# Identificação Facial Utilizando uma Rede CNN Pré-Treinada com Deep Learning

Jairo Lucas

Universidade Federal do Espirito Santo - Laboratório de Computação de Alto Desempenho - LCAD

***Resumo* - A tarefa de reconhecimento facial é uma das tarefas mais corriqueiras e naturais executadas pelos seres humanos. Apesar de simples para um humano esta tarefa tem se mostrado um grande desafio para pesquisadores nas áreas de inteligência artificial e visão computacional.**

**Nos últimos anos algoritmos que utilizam técnicas de Deep Learning, subárea do aprendizado de máquina. tem obtido relevantes resultado para a tarefa de reconhecimento facial, elevando o estado da arte para patamares em torno de 99.6% de acurácia [12] para a base de imagens LFW[11], base referência nesta área. Uma das principais características destas técnicas é a sua capacidade permitir a utilização de imensas quantidades de imagens para treinamento e obter excelentes resultados na classificação destes dados.**

**Neste trabalho utilizamos o framework Caffe (*Convolutional Architecture for Fast Feature Embedding*) [2] para avaliar uma Rede Neural Convolucional – CNN pré-treinada com técnicas de *Deep Learning* na tarefa de identificação facial, porém, aplicado a um DataSet pequeno, com poucas amostras de teste e treino.**

**Os resultados demonstram que mesmo com uma quantidade limitada de imagens de treino, a rede pré-treinada conseguiu uma taxa de acurácia de até 98.4%, taxa muito próximo ao estado da arte na tarefa de reconhecimento facial.**

***Termos - Redes Neurais Convolucionaiss, CNN, Deep Learning, Reconhecimento Facial, Caffe, Modelos pré-treinados.***

## 1.INTRODUÇÃO

Pesquisas na área de reconhecimento de objetos, e mais especificamente aquelas voltadas à face humana, têm aumentado muito nos últimos anos, principalmente devido à sua aplicabilidade em áreas tais como: segurança pública, controle de acesso, autenticação continua em redes de computadores, entre outras [14][15]. O reconhecimento facial tem obtido destaque sobre outras técnicas de biometria principalmente porque os dispositivos de captura (câmaras digitais) operam de forma não invasiva, são de baixo custo e fácil operação. Além disso, o aumento contínuo do desempenho dos processadores nas últimas décadas permitiu o uso de algoritmos mais sofisticados e robustos, como as redes neurais de aprendizado profundo (Deep Learning), que apesar da sua grande complexidade consegue oferecer uma resposta em um tempo aceitável e com uma acurácia superior a maioria das técnicas tradicionais de aprendizado.

### 1.1 – Reconhecimento Facial

O problema de reconhecimento de face pode ser caracterizado como: "dado uma imagem de entrada de uma face, comparar a face de entrada com uma biblioteca de modelos de faces conhecidas, e reportar se uma equivalência foi encontrada" [14], e se divide em dois modos distintos:

i — Verificação (*Face Verification*) : O Indivíduo fornece seus dados biométricos e um código de identificação tal como nome, CPF ou identificação funcional. O sistema examina se os dados biométricos de entrada são aqueles pertencentes ao indivíduo cuja identidade é reivindicada . Neste caso é feita uma comparação um-para-um, em que o sistema deve buscar uma resposta binária para a pergunta "Eu sou quem reivindico ser?" [15][16][17].

Computacionalmente, isto significa que não é necessário examinar toda a biblioteca de faces conhecidas a fim de verificar a veracidade da reivindicação. Alguns autores, como Yang [14], definem este processo como autenticação de face (*face authentication*).

ii — Identificação (*Face Identification*): O Indivíduo fornece somente seus dados biométricos ao sistema, que deve examinar toda a sua biblioteca de faces a fim de encontrar um individuo com características equivalentes. O sistema deve responder à pergunta "Quem sou eu?" [18].

O problema de identificação é mais complexo que o de verificação. Além de garantir a mesma acurácia exigida na verificação, a identificação deve percorrer toda a base de conhecimento, comparando a face de entrada com cada uma das faces desta, o que torna o tempo de resposta um problema a ser tratado.

