JORNAL DE ENGENHARIA, TECNOLOGIA E MEIO AMBIENTE
Nº 1. Vol. 1 2016/2

REDES NEURAIS ARTIFICIAIS E PROCESSAMENTO DE IMAGEM NO RECONHECIMENTO DE LIBRAS, USANDO O KINECT

Luiz Carlos Gonçalves[1], Rogério Bernardes Andrade[2], Roberto Duarte de Campos[3], Bruno Alves Romero[4], Eduardo Fernandes Saad[5]

[1, 2, 3, 4, 5] Faculdade de Talentos Humanos - FACTHUS, Uberaba (MG), Brasil

luizcarlos_goncalves@yahoo.com.br, rbandrade@facthus.edu.br, roberto.campos@facthus.edu.br, bruno.romero@facthus.edu.br, eduardo.saad@facthus.edu.br

RESUMO: O artigo descrito tem por objetivo, desenvolver uma programação que traduz eletronicamente o gestual da Linguagem Brasileira de Sinais (LIBRAS) através do uso do sensor Kinect (usado no videogame XBox 360 da Microsoft), redes neurais e processamento de imagens. A aplicação tem pretensão de capturar gestos e testar, realizados livremente com a mão do usuário, permitindo o treinamento e reconhecendo assim diversos tipos de gestos diferentes, utilizando o sensor Kinect, que permite a captura de uma imagem juntamente com o mapa de profundidade. A partir do rastreamento e captura de imagens, é possível distinguir os gestos obtidos, que define letras em particular. O trabalho permite o acompanhamento de elementos de uma cena à captura, a aprendizagem utilizando matrizes binárias geradas por processamento de imagens e o reconhecimento em uma Rede Neural Artificial (RNA). Os resultados foram de 88%, bem acima do mínimo aceitável para uma rede neural, satisfazendo o objetivo deste artigo.

PALAVRAS CHAVE: Gestos; LIBRAS; Neural; Rede; Kinect.

ARTIFICIAL NEURAL NETWORK AND IMAGE PROCESSING FOR LIBRAS RECOGNITION, USING KINECT

ABSTRACT: The descriptive paper aims to develop a program that electronically translates the sign of Brazilian Sign Language (Libras) by using the Kinect sensor (used in the XBox 360 video game from Microsoft), neural networks and image processing. The application has intention to capture gestures and test freely performed with the user's hand, thus permitting the training and recognizing several different types of gestures using Kinect sensor, which allows the capture of an image with the depth map. From the tracking and capturing images, you can distinguish the gestures obtained, which defines particular letters. The work allows tracking elements of a scene capture, learning using binary matrices generated by image processing and recognition in an Artificial Neural Network (ANN). The results were 88%, well above the minimum acceptable for a neural network, meeting the objective of this article.

KEYWORDS: Gestures; LIBRAS; Neural; Network; Kinect.

INTRODUÇÃO

Em nosso planeta, sempre houve a necessidade da comunicação. No universo selvagem, animais utilizam de cores, gestos, sons e demais aparatos para a sua comunicação. Morcegos emitem sonares, pavões utilizam-se das suas cores e sua visão para acasalarem-se e comunicarem-se, e assim por diante, seguindo um grande leque complexo de exemplos para que possamos apenas pensar o que é, de modo prático, o ato da comunicação. No caso do ser humano, não é diferente, desenvolvemos nossos códigos, símbolos e sinais para que nossa expressão seja plena. A variedade é imensa: desenvolvemos a expressão iconográfica por meio de quadros, esculturas, e outras linguagens artísticas; o artifício da fala e da escrita nas mais diversas linguagens em todos os cantos da terra.

