, gerando um mapa de características por filtro, compondo assim as Ativações desta camada. A Figura 4 apresenta este processo.

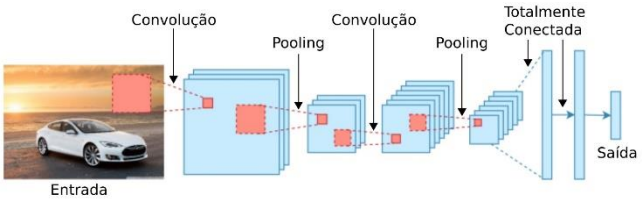

Fig. 3. Arquitetura básica de uma CNN. (Adaptado de [32])

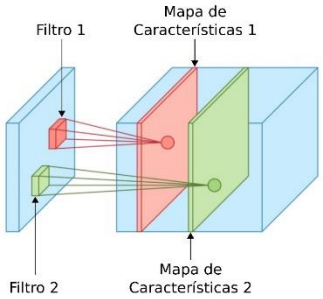

Fig. 4. Aplicação de dois filtro em uma camada. (Adaptado de [32])

A CNN *AlexNet* é composta de 25 camadas, sendo a camada 1 a camada de *Input*, na qual é aplicada a imagem que se deseja classificar, e a camada 25 a camada de *Output*, que fornece a classificação da imagem. As camadas intermediárias realizam o processamento da imagem, sendo que 5 delas são camadas convolucionais, 3 são camadas de *Pooling*, e 3 são camadas totalmente conectadas.

A CNN *AlexNet* foi originalmente pré-treinada na base de imagens *ImageNet*, composta de 1000 classes que contêm em média 1000 imagens cada, totalizando 1.461.406 imagens.

Analisando-se as Ativações das camadas intermediárias de uma CNN *off-the-shelf*, foi observado que os Mapas de Características, que compõem estas Ativações, possuem características de texturas que podem ser extraídas com um descritor. Dada esta observação, foi proposto neste trabalho a utilização dessas características de textura extraídas dos *Feature Maps* de uma CNN *off-the-shelf* para a autenticação em *smartphones*.

C. *Método Proposto.*

Os Mapas de Características de uma CNN podem ser interpretados como uma imagem em níveis de cinza. Por exemplo, ao se aplicar na CNN *AlexNet* a imagem da região periocular da Figura 5, extrai-se da camada *Conv2* os 64 Mapas de Características mostrados na Figura 6.

Fig. 5. Exemplo de imagem da região periocular.

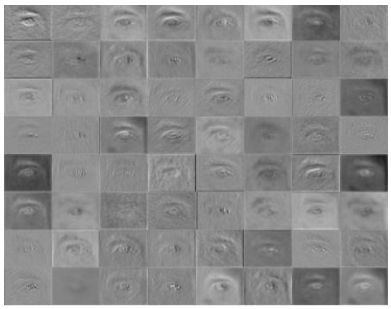

Fig. 6. Mapas de Características da camada *Conv2* da CNN AlexNet, da imagem de periocular dada na Figura 5.

Observa-se que cada mapa de características ressalta diferentes informações da região periocular. Por este motivo foi proposto a aplicação de um descritor nos Mapas de Características, de forma a extrair estas informação para serem utilizadas no processo de autenticação.

Foram propostas duas variações para o método proposto, denominadas FE-LBP-FM-Conc e FE-LBP-FM+Norm-Conc. Na variação FE-LBP-FM-Conc os Mapas de Características são extraídos em uma camada específica. O descritor uLBP é aplicado em cada mapa de características e então os histogramas uLBP são concatenados, formando assim um vetor de características. A Figura 7 exemplifica o método proposto.

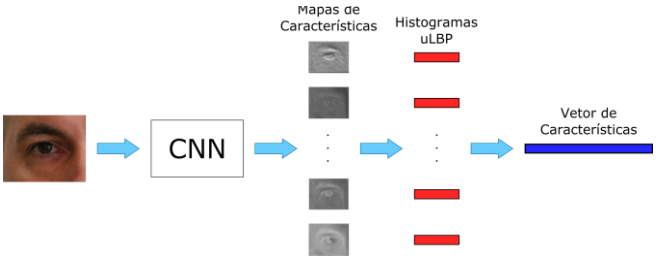

Fig. 7. Método proposto para extração de características utilizando uma CNN.

