Reconhecimento de Expressões Faciais Utilizando uma Rede CNN Pré-Treinada com Deep Learning

Jairo Lucas

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

***Resumo* - Na ultima década a análise automática de expressões faciais por um sistema autônomo tornou-se uma área de pesquisa bastante ativa em função do seu grande potencial de aplicações, visto que as expressões faciais refletem não somente as emoções, mas também outras atividades mentais e sinais fisiológicos.**

A tarefa de reconhecimento de emoções por um sistema autônomo é uma tarefa extremamente complexa. Para se ter ideia deste nível de complexidade, podemos comparar esta tarefa com outra igualmente complexa, o reconhecimento facial. Enquanto alguns trabalhos de reconhecimento facial, como os de Schroff [7] e Sun Yi [8] já reportam uma acurácia acima de 99% para bases de imagens em ambiente não controlado, o ganhador do desafio de reconhecimento de emoções promovido na *16ª. International Conference on Multimodal Interaction* [9] em 2014 obteve uma acurácia em torno de 50% para a base SFEW (Static *Facial Expressions Wild) [12],* base composta com imagens em ambiente não controlado. Na conferência do ano seguinte [10], Zhiding [11] conseguiu uma acurácia em torno de 61%

Neste trabalho utilizamos o framework Caffe (*Convolutional Architecture for Fast Feature Embedding*) [2] para avaliar o desempenho de uma rede pré-treinada para a tarefa de reconhecimento de faces, aplicada à outra tarefa, o reconhecimento de emoções. Utilizamos como base de imagens a AR Face [3] que possui imagens em um ambiente controlado (sem fundo, poses frontais e uma única face por imagem). Tentamos identificar 4 expressões faciais: Neutro, Bravo, Feliz e Assustado.

A acurácia obtida foi de 86.4%, percentual razoável para bases com imagens em ambiente controlado.

Termos - Redes Neurais Convolucionaiss, CNN, Deep Learning, Reconhecimento de expressões, Reconhecimento de faces, Caffe, Modelos pré-treinados.

1.INTRODUÇÃO

No final da década de 70 um grupo de pesquisadores descobriu que para um conjunto de emoções básicas existem expressões não-verbais distintas, universais, e provavelmente inatas. Desde então várias áreas de pesquisas se dedicaram a testar esta tese da universalidade, sendo hoje aceita a ideia que uma expressão de raiva, por exemplo, teria as mesmas características faciais no Brasil, na Europa ou na Ásia.

Um sistema computacional que pudesse avaliar corretamente, e em tempo real, as expressões faciais e o estado emocional do seu usuário, elevaria a interação entre homem e máquina para outro nível, diminuindo ou eliminado qualquer barreira de linguagem e permitindo uma interação ativa do sistema computacional com o seu usuário.

Essa possibilidade fez com que o interesse pelo tema aumentasse muito na última década, principalmente nas áreas de inteligência artificial e visão computacional.

1.2 – Técnicas de *Deep Learning*

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, complexa e supercomplexa – de acordo com o padrão de estimulo que a 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. [4]

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. 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 é o Caffe - *Convolutional Architecture for Fast Feature Embedding,* 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 utilizamos o modelo *Vgg_face ,* modelo treinada para a tarefa de reconhecimento facial.

2. Trabalhos Relacionados

Este trabalho tem como foco a identificação de expressões faciais em imagens estáticas. Os resultados dos trabalhos nesta tarefa dependem muito da base de imagens utilizada. Trabalhos mais antigos usavam bases com imagens adquiridas em ambiente controlado, são imagens in-door, sem fundo complexo, em poses frontais e apenas uma face por foto. Para este tipo de base de imagens a acurácia relatada é próxima da perfeição como relatado no trabalho de Zhao[13] e no trabalho de Katsia e Pitas[14].