Este trabalho trata do problema de identificação facial.

### 1.2 – Técnicas de *Deep Learning*

Nos últimos anos as técnicas baseadas em *Deep Learning* -

ganharam destaque entre os pesquisadores de várias áreas da inteligência artificial, principalmente aquelas voltadas para visão computacional. A grande maioria destas técnicas consiste em um conjunto de múltiplas tarefas de aprendizado de máquina que lida com diferentes tipos de abstrações. A técnica utilizada neste trabalho foi a de Redes Neurais Convolucionais – CNN.

### 1.2.1 – Redes CNN – Redes Neurais Convolucionais

Com a popularização dos celulares com câmaras digitais e a explosão de redes sociais como o Facebook e Instagram, o volume de informações em formato de imagens, fotos e vídeos disponíveis na internet vem aumentando de forma exponencial. O crescimento deste volume de informação demanda a criação de novas técnicas de buscas que não sejam baseadas em texto, e sim capazes de inferir informações diretamente destas mídias. As redes convolucionais foram projetadas para atuar neste tipo de problema, que envolve a detecção e reconhecimento de objetos, pessoas ou animais em uma determinada cena.

As CNN's – Redes Neurais Convolucionais - foram projetadas inspiradas na arquitetura biológica do celebro. Em 1968 Hubel e Wiesel realizaram experimentos com gatos e macacos e mostraram que o córtex visual é formado por um conjunto hierárquico de células sensíveis a pequenas sub-regiões chamadas de campos receptivos, de forma que cada célula é "especialista" em monitorar (e ser ativada) por uma pequena região. Hubel classificava estas células em categorias - simples, complexas e supercomplexas – de acordo com o padrão de estimulo que as ativam. Células simples são ativadas quando são apresentados padrões simples para o animal, como linhas. As células complexas e supercomplexas são ativadas quando padrões mais elaborados são apresentados ao animal.

A partir deste estudo surge a hipótese que uma boa representação interna para uma rede neural para reconhecimento de imagens seria uma estrutura hierárquica, onde os pixels formam arestas, as arestas formam padrões, os padrões combinados formam as partes, as partes combinadas formam os objetos e os objetos formam a cena [19].

Esta estrutura considera que o mecanismo de reconhecimento necessita de vários estágios de treinamento empilhados uns sobre os outros, um para cada nível de hierarquia [19]. As redes CNN's seguem este conceito, representando arquiteturas multi-estágios capazes de serem treinadas.

### 1.3 – Caffe - *Convolutional Architecture for Fast Feature Embedding*

O framework utilizado neste trabalho foi o Caffe - *Convolutional Architecture for Fast Feature Embedding* [2] , que foi projetado e desenvolvido pelos pesquisadores do *Berkeley Vision and Learning Center* (BVLC) da Universidade da California.

O Caffe é um framework de código aberto que oferece uma série de modelos e exemplos de redes pré-treinadas com deep learning. Possui uma comunidade bastante ativa [2] onde podem ser disponibilizados novos modelos e conhecimentos relevantes. Os modelos disponibilizados podem ainda ser adaptados ou parametrizados para uso em diversas aplicações.

Neste trabalho foi utilizado o modelo *VGG-Face CNN Descriptor,* descrito em [1].

## 2. Trabalhos Relacionados

Este trabalho tem como foco a identificação facial em imagens. Dentre os vários métodos propostos na literatura para o problema em foco, é possível distinguir os que são baseados em aprendizado profundo (Deep Learning) e os baseados em aprendizado superficial (Shallow) [1].