Utilizando-se os vetores de características, calcula-se a distância Euclidiana entre a amostra e todos os elementos da suposta classe. O *score* é definido como a menor distância encontrada. A diferença da variação FE-LBP-FM+Norm-Conc é que o descritor uLBP é aplicado em cada mapa de características após normalização. Isso foi proposto após verificação de grande variação nos valores numéricos das características.

O desempenho do método proposto foi comparado com extração de características utilizando CNNs (sem a aplicação de descritor nos Mapas de Características), aqui denominado FE-raw. Foi também comparado com o desempenho dos métodos tradicionais em Visão Computacional uLBP, HOG, SURF e SIFT.

Na técnica FE-raw os Mapas de Características são extraídos em uma camada específica. As linhas de cada mapa são concatenadas e então os mapas são concatenados, formando assim um vetor de características.

Para comparação de performance, a imagem original é subdividida em 9 células. O histograma uLBP é extraído em cada célula e então os histogramas são concatenados, gerando o vetor de características. Da mesma maneira anterior, o *score* é definido como a menor distância Euclidiana encontrada.

Aplica-se também o método HOG na imagem original extraindo-se o vetor de características. Calcula-se a distância

Euclidiana entre a amostra e todos os elementos da suposta classe. O *score* é definido como a menor distância encontrada.

Em cada imagem original, os SURF *points* são detectados, e então um vetor de caraterísticas é extraído em cada ponto. Os vetores de características da amostra são comparados com os vetores de características de cada elemento da suposta classe. Os vetores de características similares são encontrados. É calculada a distância Euclidiana entre os SURF *points* com vetores de características similares, e então é feita a média destas distâncias. O *score* é dado pelo menor valor de média de distâncias.

Em cada imagem original os SIFT *frames* são detectados, assim como o vetor de característica de cada SIFT *frame*. Os vetores de características da amostra são comparados com os vetores de características de cada elemento da suposta classe. Os vetores de características similares são encontrados. É calculada a distância Euclidiana entre os SIFT *frames* com vetores de características similares, e então é feita a média destas distâncias. O *score* é dado pelo menor valor de média de distâncias.

Para se comparar os métodos, os scores dos genuinos e impostores foram utilizados para determinar a Taxa de Erro Igual (*Equal Error Rate* - EER), que é o ponto de operação do sistema onde a Taxa de Falsa Rejeição (*False Rejection Rate* - FRR) é igual à Taxa de Taxa de Falsa Aceitação (*False Acceptance Rate* - FAR).

IV. Resultados

Utilizando os vetores de características extraídos dos Mapas de Características de cada camada, foram calculados os *scores* genuinos e impostores e calculado o EER em cada camada. O mesmo processo foi realizado com o método tradicional de extração de características utilizando apenas a CNN (FE-raw). A Tabela II apresenta o menor EER obtido e a camada correspondente, nos diferentes *smartphones* e condições de iluminação da base de dados VISOB, para cada uma das variações do método proposto. Os valores em **negrito** mostram o melhor resultado obtido em cada linha.

Observa-se que não é possível definir uma camada que seja a mais eficiente para extração de características, visto que diferentes camadas fornecem o menor EER para diferentes combinações de *smartphone* e condição de iluminação. Mas, o método proposto FE-LBP-FM+Norm-Conc obteve os melhores resultados em 16 dos 18 experimentos realizados.

TABELA II. Menor EER e camada da CNN AlexNet correspondente, nos diferentes smartphones e condições de iluminação da base de dados VISOB

