APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS EM DADOS CAPTURADOS PELO KINECT: O ESTUDO DE CASO DO RECONHECIMENTO DE SINAIS DE LIBRAS

LUCAS TORTELLI[1]; SIMONE RUTZ[2]; MARILTON AGUIAR[3]

[1]*Universidade Federal de Pelotas – lmtortelli@inf.ufpel.edu,br*
[2]*Universidade Federal de Pelotas – sdrutz@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – marilton@inf.ufpel.edu.br*

1. INTRODUÇÃO

Durante o histórico da evolução humana, estes necessitavam comunicar-se através da fala, escrita e o mais instintivo de todos, o gesto. Atualmente a tecnologia baseia-se na comunicação escrita, através de dispositivos como teclado para realizar a troca de informações. Porém uma categoria quem tem ganhado destaque são as interfaces gestuais. Porque contém diversas aplicações como: telas sensíveis ao toque em *tablets*, área de jogos com o Microsoft Kinect™, movimento do controle do Nintendo Wii e, especificamente na área de tecnologias assistidas (ALVARENGA,2012)

Segundo censo realizado em 2010 pelo IBGE (Instituto Brasileiro de Geografia e Estatística), cerca de 9,7 milhões de brasileiros possuem deficiência auditiva, sendo que destes afetados 1 milhão são crianças e jovens de até 19 anos (IBGE,2012). Estes são introduzidos nas LIBRAS (linguagem de sinais brasileira) sem haver contato com o português, ou este sendo escasso. A dificuldade de entendimento da LIBRAS deve-se ao fato dessa comunicação ser expressa por sinais que muitas vezes são complexos.

A integração de técnicas de mapeamento, reconhecimento de padrões e aprendizagem de máquina pode ser utilizada para interpretação de gestos específicos. É neste contexto que o projeto LIKI – Libras e Kinect™ Aplicados às Tecnologias Assistivas, em desenvolvimento pelo grupo de pesquisa, está inserido. O projeto busca propor modelos e ferramentas computacionais para auxiliar no processo de ensino-aprendizagem da Língua Brasileira de Sinais (LIBRAS) para crianças com algum tipo de deficiência auditiva.

Mais especificamente, uma das temáticas abordadas pelo projeto trata do desenvolvimento e a aplicação de técnicas de reconhecimento de padrões para a identificação dos gestos. Outras temáticas abordadas no projeto envolve capturar, especificar e armazenas de modo apropriado os movimentos do usuário utilizando o sensor Kinect™ (versão 2) e, então, integra-las em um software que será utilizado em sala de aula.

Assim, este artigo pretende apresentar os resultados preliminares com a aplicação de redes neurais convolucionais (CNN) (WANG,2014) para o reconhecimento de gestos em LIBRAS capturados em imagens 2D. As imagens utilizadas contêm os gestos estáticos, obtidas com o sensor Microsoft Kinect™.

2. METODOLOGIA

LIBRAS consiste em uma língua que possui sua própria estrutura gramatical. Ao contrário da opinião pública a LIBRAS não contém relação gramatical com o Português, e sim essa está relacionada a linguagem gestual francesa, desta maneira assemelha-se a outras linguagens de sinais europeias (SOUSA,2012).

Assim como todas as demais línguas, LIBRAS contém diversos níveis linguísticos como: fonologia, morfologia, sintaxe e semântica, desta maneira a linguagem de sinais não compreendem somente sinais, e sim conexões que a transformam em uma forma de comunicação complexa. Os sinais como são chamados consistem em gestos utilizando-se das mãos, expressões faciais, orientação do corpo e entre outras.

Sendo a primeira língua aprendida por pessoas com deficiência auditiva, lamentavelmente, grande parte das pessoas fora da comunidade surda não a compreendem. Esta realidade dificulta a inclusão do surdo em vários ambientes (MONTEIRO,2013). Caracterizando desta maneira uma maior exclusão da interação deste grupo de pessoas, com o restante que não partilham da mesma deficiência de comunicação, causando assim uma maior segregação da sociedade.