Entre os métodos que utilizam aprendizado "superficial" podemos citar os trabalhos de Zang [8] que propõem uma melhoria no algoritmo de codificação esparsa e alcança uma taxa de aproximadamente 98,5% de acurácia para a base AR Face[4], Yang [3] que utiliza algoritmo de transformação randômica e obtém uma acurácia de 95.8% para a base[4] , Tian [6] utiliza variações da redução de dimensionalidade com acurácia de 94.5% para a base [4] e Xu [7], que utiliza um algoritmo de regressão linear e obtém 95,52% para a mesma base. Existem ainda os métodos semiautomáticos de reconhecimento, onde a posição dos olhos, nariz e/ou boca são fornecidos junto com a imagem da face. De Souza [20] utilizou esta abordagem com um classificador baseado em Redes Neurais sem Peso e reportou uma taxa de acerto de até 99.3% para a base de imagens AR Face [4]. Também usando um método semiautomático, Park [21] utilizou um modelo onde informações geométricas estruturais da face são codificadas em um gráfico ARG (Attributed Relational Graph) e obteve resultados de até 98.5% para esta mesma base.

Os métodos baseados em aprendizagem profunda geralmente utilizam bases de dados maiores, não sendo encontrados trabalhos com estes métodos que reportem o uso da base AR Face. Entre os trabalhos relevantes que utilizam métodos de aprendizado profundo podemos citar Parkhi [10] que utiliza uma arquitetura baseada em rede neural profunda de 16 camadas e consegue uma acurácia de 98.8% na base de dados LFW [11], base esta composta de aproximadamente 13.000 imagens e cerca de 5.700 pessoas diferentes. Já Sun [13] propôs utilizar duas arquiteturas de redes neurais muito profundas e conseguiu uma acuraria de 99,53% neste mesma base. Schroff [12] utilizou uma variação de redes neurais convolucionais (CNN) e obteve uma acurácia de 99.63%, considerado o recorde para a base de dados LFW [11] [12].

## 3.Metodologia

### 3.1 – Base de Imagens

Este trabalho usou como base de imagens a base AR Face descrita em [11]. A versão utilizada conta com 3.314

imagens faciais com tamanho de 576 x 768. A base retrata 135 pessoas diferentes, sendo 76 homens e 59 mulheres. A base possui imagens com o rosto parcialmente oculto, variação na iluminação e diferentes expressões faciais.

3.2 – DataSet

Para uma melhor avaliação dos resultados, a partir da base principal, foram criados 4 conjuntos de imagens para treino da rede e 1 conjunto para teste. Os conjuntos são especificados abaixo.

3.2.1 – Conjuntos de Treino

*Treino-B2 :* Conjunto com 270 imagens, sendo duas imagens por pessoa, com variação na expressão facial e sem partes ocultas. A fig. 1 mostra as imagens deste conjunto.

Fig. Imagens do conjunto de treino *Treino-B2*

*Treino-B4 :* Conjunto com 540 imagens, sendo quatro imagens por pessoa, com variação na expressão facial, variação na iluminação e sem partes ocultas. A fig. 2 mostra as imagens deste conjunto.

Fig 2. Imagens do conjunto de treino *Treino-B4*

*Treino-B8 :* Conjunto com 1080 imagens, sendo oito imagens por pessoa, com variação na expressão facial, variação na iluminação e sem partes ocultas. A fig. 3 mostra as imagens deste conjunto.