Nesse cenário, fez-se necessária a criação de uma linguagem, tão complexa e importante quanto qualquer outra: as linguagens envolvendo as necessidades especiais (como o Braile e as línguas para surdos, por exemplo). Dada a constante evolução no campo das tecnologias da informação é importante destacar a inclusão de questões envolvendo a acessibilidade, e a partir daí, a criação de meios para a comunicação e expressão, no caso deste projeto, de pessoas que apresentam surdez (SILVA; GESUELI; KAUCHAKJE, 2003). A teia tecida neste presente projeto não é nova, já que a vertente cresce exponencialmente no Brasil e no mundo no que diz respeito a pesquisas acadêmicas, ressaltando as suas contribuições sociais. A ausência de fala, por falta de contato com a língua portuguesa, por surdez ou por deficiência nas cordas vocais, possibilitou uma dada parcela da sociedade mundial que criasse a sua própria língua de sinais baseada na falada, ou seja, duas linguagens oficiais dentro de um próprio país. No Brasil, temos a LIBRAS (Língua Brasileira de Sinais) que envolve alfabeto, números, gestos e expressões corporais a fim de promover a comunicação entre indivíduos (HAGEN, Rose-Marie; HAGEN, Rainer. Goya, 2005). Tal como em nosso país, podemos pensar também as possíveis aplicações em interfaces semelhantes apresentadas aqui, em outras palavras, há a possibilidade de tais aplicações em reconhecimento de outras linguagens de sinais, como no

caso da ASL (American Sign Language) (Starner, 1998; Rung-Huei & Ming, 1998; Peres et al., 2006).

Ressalta-se, que o objetivo deste artigo é criar um protótipo de software que, através da captura de imagens gestuais via Kinect e treino da Rede Neural Artificial (RNA), entrelaçados em uma gama de conceitos e métodos que aqui serão discorridos, produzirão resultados que nos permitirão uma análise conclusiva das relações envolvendo homem e computador. Tendo como base trabalhos anteriores na área, os quais apresentaram positivos resultados. Esse trabalho inspira-se no intuito de contribuir com o campo de pesquisa e, consequentemente, com o meio social que vivemos.

MÉTODOS

O aplicativo foi desenvolvido utilizando a linguagem de programação Java e a linguagem C# nas plataformas NetBeans IDE 8.0.2 e Microsoft Visual Studio Express 2013. Os testes foram executados após o desenvolvimento da aplicação em um notebook com processador Core i7 e 8 GB de memória. O software foi desenvolvido em plataformas distintas. Foi buscada a bibliografia referente à essas linguagens, bem como pesquisa sobre o funcionamento da tecnologia do sensor de reconhecimento visual de gestos existentes atualmente no mercado.

DESENVOLVIMENTO

O campo de pesquisa relacionando homem-computador e a utilização da captura de imagens para tal, abriu várias portas na subárea, envolvendo jogos e experiências das mais diversas realidades virtuais aumentadas, como exemplo (Alvarenga; Corrêa; Osório, 2012). Para a execução deste trabalho, foram capturadas, via Kinect, as imagens gestuais estáticas, gestos do alfabeto de LIBRAS que exigem movimentos como os que representam as letras “Z” e “J”, não sendo utilizados para o aprendizado da RNA. Porém, é preciso que seja apresentado um dos objetos essenciais para a condução desse projeto. Tendo em mente que há diversas formas de se conseguir material iconográfico a ser utilizado no treino da Rede Neural Artificial, neste trabalho, porém foi utilizado a partir da captura de imagens através do Kinect. Trata-se de um sensor desenvolvido pela Microsoft e utilizado no vídeo game Xbox 360 que propicia a captura de elementos em terceira dimensão (3D), capturando movimentos do corpo humano com maior precisão em relação a dispositivos como o Wiimote, entre outros.

MODELAGEM DO PROTÓTIPO

Para a aprendizagem e reconhecimento dos gestos, foram criados 3 módulos de software que integram técnicas de rastreamento de movimento, processamento de imagens e inteligência artificial. A operação do software é dividida em etapas distintas que são executadas para atingir um objetivo em comum.

KINECT

Diante das inovações tecnológicas e demandas que surgem hora após hora, fez-se necessário por parte da indústria do entretenimento a criação de um dispositivo para que houvesse uma experiência plena entre o usuário e uma outra dimensão, no caso, nos referimos ao homem e aos videogames, ou ainda, a relação entre o real e o virtual.

Com o propósito de sugerir essa experiência, foi criado o Kinect (Figura 1), dispositivo para a captura de movimentos, utilizado no videogame Xbox 360, fabricado pela Microsoft em 2010, batendo recordes de venda e apresentando uma série de possibilidades como ferramenta utilizada em diversos campos da tecnologia.

Esse dispositivo conecta-se por meio de uma interface USB 2.0 com o Xbox, Linux ou Microsoft Windows; possui ainda um motor de angulação de câmera – o sistema RGB-D – mapeamento de ambientes em 3D, captação de som ambiente, e sensibilidade à profundidade infravermelha.