Trabalhos mais recentes utilizam bases de imagens mais próximas do "mundo real". São imagens out-door, adquiridas em situações reais, com fundo complexo e algumas vezes com mais de uma de uma face por imagem. Neste tipo de base de imagens a acurácia é bem menor, embora tenha melhorado significativamente nos últimos anos. Entre os trabalhos que utilizam este tipo de bases de imagens podemos citar Levi [1] que utilizou Redes CNN com mapeamento de padrões binários obtendo uma acurácia de 54% para a base SFEW, e Zhiding[11] que utiliza múltiplas redes Deep Learning e obtém um acurácia de 61% para esta mesma base.

3.Metodologia

3.1 – Base de Imagens

Este trabalho usou como base de imagens a base AR Face descrita em [3]. A versão utilizada neste trabalho conta com 3.314 imagens faciais com tamanho de 576 x 768. A base retrata 4 diferentes expressões faciais: Neutra, brava, assustada e feliz. A partir desta base foram criados dois conjuntos de imagens que são descritos na sequencia.

3.1.1 – Conjuntos de Treino

O conjunto de treino foi montado usando 1.333 imagens aleatórias do conjunto principal. As imagens possuem variação de iluminação, sem ocultamento de parte da face e sem óculos. Neste conjunto estão representadas 4 expressões faciais, conforme mostrado na figura 1.

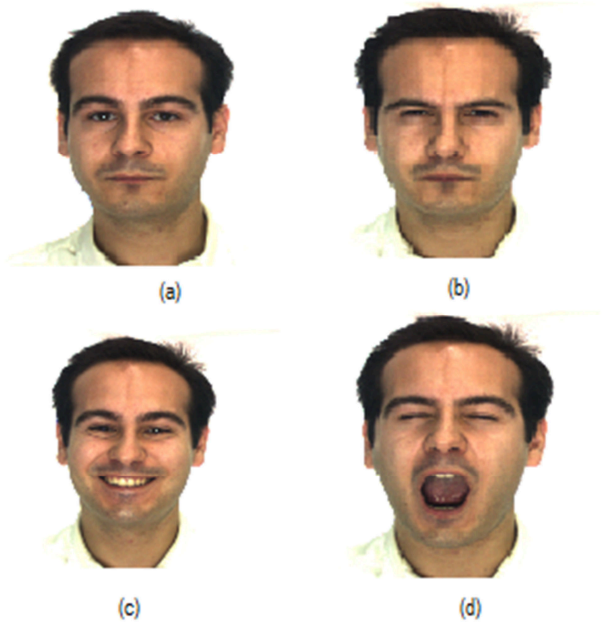

Fig. Expressões do conjunto de treino. (a) – Neutro, (b) – Bravo, (c)-Feliz, (d) assustado.

2.2 – Conjuntos de Teste

O conjunto Teste foi formando retirando aleatoriamente 447 imagens banco principal. Todas as expressões do conjunto de teste possuem uma correspondência no conjunto de treino (dataset closed). A imagem (pose) utilizada no conjunto de Teste não faz parte do conjunto de treino. A figura 2 mostra exemplos das imagens deste conjunto.

Figura 2 – Exemplo de imagens da base de Teste

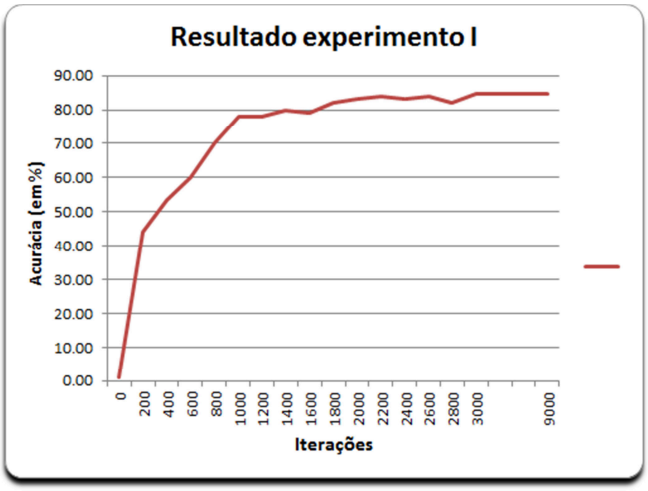

Fig.7. Resultados do experimento I

3.2 – Hardware

O Hardware utilizado neste trabalho foi um Intel Xeon com 4 core, 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 ao no experimento principal

No experimento é apresentado ao classificador uma face com uma expressão e o mesmo deve reconhecer a expressão baseado nas expressões apresentadas na fase de treino da rede.