			FE-raw		FE-LBP-FM-Conc		FE-LBP-FM+Norm-Conc	
			EER [%]	Camada	EER [%]	Camada	EER [%]	Camada
iPhone	Day Light	Left	5,84	conv5	7,76	relu4	**3,37**	conv3
		Right	6,47	conv3	8,22	pool2	**3,96**	conv5
	Dim Light	Left	**5,54**	conv5	10,35	relu4	5,94	conv5
		Right	**5,59**	conv5	12,23	relu3	5,98	conv2
	Office Light	Left	7,78	conv5	9,77	relu3	**5,78**	conv4
		Right	8,93	conv5	11,35	relu3	**6,38**	conv5
Oppo	Day Light	Left	10,03	conv5	11,56	pool2	**9,77**	conv5
		Right	9,91	pool2	12,30	pool2	**9,34**	conv2
	Dim Light	Left	4,57	conv3	7,38	pool2	**4,09**	conv3
		Right	5,13	conv3	8,62	pool2	**4,00**	conv3
	Office Light	Left	9,29	conv5	11,35	relu4	**7,63**	conv2
		Right	11,91	drop6	13,36	relu3	**9,55**	conv3
Samsung	Day Light	Left	9,46	conv5	11,53	relu3	**7,09**	conv4
		Right	7,87	conv5	11,32	relu3	**5,59**	conv3
	Dim Light	Left	6,99	conv5	8,15	relu3	**4,07**	conv5
		Right	5,54	conv5	6,73	relu2	**2,71**	conv4
	Office Light	Left	16,71	conv5	13,37	relu3	**7,17**	conv5
		Right	13,31	conv5	12,88	relu2	**5,61**	conv2

TABELA III. Avaliação de desempenho de cada método comparado nos diferentes smartphones e condições de iluminação da base de dados VISOB

EER [%]																		
Phone	***iPhone***		***Oppo***		***Samsung***		***iPhone***		***Oppo***		***Samsung***		***iPhone***		***Oppo***		***Samsung***	
Eye Side	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***	***Left***	***Right***
Light	***Day Light***						***Dim Light***						***Office Light***					
uLBP	30,65	29,08	50,73	42,91	19,85	29,33	53,22	48,91	25,80	27,32	27,45	27,74	41,66	35,54	32,15	30,19	26,72	30,44
HOG	30,02	29,63	47,43	43,97	27,27	33,08	44,06	44,04	31,67	30,69	31,97	36,62	30,89	33,70	33,18	32,11	30,61	40,34
SURF	18,09	18,52	19,40	21,29	20,68	20,60	17,01	16,66	19,34	17,71	22,31	21,62	26,45	25,90	27,72	27,84	27,16	27,22
SIFT	7,79	8,50	**9,26**	10,78	8,62	7,22	7,93	7,68	7,73	7,35	8,32	8,94	8,31	9,86	14,37	15,69	12,41	10,08
FE-raw[a]	5,84	6,47	10,03	9,91	9,46	7,87	**5,54**	**5,59**	4,57	5,13	6,99	5,54	7,78	8,93	9,29	11,91	16,71	13,31
FE-LBP-FM-Conc[a]	7,76	8,22	11,56	12,30	11,53	11,32	10,35	12,23	7,38	8,62	8,15	6,73	9,77	11,35	11,35	13,36	13,37	12,88
FE-LBP-FM+Norm-Conc[a]	**3,37**	**3,96**	9,77	**9,34**	**7,09**	**5,59**	5,94	5,98	**4,09**	**4,00**	**4,07**	**2,71**	**5,78**	**6,38**	**7,63**	**9,55**	**7,17**	**5,61**

[a.] Resultado da melhor camada

Outra observação importante é que usualmente, no método tradicional de extração de características utilizando CNN (FE-raw), os vetores de caracterísicas são extraídos das camadas totalmente conectadas, porém, em apenas um caso (*Oppo/Office Light/Right*) uma camada totalmente conectada (*drop6*) apresentou o menor EER, o que pode indicar que as camadas totalmente conectadas não sejam as melhores camadas de extração para este método.

O método proposto com normalização dos Mapas de Características (FE-LBP-FM+Norm-Conc) apresenta consistentemente melhores resultados do que a variação sem normalização dos Mapas de Características (FE-LBP-FM-Conc). Esta tem resultado inferior até que o FE-raw.

A Tabela III mostra o desempenho de cada método comparado, para os diferentes *smartphones* e condições de iluminação da base de dados VISOB. Os valores em **negrito** ressaltam cada melhor resultado obtido para cada *smartphone* e iluminação. Observa-se que dentre os descritores tradicionais (*handcrafted*), apenas o SIFT tem resultados similares à CNN, e em apenas uma ocasião apresenta melhor resultado.