Para a comunicação do Kinect para Xbox 360 com o computador, foi utilizada uma fonte adaptadora para o uso em um laptop, e a utilização do SDK 1.8 (kit para desenvolvimento de software) contendo APIs, amostras de códigos e drivers necessários para que os desenvolvedores criem suas próprias aplicações ao fazerem uso do sensor de movimento.

Figura 1 - Sensor Kinect.

RASTREAMENTO DA MÃO

Quando um ser humano está presente na frente do Kinect, o quadro de esqueleto é capturado. Conforme (MSDN, 2015), a distância mínima entre o Microsoft Kinect e a pessoa a ser rastreado é de 6 pés (aproximadamente 1,83m). A pessoa que está perto do Kinect é tomada como a pessoa a ser rastreada de acordo com “Z” que é o eixo de profundidade (é considerado o tridimensional com 3 eixos x, y e z) tomado como esqueleto primário. Até vinte juntas do corpo são rastreadas para dentro do corpo da pessoa. A partir do esqueleto primário, os pontos do pulso (WristJoint) e os pontos da mão (HandJoint) são obtidos e o rastreamento da mão próxima ao usuário é efetuado.

Após o rastreamento da mão do usuário, subtrai-se a mão do fundo (backgroud), e o menor valor de

profundidade é obtido e registrado como a profundidade mínima.

A partir dos dois pontos (ponto de pulso e da mão) do esqueleto, são convertidos em um ponto de profundidade. Os pixels fora do alcance da profundidade (valor mínimo e máximo da profundidade) limiar não são registrados e são marcados com valores de pixels preto (MSDN,2015).

Figura 2 - Rastreamento da mão.

PROCESSAMENTO DE IMAGEM

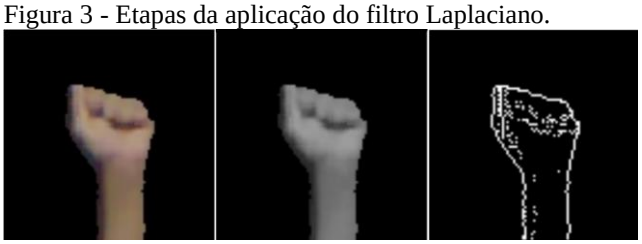
O formato de imagem Jpeg ou Jpg (Joint Photographics Experts Group, um método de compressão de imagens) foi a extensão utilizada no processamento das imagens, pois ele apresenta as mesmas características do modelo bitmap, porém com tamanho reduzido. "Os bitmaps nada mais são do que o conjunto de pixels da imagem digital, onde o valor de cada pixel representa as suas características de luminosidade e cor. Desse modo, os dados mapeiam a imagem, dando origem ao nome bitmap" (MARQUES et al., 1999, p. 243). O padrão adotado especifica a quantidade de bits utilizada dentro de cada pixel. Nas imagens no padrão RGB (Red, Green e Blue, em inglês) utiliza-se 8 bits para representar cada cor. Desta forma, para cada pixel deste padrão de imagem, são necessários 24 bits (SOLOMON et al., 2013, p. 9).

No trabalho o processamento é usado para realçar as bordas da imagem em questão, o conjunto de imagens utilizadas tem como padrão 95x95 pixels e divididos em quadrantes para manipulação dos mesmos.

PRÉ-PROCESSAMENTO

Para efetuar essa etapa, esta foi dividido em duas partes, sendo elas escala de cinza e o filtro laplaciano. A escala de cinza foi utilizada primeiramente para prosseguir com o processo, nesta etapa as camadas de cores (RGB) são transformadas em apenas uma soma ponderada das cores. De acordo com Solomon *et al.* (2013, p. 12) "a imagem em escala de cinza é, basicamente, uma soma ponderada dos canais de cor vermelho, verde e azul". A escala de cinza possui pixels que variam entre 0 e 255 totalizando 256 níveis, sendo os valores próximos de zero a tonalidade mais próxima de preto e os valores mais próximos de 255 são aqueles que mais se aproximam da cor branca. A próxima parte é a aplicação do filtro laplaciano, que por sua vez destaca as bordas contidas na imagem, onde a imagem utilizada já é capturada pelo Kinect com o background removido e transformado em pixels pretos, contendo apenas a mão em destaque, conforme a Figura 3.