O Microsoft Kinect™ é um dispositivo projetado para facilitar a iteração entre o usuário com a máquina. O dispositivo faz uso de gestos e comandos falados, revelando um novo modo de comunicação com o computador. O Kinect™ é capaz de reconhecer uma pessoa sob seu campo de visão, acarretando assim a descoberta de 25 junções no total, para determinação dos gestos (SOUSA,2012). O Kinect™ fornece diversas informações para realizar a análise de gestos como: Imagem do esqueleto com 25 junções, profundidade dos elementos no seu campo de visão, uma imagem em *full HD* (*High Definition*) e uma imagem em infravermelho.

A CNN trabalha com células simples e complexas, utilizadas para extração implícita de características dos padrões visuais apresentados como entrada, que posteriormente são integradas a uma rede neural *Multi Layer Perceptron*.

Uma entrada em uma CNN, transforma-se em uma série de camadas ocultas (DEEPLEARNING,2013). Em que cada camada oculta é composta por um conjunto de neurônios, em que cada neurônio é totalmente conectado aos neurônios da camada anterior, criando assim uma propagação de uma característica presente na imagem. Estas camadas podem sofrer alterações quando ocorrer o reconhecimento de uma nova característica. A arquitetura de uma CNN pode ser observada na Figura 1.

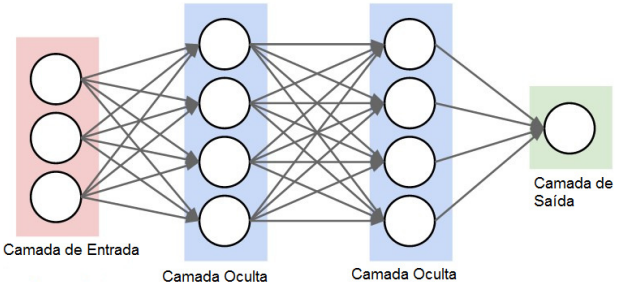

Figura 1. Arquitetura de uma CNN

Um mapa característico é formado para cada camada oculta, criando uma correlação entre os pontos de relevância da imagem. A aplicação de filtros lineares em sub-regiões desta mesma imagem, adicionando a polarização é chamado de convolução. Como pode ser visto todo esse processo acarreta em considerável esforço computacional, uma vez que cria diversas camadas de tamanho igual o da imagem, e aplica funções a cada um deles. Para isto surge a necessidade de inibir esse aumento de complexidade, porém sem perder informações. Desta maneira utiliza-se um conceito chamado de *max-pooling*, que é uma forma de amostragem, que visa particionar as sub-regiões de cada camada, ressaltando somente a

característica mais importante. Um exemplo desta técnica pode ser visto na Figura 2.

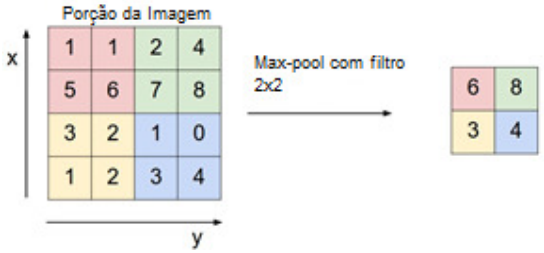

Figura 2: Exemplo de aplicação de *max-pooling*

Após todos os processos compreendidos na CNN, a imagem resultante consiste em uma sobreposição de todas as camadas intermediárias criadas na execução. Por fim é inserida em uma MLP e seu vetor de pesos é gerado para o melhor modelo encontrado.

3. RESULTADOS E DISCUSSÃO

Para a realização deste trabalho, foi utilizada a implementação de CNN disponibilidade na biblioteca de aprendizado de máquina em *Python* chamada Theano(BASTIEN,2012). Esta consiste em uma biblioteca própria para avaliar expressões matemáticas envolvendo vetores multidimensionais de forma eficiente, podendo ou não utilizar GPU (*Graphics Process Unity*) para realizar o processamento.

Foram utilizadas imagens de um banco de dados para posturas das mãos de forma estática (MARCEL,20120) com 6 gestos (a, b, c, ponto, cinco e v) produzidos por 380 pessoas, sendo cada pessoa realizando o mesmo sinal aproximadamente 10 vezes, totalizando 3800 imagens em RGB (*Red – Green-Blue*) de tamanhos variados. Primeiramente realizou-se a etapa de pré-processamento dos dados, em que todas as imagens foram padronizadas para um tamanho 66x76, visando não afetar a integridade das informações presentes na mesma. As imagens também sofreram alteração no seu sistema de cores, executando um filtro para deixa-las em escala de cinza. Este último diminui consideravelmente o esforço computacional.