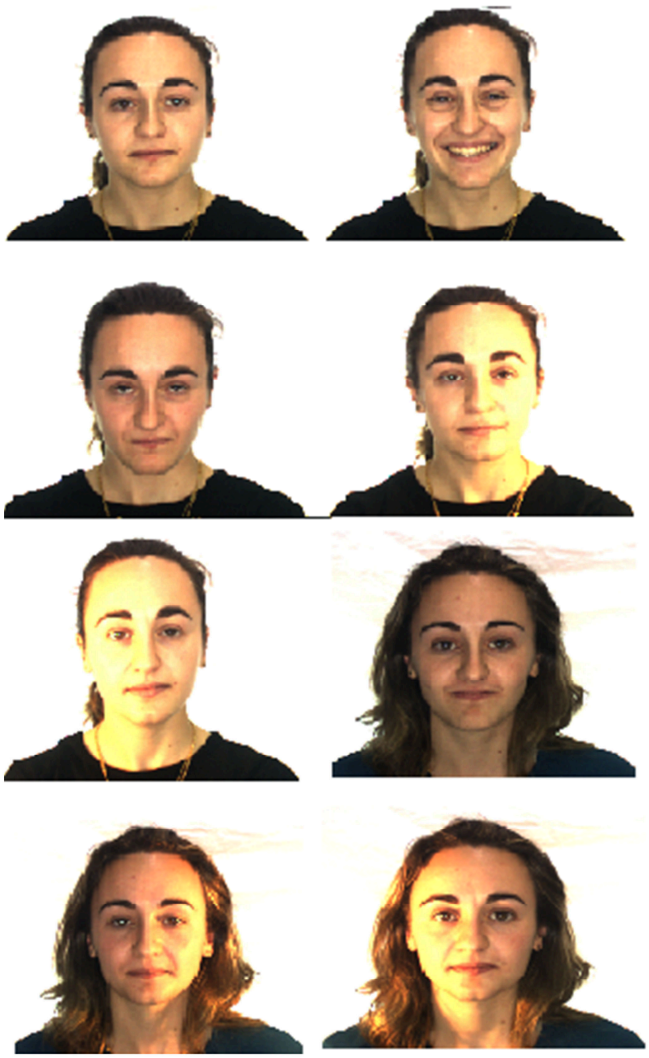

Fig 3. Imagens do conjunto de treino *Treino-B8*

*Treino-Base-Total :* Conjunto com todas a imagens da base, sendo aproximadamente 26 imagens por pessoa , com variação na expressão facial, variação na iluminação, com parte ocultas do rosto e com óculos de sol. A fig. 4 mostra exemplos de alguma das imagens que fazem parte deste conjunto.

Fig.4. Exemplo de Imagens do conjunto de treino *Treino-Base-Total*

3.2.2 – Conjuntos de Teste

*Teste-1:* O conjunto Teste-1 é formando por 120 imagens, 1 por pessoa, utilizando uma expressão facial neutra, sem óculos e sem ocultamento do rosto. Todas as pessoas deste conjunto possuem uma correspondência no conjunto de treino (dataset closed). A imagem (pose) utilizada no conjunto de

Teste-1 não faz parte do conjunto de treino. A figura 5 mostra exemplos das imagens deste conjunto.

Fig.5. Exemplo de imagens do conjunto ***Teste-1.***

3.2 – Hardware

O Hardware utilizado neste trabalho foi um Intel Xeon com 4 cores, 2.1 Ghz, 12 gb de memoria Ram e placa de vídeo Nvideo Tesla C2050.

3.3 – Software

Neste trabalho foi utilizado o framework Caffe descrito em [14] e o sistema operacional Linux.

## 4. Experimentos e Resultados

Após vários procedimentos de testes de calibração e configuração da rede, foram fixados os melhores parâmetros encontrados e aplicados a todos os conjuntos descritos anteriormente.

Em todos os experimentos é apresentado ao classificador uma face e o mesmo retorna uma lista com as faces mais semelhantes. Um classificador perfeito retornaria a identidade correta no top da lista para todas as faces apresentada.

Nos resultados são considerados dois valores, a acurácia denominada ***Top 1***, quando a identidade correta é a primeira da lista, e a acurácia denominada *Top 5*, quando a identidade correta é retornada entre a 2ª. e 5ª. posição da lista.

Os resultados e análise de cada teste são descritos na sequencia:

4.1 – Experimento I

Conjunto de treino : Treino-B2
Conjunto de teste : Teste-1
Numero de iterações utilizadas : 3.000
Tempo aproximado para o resultado: 1:15h

Neste conjunto a acurácia foi extremamente baixa, ficando em torno de **32%** para a Top 1 e **59%** para a Top 5, com variação de +/- 1.5 pontos. Este resultado era previamente esperado, visto que redes CNN tem um desempenho melhor com grande quantidade de imagens de treino, e neste conjunto a rede foi treinada com apenas 2 imagens por pessoa. Na figura 7 é mostrado um gráfico com o resultado deste conjunto.