Figura 3 - Etapas da aplicação do filtro Laplaciano.

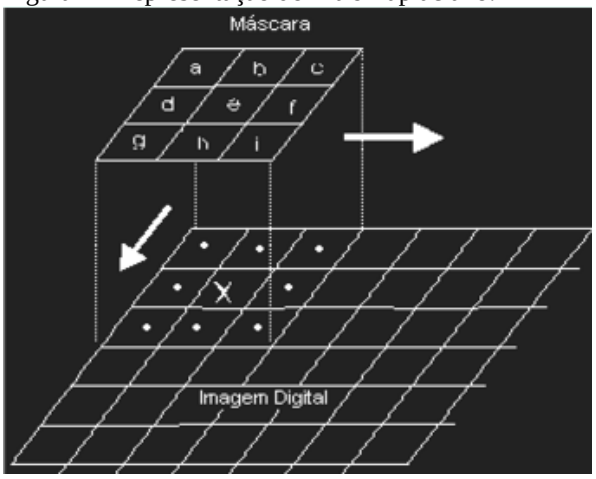


FILTRO LAPLACIANO

Nessa etapa, são destacadas as bordas da imagem, pois são os limites e os contornos que contrastam o fundo do objeto contido na imagem. O método laplaciano entra nesse processo como detector de bordas. No processo realiza-se uma aplicação de uma matriz de convolução denominado filtro laplaciano, que gera uma nova imagem com as bordas da imagem ou objetos da imagem original com fundo preto. Utiliza-se essa matriz para calcular a intensidade do pixel utilizando uma operação lógica e aritmética orientada a vizinhança (MARQUES *et al*., 1999, p. 33). Na parte de convolução do filtro, os valores de todos os pixels da imagem original e seus vizinhos (3 x 3) são multiplicados pela máscara "{{0, -1, 0 }, { -1, 4, -1 }, { 0, -1, 0 }}".

Os valores são somados e cada pixel e sua vizinhança se transformam em um novo pixel que é transportado para uma nova imagem de mesmo tamanho em relação à imagem original.

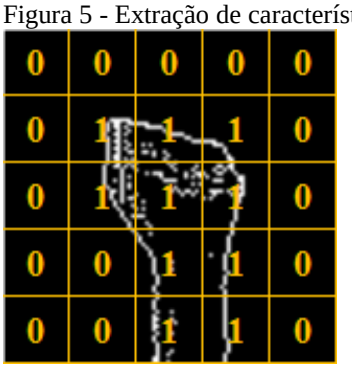
Figura 4 - Representação do filtro Laplaciano.



EXTRAÇÃO DE CARACTERÍSTICAS

No processo de extração de característica, a imagem resultante do filtro com tamanho total de 95x95 pixels é dividida em uma matriz de 5x5, onde cada quadrante possui tamanho de 19 pixels, totalizando 25 quadrantes, e a detecção de pontos brancos e pretos nestes quadrantes geram uma matriz de características, onde é identificado se o quadrante possui a cor branca; am caso positivo, atribui-se o valor 1, caso contrário atribui-se 0, essa matriz de características é usada posteriormente. A Figura 5 mostra de forma representativa como a matriz é gerada.

Figura 5 - Extração de características.



RECONHECIMENTO DOS GESTOS

Foi utilizada uma rede neural artificial (RNA) denominada *perceptron* de múltiplas camadas (PMC) para o reconhecimento dos gestos. A PMC constitui-se de pelo menos uma camada de entrada, uma camada de saída e pelo menos uma camada escondida, podendo ter mais de uma camada escondida, se necessário. Uma rede PMC aprende e reconhecer padrões a partir de treinamentos. Segundo Silva et al. (2010, p. 91)

> [...]as redes PMC são ainda caracterizadas pelas elevadas possibilidades de aplicações em diversos tipos de problemas relacionados com as mais diferentes áreas do conhecimento [...]".