V. Conclusões

Neste trabalho foram investigadas as características texturais visualizadas nos Mapas de Características de uma CNN. Foi utilizada a CNN AlexNet pré-treinada na base de imagens *ImageNet* como extrator de características, porém, ao invés de utilizar diretamente as Ativações como vetores de características, foi aplicado o descritor uLBP nos Mapas de Características das Ativações, extraindo assim características texturais. O método proposto foi avaliado para autenticação em *smartphones* por meio de imagens da região periocular, utilizando a base de dados VISOB.

O método foi proposto com duas variações, com e sem normalização dos Mapas de Características antes da aplicação do descritor. Foi constatado a importância da normalização dos Mapas de Características devido à grande variação numérica dos valores dos mapas.

O método proposto foi aplicado em todas as camadas da CNN *AlexNet*, para que se pudesse analisar o menor ERR em cada uma. Não foi possível definir uma camada como sendo a que fornece o melhor desempenho, considerando os diferentes cenários. No entanto, a variação proposta com normalização mostrou os melhores resultados só que em diferentes camadas.

Dentre métodos tradicionais utilizando os descritores uLBP, HOG, SURF e SIFT, o que obteve resultados mais próximos aos obtidos utilizando CNN foi o SIFT, porém, na maioria dos cenários os métodos *handcrafted* apresentaram resultados inferiores.

Como estabelecido em nossa hipótese, os resultados obtidos mostram que os Mapas de Características de uma CNN possuem informações texturais discriminantes, que podem ser extraídas usando-se um descritor de texturas. Infelizmente não foi possível determinar qual camada da CNN apresenta o melhor desempenho para ser usada no processo de autenticação, independente de tipo de *smartphone* e condição de iluminação. Em trabalhos futuros, pretende-se analisar o método aplicado em outras bases de dados, com outras arquiteturas de CNN e com outros descritores de textura.

References

[1] Gupta, S.; Buriro, A. and Crispo, B., 2018. Demystifying Authentication Concepts in Smartphones: Ways and Types to Secure Access. Mobile Information Systems, 2018, pp.1-16.

[2] Sharif, M.; Naz, F.; Yasmin, M.; Shahid, M. and Rehman, A., 2017. Face Recognition: A Survey. Journal of Engineering Science and Technology Review, 10(2), pp.166-177.

[3] Alonso-Fernandez, F. and Bigun, J., 2016. A survey on periocular biometrics research. Pattern Recognition Letters, 82, pp.92-105.

[4] Smereka, J. M.; Boddeti, V. N. and Kumar, B. V., "Probabilistic Deformation Models for Challenging Periocular Image Verification," in IEEE Transactions on Information Forensics and Security, vol. 10, no. 9, pp. 1875-1890, Sept. 2015, doi: 10.1109/TIFS.2015.2434271.

[5] Mahalingam, G.; Ricanek, K. and Albert, A M., "Investigating the Periocular-Based Face Recognition Across Gender Transformation," in IEEE Transactions on Information Forensics and Security, vol. 9, no. 12, pp. 2180-2192, Dec. 2014, doi: 10.1109/TIFS.2014.2361479.

[6] Juefei-Xu, F.; Luu, K.; Savvides, M.; Bui, T. and Suen, C.; 2011. Investigating age invariant face recognition based on periocular biometrics. 2011 International Joint Conference on Biometrics (IJCB),.

[7] Lecun, Y.; Bottou, L.; Bengio, Y. and Haffner, P., "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[8] Krizhevsky, A.; Sutskever, I. and Hinton, G., 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), pp.84-90.

[9] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), pp.211-252.

[10] He, K.; Zhang, X.; Ren, S. and Sun, J., "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[11] Girshick, R.; Donahue, J.; Darrell, T. and Malik, J., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[12] Hariharan, B.; Arbeláez, P.; Girshick, R. and Malik, J., 2014. Simultaneous Detection and Segmentation. Computer Vision – ECCV 2014, pp.297-312.

[13] Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A. and Torralba, A., "Places: A 10 Million Image Database for Scene Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452-1464, 1 June 2018, doi: 10.1109/TPAMI.2017.2723009.

[14] Razavian, A. S.; Azizpour, H.; Sullivan, J. and Carlsson, S., "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 512-519, doi: 10.1109/CVPRW.2014.131.