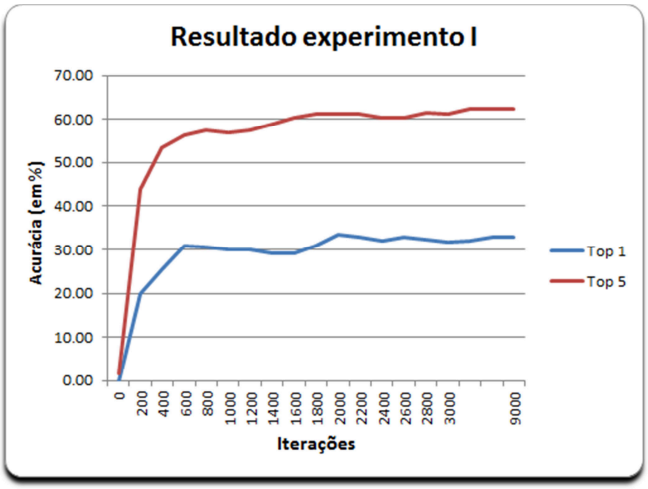


Fig.7. Resultados do experimento I

4.2 – Experimento II

Conjunto de treino : Treino-B4
Conjunto de teste : Teste-1
Numero de iterações utilizadas : 4.000
Tempo aproximado para o resultado: 1:44h

Neste conjunto de dados a acurácia melhorou significativamente, ficando em torno de **52%** para o top 1, com picos de 54% e **75%** para o top 5, com picos de 77%.

O aumento da acurácia era esperado em função do aumento do número de imagens de treino, que neste conjunto foi de 4 imagens para cada pessoa. Na figura 8 é mostrado o gráfico com os resultados do experimento II.

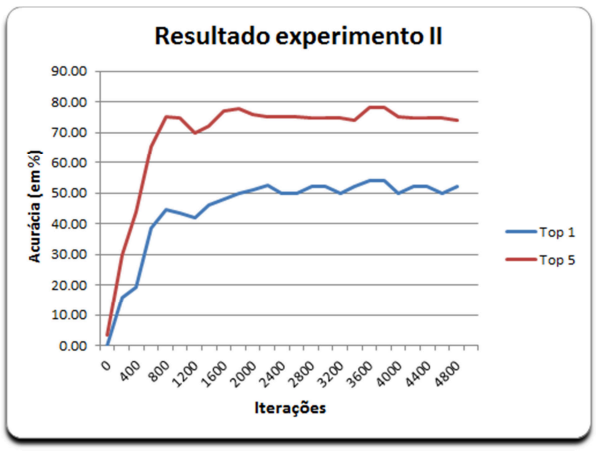


Fig.8. Resultado do experimento II

4.3 – Experimento III

Conjunto de treino : Treino-B8
Conjunto de teste : Teste-1

Numero de iterações utilizadas : 2.000
Tempo aproximado para o resultado: 1:20h

Neste conjunto de dados a acurácia também melhorou significativamente em relação ao conjunto anterior. Rapidamente e com pouco mais de 2.000 iterações a acurácia se estabilizou em torno de **94.6%** para o top 1 e **98.4%** para o top 5. Se prolongarmos os testes, a acurácia top 1 chega a 95,3% depois de 10.000 iterações e aproximadamente 3 horas de processamento. Já a acurácia top 5 não tem nenhum ganho com o aumento do número de iterações.

Na figura 9 é mostrado o gráfico com os resultados do experimento III.

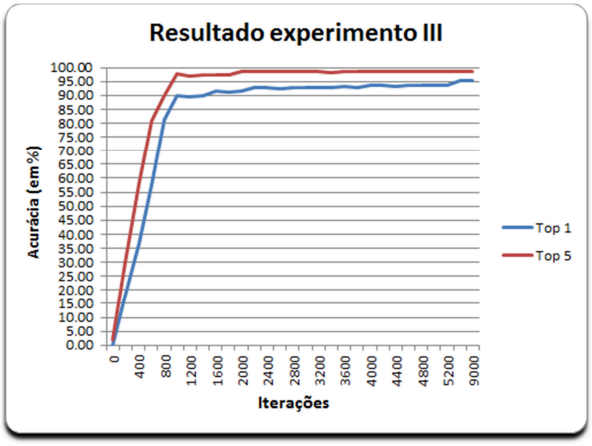


Fig.9. Resultado do experimento III

4.3 – Experimento IV

Conjunto de treino : Treino-Base-Total
Conjunto de teste : Teste-1
Numero de iterações utilizadas : 4.000
Tempo aproximado para o resultado: 2:30h