As redes neurais artificiais são representações simplificadas dos neurônios biológicos humanos, nos quais os neurônios são conectados entre si. Os neurônios recebem valores iniciais diferentes em suas entradas e cada valor recebe uma ponderação, os resultados são acumulados e transferidos para a camada seguinte utilizando uma função de ativação. Para o desenvolvimento do trabalho, foi utilizado a função de ativação sigmoide por apresentar uma melhor utilização em relação as demais funções existentes, pois sempre resultam em zero e um, onde os dados existentes no trabalho, tanto as entradas da rede quanto as saídas são números binários. A expressão matemática da função sigmoide é dada pela equação 2.

$$g(u) = \frac{1}{1 + e^{-u}} \qquad (1)$$

TREINAMENTO DA REDE NEURAL

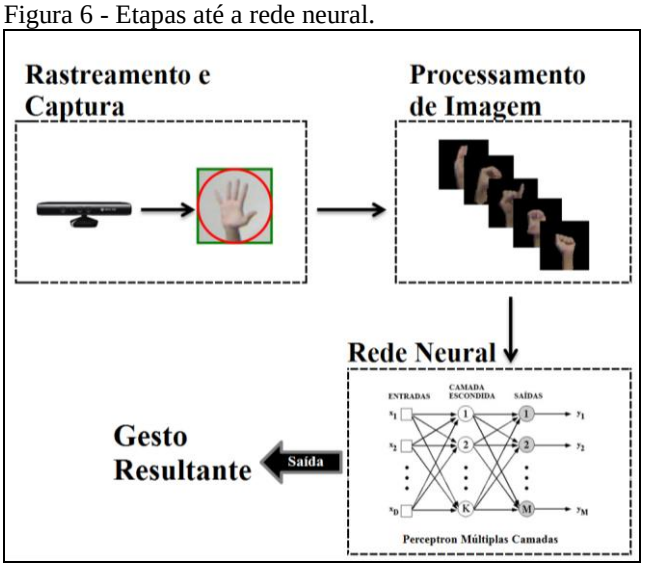
A perceptron de múltiplas camadas é treinada com o algoritmo back-propagation, onde as amostras são apresentadas a rede com suas respectivas saídas, o qual são calculadas e comparadas com as saídas desejadas. Os erros encontrados são retro propagados para os neurônios das camadas anteriores até a camada de entrada. É efetuada uma soma ponderada dos gradientes dos neurônios da próxima camada e é gerado o gradiente local (HAYKIN, 2001, p. 194). Repete-se o processo todo até que o erro médio quadrático seja menor que a precisão estipulada. O gradiente local do *j*-ésimo neurônio de saída é calculado conforme a equação 2.

$$\delta_j(n) = (d_j(n) - o_j(n))o_j(n)(1 - o_j(n)) \qquad (2)$$

As épocas de treino são quando todas as amostras passam pelo treinamento uma vez, as épocas podem variar de acordo com os parâmetros usados, se os parâmetros forem diferentes resultarão em uma nova quantidade de épocas.

Foi abordada neste artigo a utilização do treinamento supervisionado, tendo a disposição as amostras de treino e suas respectivas saídas. Foram utilizados os valores gerados pelo filtro contidos em um arquivo de extensão TXT, onde as entradas e as saídas são colocadas manualmente no software de treino e executado o processo de treinamento supervisionado. O processo de teste é realizado posteriormente, com os mesmos procedimentos utilizados no treino na manipulação do arquivo de amostras. OI processo geral esta representado na Figura 6.

Figura 6 - Etapas até a rede neural.



RESULTADOS

Para execução dos testes feitos, foram utilizadas 420 imagens diferentes da mão.

As letras em Libras "R" e "U" e as letras "T" e "F" possuem semelhanças. Após o processo de extração de

características, geraram valores binários muito parecidos, dificultando a difenrenciação das letras.

Para os testes realizados, foram tiradas das amostras as letras “U” e “T”, e também letras que necessitam de movimento como por exemplo as letras “H”, “J” e “Z” por não representarem a abordagem real que seria o reconhecimento de gestos estáticos.

Como método de padronização, foram gerados pesos iniciais aleatórios e utilizados em todos os treinos nas variações dos parâmetros; caso não processe dessa forma, os resultados sofreriam alterações por não serem iguais, gerando diferentes resultados para os mesmos parâmetros.

Devido ao tempo gasto no treino ser elevado, foi utilizado uma variação no algoritmo de treinamento para diminuir o tempo na convergência da rede. Na utilização de taxa de aprendizagem dinâmica ela sofre variações de acordo com o próprio histórico de treinamento se ajustando automaticamente. A taxa de aprendizagem é ajustada para 0,1 e, durante o processo após 10 épocas, se o erro global da rede diminuir, acrescenta-se 10% à taxa de aprendizagem, mas se aumentar, o erro global é decrescido em 10%.