O resultado e a análise do experimento são descrito na sequencia:

Conjunto de treino : Treino.txt
Conjunto de teste : Teste.txt
Numero de iterações utilizadas : 3.000
Tempo aproximado para o resultado: 1:35h

4 – Conclusões

Neste trabalho utilizamos uma rede CNN pré-treinada para a tarefa de reconhecimento de faces em uma tarefa de reconhecimento de expressões. O experimento utilizou a base de imagens AR Face. Esta base é formada por imagens adquiridas em ambiente controlado, in-door, onde todas as imagens não possuem fundo, são todas frontais e existe uma única face por imagem.

Os resultados relatados nos trabalhos nesta tarefa dependem muito do padrão das imagens utilizadas. Trabalhos mais antigos utilizavam bases com imagens adquiridas em ambiente controlado - mesmo padrão das imagens deste experimento - para este tipo de base de imagens a acurácia relatada é próxima da perfeição conforme descrito nos trabalhos de Zhao[13] e Katsia e Pitas[14].

Trabalhos mais recentes utilizam bases de imagens mais próximas do "mundo real". São imagens out-door, adquiridas em situações reais, com fundo complexo e algumas vezes com mais de uma de uma face por imagem. Neste tipo de base de imagens a acurácia é bem menor, Levi [1] relata uma acurácia de 54% para a base SFEW, e Zhiding[11] obtém um acurácia de 61% para esta mesma base. Para maiores detalhes ver a seção de trabalhos relacionados.

Em nosso experimento obtivemos um resultado de 86.4% de acurácia nos testes. Apesar do resultado estar abaixo dos relatados para o reconhecimento de expressões em bases de imagens em ambiente controlado, o mesmo comprova que é possível, dentro de certos limites, utilizar uma rede treinada para uma tarefa em outra tarefa mais especifica.

5 -Bibliografia

[1] Levi, Gil and Tal Hassner; ***Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns*** ; ACM

International Conference on Multimodal Interaction (ICMI), Seattle, . 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] A. M. Martinez. **The AR face database,** CVC Technical Report, 1998,

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

[5] Kobayashi, H. and Hara, F. ***The Recognition of Basic Facial Expressions by Neural Network***. Proc. IJCNN 1991, IEEE Computer Society (1991), 460-466.

[6] Happy ,S L ; Routray, A; ***Automatic Facial Expression Recognition Using Features of Salient Facial Patches***; IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, N. 1, JANUARY-MARCH 2015

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

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

[9] Dhall, Abhinav; Goecke,Roland; Joshi, Jyoti; ***Emotion Recognition In The Wild Challenge 2014:Baseline, Data and Protocol***; (EmotiW), 2014.

[10] Dhall, Abhinav; Goecke,Roland; Murthy, O. V. Ramana; ***Emotion Recognition In The Wild Challenge 2015:Baseline, Data and Protocol***; (EmotiW), 2015.

[11] Zhiding Yu, Cha Zhang ; ***Image based Static Facial Expression Recognition with Multiple Deep Network Learning***; Wild Challenge (EmotiW), 2015.

[12] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, ***Static Facial Expression in Tough Conditions - EFEW***: Data, Evaluation Protocol and Benchmark, IEEE ICCV BEFIT Workshop 2011

[13] Zhao, J., and G. Kearney. "***Classifying facial emotions by backpropagation neural networks with fuzzy inputs***" International Conference on Neural Information Processing. Vol. 1. 1996.

[14] Kotsia, Irene, and Pitas, Ioannis. "***Facial expression recognition in image sequences using geometric deformation features and support vector machines***." IEEE Transactions on Image Processing 16.1 (2007): 172-187.