Neste conjunto foram utilizadas todas as imagens disponíveis para o treino da rede, sendo em média 24 imagens por pessoa.

Os resultados mostram que a rede é bastante estável e coerente com os resultados anteriores. Depois de aproximadamente 4.000 iterações, tanto a acurácia top 1 quanto a top 5 convergem para uma taxa de **98.4%** , mostrando que este é o limite de reconhecimento da rede montada.

A figura 10 mostra o gráfico com o resultado do experimento IV.

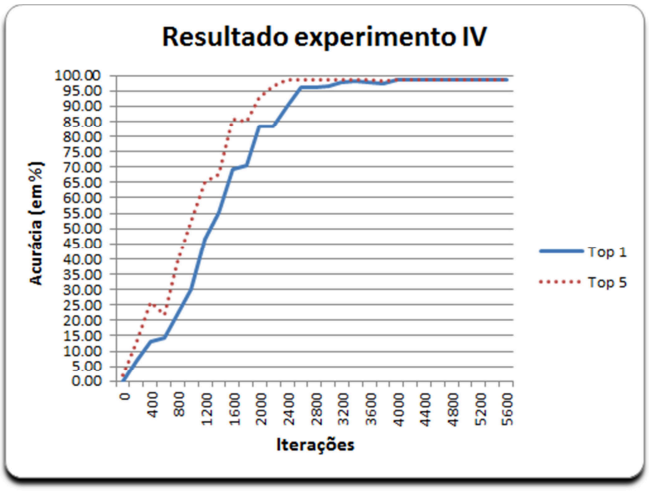


Fig.10. Resultado do experimento IV

## 5. Conclusões

Este trabalho testou a tarefa de identificação facial para a base AR Face utilizando redes CNN pré-treinadas com aprendizado profundo e o FrameWork Caffe.

Os trabalhos que relatam o uso desta base de imagens são todos baseados em aprendizagem "superficial". Entre eles podemos citar Zang [8] que relatada uma acurácia de aproximadamente 98%, Yang [3] com uma acurácia de 95.8% e Xu [7], que obtém uma acurácia de 95,52% para a mesma base. Existem ainda os métodos semiautomáticos de reconhecimento, onde a posição dos olhos, nariz e/ou boca são fornecidos junto com a imagem da face. De Souza [20] utilizou esta abordagem com um classificador baseado em Redes Neurais sem Peso e reportou uma acurácia de até 99.3% para a base de imagens AR Face [4]. Também usando um método semiautomático, Park [21] obteve resultados de até 98.5% para esta mesma base. Para maiores detalhes ver a seção de trabalhos relacionados.

Os métodos baseados em aprendizagem profunda, como o utilizado neste trabalho, geralmente utilizam bases de imagens maiores, não sendo encontrados trabalhos com estes métodos que reportam o uso da base AR Face. Entre os trabalhos relevantes podemos citar Parkhi [10] com uma acurácia de 98.8% na base de dados LFW [11], Sun [13] com uma acuraria de 99,53% nesta mesma base e Schroff [12] que obteve uma acurácia de 99.63%, e é considerado o recorde para a base de dados LFW [11] [12].

Neste experimento conseguimos um acurácia de **98.4%,** sendo um dos melhores, senão o melhor, resultado reportado para a tarefa de reconhecimento facial automático usando a base AR Face.

Os resultados mostram que a rede foi bastante coerente e estável nos vários experimentos realizados, e a acurácia obtida é comparável com o estado da arte na tarefa de reconhecimento facial. Com isso, é possível concluir que o uso das redes CNN pré-treinadas é viável na tarefa de reconhecimento facial, mesmo em bases de imagens pequenas.