Na primeira variação no treinamento da rede alfabética de libras, foi utilizado uma camada escondida com 10 neurônios e 3 de precisão (número de casas decimais), variando as quantidades de amostras de treinamento afim de descobrir o melhor parâmetro para atingir o melhor acerto possível.

Conforme a tabela 1 foi identificada que com 420 amostras, a rede acertou 85% das 42 amostras de teste que foram utilizadas em todo o processo.

Tabela 1 - Treinamento da rede alfabética de libras.

Camada Escondida	Amostras	Precisão	Nível de Acerto
10	105	3	83,00%
10	210	3	83,00%
10	315	3	80,00%
10	420	3	85,00%

Fonte - Dados obtidos pelo autor (2015).

As quantidades de amostras utilizadas foram relevantes na obtenção do melhor resultado. A Figura 7 exibe o gráfico de acertividade em relação à quantidade de amostras.

Na segunda variação no treinamento da rede alfabética de libras, foi utilizada uma camada escondida com 10 neurônios e 420 amostras por apresentar uma maior assertividade, e foram variadas as precisões da rede, acrescentando de 3 em 3, no qual a precisão de 12 mostrou ser superior, acertando 88% conforme Tabela 2.

A modificação da precisão foi relevante na obtenção de uma maior assertividade. A Figura 8 exibe o gráfico de assertividade em relação à variação da precisão.

Figura 7 - Comparativo de assertividade em relação à quantidade de amostras.

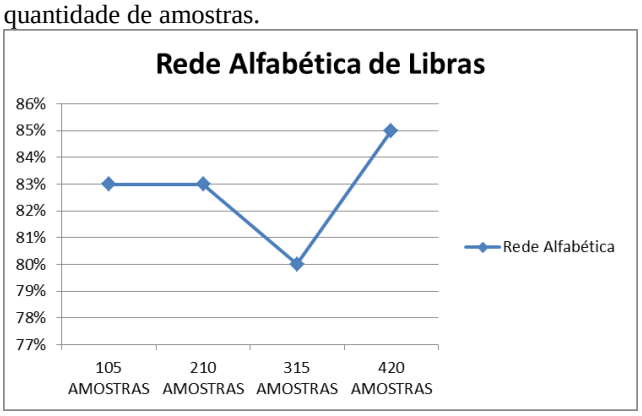
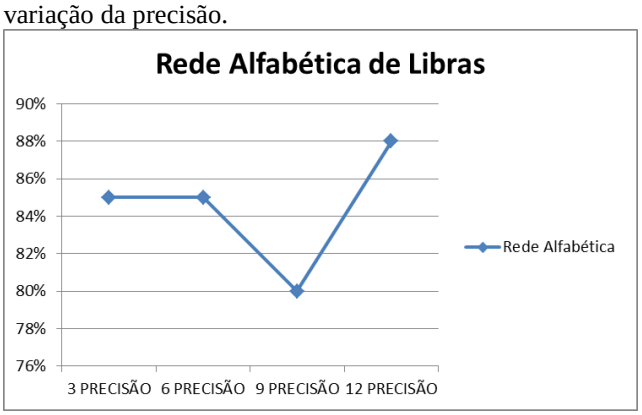


Tabela 2 - Treinamento da rede alfabética de libras.

Camada Escondida	Amostras	Precisão	Nível de Acerto
10	420	3	85,00%
10	420	6	85,00%
10	420	9	80,00%
10	420	12	88,00%

Figura 8 - Comparativo de assertividade em relação à variação da precisão.



Fonte - Dados obtidos pelo autor (2015).