[15] Raja, K. B.; Raghavendra, R.; Stokkenes, M. and Busch, C., "Smartphone authentication system using periocular biometrics," 2014 International Conference of the Biometrics Special Interest Group (BIOSIG), 2014, pp. 1-8.

[16] Raja, K.; Raghavendra, R. and Busch, C., 2016. "Collaborative representation of deep sparse filtered features for robust verification of smartphone periocular images." 2016 IEEE International Conference on Image Processing (ICIP),.

[17] Raghavendra, R. and Busch, C., 2016. Learning deeply coupled autoencoders for smartphone based robust periocular verification. 2016 IEEE International Conference on Image Processing (ICIP),.

[18] Ahuja, K.; Bose, A.; Nagar, S.; Dey, K. and Barbhuiya, F., 2016. ISURE: User authentication in mobile devices using ocular biometrics in visible spectrum. 2016 IEEE International Conference on Image Processing (ICIP),.

[19] Rattani, A.; Derakhshani, R.; Saripalle, S. and Gottemukkula, V., 2016. ICIP 2016 competition on mobile ocular biometric recognition. 2016 IEEE International Conference on Image Processing (ICIP),.

[20] Raja, K.; Raghavendra, R.; Stokkenes, M. and Busch, C., 2015. Multi-modal authentication system for smartphones using face, iris and periocular. 2015 International Conference on Biometrics (ICB),.

[21] Santos, G.; Grancho, E.; Bernardo, M. and Fiadeiro, P., 2015. Fusing iris and periocular information for cross-sensor recognition. Pattern Recognition Letters, 57, pp.52-59.

[22] Ahmed, N.; Cvetkovic, S.; Siddiqi, E.; Nikiforov, A. and Nikiforov, I., 2016. Using fusion of iris code and periocular biometric for matching visible spectrum iris images captured by smart phone cameras. 2016 23rd International Conference on Pattern Recognition (ICPR),.

[23] Alonso-Fernandez, F.; Raja, K.; Busch, C. and Bigun, J., 2017. Log-likelihood score level fusion for improved cross-sensor smartphone periocular recognition. 2017 25th European Signal Processing Conference (EUSIPCO),.

[24] Stokkenes, M.; Ramachandra, R.; Raja, K.; Sigaard, M. and Busch, C., 2017. Feature level fused templates for multi-biometric system on smartphones. 2017 5th International Workshop on Biometrics and Forensics (IWBF),.

[25] Aginako, N.; Castrillón-Santana, M.; Lorenzo-Navarro, J.; Martínez-Otzeta, J. and Sierra, B., 2017. Periocular and iris local descriptors for identity verification in mobile applications. Pattern Recognition Letters, 91, pp.52-59.

[26] Rattani, A. and Derakhshani, R., 2017. On fine-tuning convolutional neural networks for smartphone based ocular recognition. 2017 IEEE International Joint Conference on Biometrics (IJCB),.

[27] Ahuja, K.; Islam, R.; Barbhuiya, F. and Dey, K., 2017. Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recognition Letters, 91, pp.17-26.

[28] Reddy, N.; Rattani, A. and Derakhshani, R., 2018. Comparison of Deep Learning Models for Biometric-based Mobile User Authentication. 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS),.

[29] Rattani, A.; Reddy, N. and Derakhshani, R., 2018. Multi-biometric Convolutional Neural Networks for Mobile User Authentication. 2018 IEEE International Symposium on Technologies for Homeland Security (HST),.

[30] Kondapi, L.; Rattani, A. and Derakhshani, R., 2019. Cross-illumination Evaluation of Hand Crafted and Deep Features for Fusion of Selfie Face and Ocular Biometrics. 2019 IEEE International Symposium on Technologies for Homeland Security (HST),.

[31] De Marsico, M.; Nappi, M. and Proença, H., 2017. Results from MICHE II – Mobile Iris CHallenge Evaluation II. Pattern Recognition Letters, 91, pp.3-10.

[32] Dertat, A., 2017. Applied Deep Learning - Part 4: Convolutional Neural Networks. [online] Towards Data Science. Available at: <https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2> [Accessed 27 September 2021].