## 6. BIBLIOGRAFIA


[1] O. M. Parkhi, A. Vedaldi, A. Zisserman; ***Deep Face Recognition ;*** British Machine Vision Conference, 2015

[2] Jia, Yangqing and Shelhamer, Evan and Donahue, Jeff and Karayev, Sergey and Long, Jonathan and Girshick, Ross and Guadarrama, Sergio and Darrell, Trevor; ***Caffe: Convolutional Architecture for Fast Feature Embedding***; arXiv preprint arXiv:1408.5093; 2014

[3] Yang, Meixia and Cao, Zhuming; ***Face recognition based on Radon transform;*** 4th National Conference on Electrical, Electronics and Computer Engineering (NCEECE 2015); China;2015

[4] A. M. Martinez. **The AR face database,** CVC Technical Report, 1998,

[5] Hu, Changhui and Lu Xiaobo; ***Single Sample Face Recognition via Lower-upper Decomposition;*** School of Automation, Southeast University, Nanjing, China ;2015

[6] Tian, Liang ***An Embedded Feature Extraction Algorithm with MMC for Face Recognition;*** International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.5 (2016), pp.309-320

[7] Xu, Li, Kan, Xie and Wang , Zhiyu; ***Nonnegative Linear Reconstruction Measure based Face Recognition System;*** 5th International Conference on Information Science and Technology (IC1ST);Changsha, China; 2015

[8] Zhang, Jun-Kai and Gu Xiao-Ya; ***Face Recognition based on Improved Robust Sparse Coding Algorithm;*** International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.8, No.9 (2015), pp.339-346

[9] Zhang, Jun-Kai and Gu Xiao-Ya; ***Face Recognition based on Improved Robust Sparse Coding Algorithm;*** International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.8, No.9 (2015), pp.339-346

[10] O. M. Parkhi, A. Vedaldi, A. Zisserman; ***Deep Face Recognition;*** British Machine Vision Conference, 2015

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: ***A database for studying face recognition in unconstrained environments***. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[12] Schroff , Florian; Kalenichenko, Dmitry and Philbin , James ; ***FaceNet: A Unified Embedding for Face Recognition and Clustering***. In Proc. CVPR, 2015.

[13] Sun ,Yi; Liang ,Ding; Wang, Xiaogang; Tang, Xiaoou ; ***DeepID3: Face Recognition with Very Deep Neural Networks***.; CoRR, abs/1502.00873, 2015

[14] YANG, M. H.; KRIEGMAN, D. J.; AHUJA, N. ***Transactions on Pattern Analysis and Machine Intelligence,*** *IEEE* vol. 24, n. 1, 2002.

[15] ZHAO, W.; CHELLAPPA, R.; PHILLIPS, J.; ROSENFELD, ***A. Face recognition: a literature survey***, ACM Computing Surveys, v. 35, n. 4, p. 399–458, 2003.

[16] HONG, L.; JAIN, A. ***Integrating faces and fingerprints for personal identification,*** IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 12, p. 30-36,1998.

[17] JAIN, A.; ROSS, A.; PRABHAKR, S. ***An introduction to biometric recognition***, IEEE Transactions on Circuits and Systems for Video Technology, Special issue on image and video - Based Biometrics, v. 14, n. 1, 2004.

[18] PHILLIPS, J.; SYED, R.; HYEONJOON, M. ***The FERET verification testing protocol for face recognition algorithms***, Technical Report, out. 1998.

[19] LECUN, Yann et al. ***Convolutional networks and applications in vision.*** In: ISCAS. 2010. p. 253-256.

[20] DE SOUZA, Alberto F. et al. ***VG-RAM weightless neural networks for face recognition***. INTECH Open Access Publisher, 2010.

[21] PARK, Bo-Gun; LEE, Kyoung-Mu; LEE, Sang-Uk. **Face recognition using face-ARG matching. IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 12, p. 1982-1988, 2005.