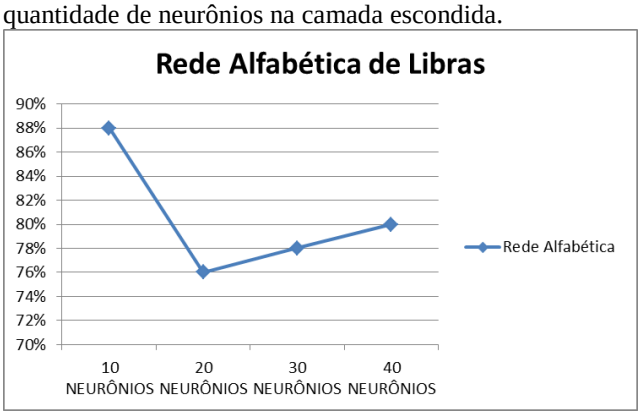
Na terceira variação no treinamento da rede alfabética de libras, foram utilizadas 12 casas decimais de precisão e 420 amostras. Esses parâmetros foram usados por apresentarem maior assertividade, e foram variadas as quantidades de neurônios na camada escondida para analisar qual seria o melhor parâmetro possível. Os neurônios foram acrescidos de 10 em 10, no qual observou-se que, à medida que os neurônios aumentavam, o nível de acerto caía conforme mostra a Tabela 3. Foi identificado que com 10 neurônios, a rede obteve uma assertividade superior às demais modificações feitas nas quantidades de neurônios no treinamento acertando 88% das 42 amostras de teste.

Tabela 3 - Treinamento da rede alfabética de libras.

Camada Escondida	Amostras	Precisão	Nível de Acerto
10	420	12	88,00%
20	420	12	76,00%
30	420	12	78,00%
40	420	12	80,00%

As modificações nas quantidades de neurônios não foram relevantes na obtenção de uma maior assertividade. A Figura 9 contém o gráfico em relação à quantidade de neurônios na camada escondida.

Figura 9 - Comparativo de assertividade em relação à quantidade de neurônios na camada escondida.



CONCLUSÃO

Este projeto mostrou ser capaz de efetuar o aprendizado e o reconhecimento de gestos estáticos através do uso de um Kinect, Processamento de Imagens e de Redes Neurais Artificiais. Os resultados foram bem satisfatórios, com 88% de acertos, porém existem possibilidades de efetuar melhorias no sistema para obtenção de resultados ainda melhores no reconhecimento dos gestos.

Como trabalhos futuros, espera-se aumentar a quantidade de gestos para uma melhor experiência, tão bem como reconhecer os gestos de forma automática em tempo real.

REFERÊNCIAS

ALVARENGA, M. L. T.; CORREA, D.S.O.; OSORIO, F. S. **Redes Neurais Artificiais aplicadas no Reconhecimento de Gestos usando o Kinect.** In: Computer On The Beach 2012, Florianópolis. Anais do Evento Computer On The Beach, 2012. p. 347-356.

HAYKIN, S.; **Redes Neurais**: Princípios e Práticas. Porto Alegre: Bookman, [2001]. 900 p.

LIBRAS em Contexto: Curso Básico: Livro do Professor. 4. ed. Rio de Janeiro: LIBRAS Editora Gráfica, 2005. HAGEN, Rose-Marie & HAGEN, Rainer. Goya.

MARQUES FILHO, O.; VIEIRA NET, H. **Processamento Digital de Imagens**. Rio de Janeiro: Brasport, [1999].

MSDN, **Microsoft Kinect.** Disponível em: <<http://msdn.microsoft.com/enus/library/hh855419.aspx>>. Acesso em 27 out 2015.

MSDN, **Microsoft Kinect**. Disponível em: <<http://msdn.microsoft.com/en-us/library/hh855352.aspx>>. Acesso em 27 out 2015.

PERES, Sarajane Marques; FLORES, Franklin Cesar; VERONEZ, Denise; OLGUIN, Carlos Jose Maria. **LIBRAS Signals Recognition: a study with Learning Vector Quantization and Bit Signature**. Ninth Brazilian Symposium on Neural Networks, 2006. SBRN '06. pp. 119 – 124. DOI: 10.1109/SBRN.2006.26

SILVA, I .R, KAUCHAKJE, S., GESUELI, Z. M. **Cidadania**, **Surdez e Linguagem** - Desafios **e** Realidades. São Paulo: Plexus Editora, 2003.

STARNER, T.; WEAVER, J.; PENTLAND, **A. Real-time American sign language recognition using desk and wearable computer based video.** In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Dec. 1998 Vol.20,N.12. pp. 1371 – 1375. DOI: 10.1109/34.735811

SOLOMON, C.; BRECKON, T. **Fundamentos de Processamento Digital de Imagens**: Uma Abordagem Prática com Exemplos em Matlab. Rio de Janeiro: LTC, [2013]. 289 p. http://www.dpi.inpe.br/~carlos/Academicos/Cursos/Pdi/pdi_filtros.htm