Aplicando as imagens pré-processadas para realizar a execução, obtiveram-se os seguintes resultados de classificação de cada gesto, apresentados na Tabela 1.

Tabela 1: Classificação das imagens pelo algoritmo de CNN

Max-Pooling	**Tx. Aprendizado**	**Erro de Validação**	**Erro do Melhor Modelo**
2	0.0230	45,67%	42,82%
1	0.0114	14.358%	11.025%

Como pode ser observado na Tabela 1, utilizando um número menor de *Max-Pooling* conserva uma taxa de classificação com menor erro em comparação as demais, porém os custos computacionais para realizar a aprendizagem sem diminuir as dimensões da imagem são consideravelmente maiores. O número de épocas introduzido verifica o limite máximo de treinamento que a rede terá.

Tratando-se de reconhecimento de gestos da LIBRAS, os resultados alcançados foram satisfatórios levando em consideração a quantidade de imagens trabalhadas, atingindo erro de 14.358%.

4. CONCLUSÕES

A utilização de técnicas de aprendizagem para o reconhecimento de padrões em gestos de LIBRAS é de suma importância para a criação de uma ferramenta automatizada. Vale ressaltar que poucos trabalhos realizados conseguiram tratar este problema de forma íntegra. A CNN por tratarem-se de realizar o aprendizado a partir de imagens, podem mensurar resultados satisfatórios, uma vez que realiza suas técnicas extraindo características destas.

Observando o reconhecimento dos gestos selecionados obteve-se uma taxa satisfatória de erro de validação e do melhor modelo encontrado respectivamente, 14.358% e 11.02%. Desta maneira para trabalhos futuros estimasse realizar a execução do algoritmo para um maior conjunto de gestos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALVARENGA, M.L.T.,CORREA, D. S. O.,OSÓRIO, F. S., Redes Neurais Artificiais aplicadas no Reconhecimento de Gestos usando o Kinect, **Computer on the Beach**, Universidade do Vale do Itajaí, p.347-356,2012.

IBGE. **Deficiência auditiva atinge 9,7 milhões de brasileiros**. WinAudio, Curitiba, 30 mar. 2012. Acessado em 17 julho. 2015. Online. Disponível em: http://www.winaudio.com.br/produtos-e-servicos/noticias-em-audiologia/3704-deficiencia-auditiva-atinge-98-milhoes-de-brasileiros.html

MONTEIRO, O.M., ANTUNES, L. L., Estudo do uso do Kinect para interpretação de gestos visando LIBRAS, **13º Congresso Nacional de Iniciação Científica**, Faculdade Anhanguera de Belo Horizonte, v.1, 2013.

ANDERSSON, V. O.; GRAÑA, G.; ARAÚJO, R. M. Investigando o Uso do Kinect para Biometria Através do Caminhar Humano. **Artigo de Pós-Graduação em Computação**, Universidade Federal de Pelotas,2011.

WANG,J.,SONG, Y., LEUNG,T., ROSENBERG,C.,WANG,J.,PHILBIN,J., CHEN,B. ,WU,Y., Learning Fine-grained Image Similarity with Deep Ranking, **Cornell University Library,** 2014.

SOUSA, A.P.A. **Interpretação da língua gestural Portuguesa.** 2012. Dissertação de Mestrado em Engenharia Informática, Faculdade de Ciências, Universidade de Lisboa.

DEEPLEARNING. **Convolutional Neural Network (LeNet)**. DeepLearning, 2013, Acessado em 17 julho. 2015. Online. Disponível em: http://deeplearning.net/tutorial/lenet.html

BASTIEN, F.,LAMBLIN, P.,PASCANU, R., BERGSTRA, J., GOODFELLOW, I., BERGERON, A., BOUCHARD, N.,WARDE, D., BENGIO, Y., “Theano: new features and speed improvements”. **NIPS deep learning workshop**,2012.

MARCEL,S., Hand posture recognition in a body-face centered space., *Proceedings of the Conference on Human Factors in Computer Systems, Chile, 1999.*