ESTUDO DAS ARQUITETURAS FP E LSTM DE REDES CNN APLICADAS EM DADOS CAPTURADOS PELO KINECT™: O ESTUDO DE CASO DO RECONHECIMENTO DE SINAIS DE LIBRAS

MONIQUE FARIA; LUCAS TORTELLI, SIMONE RUTZ; MARILTON AGUIAR

Universidade Federal de Pelotas – {mnfaria, lmtortelli, sdrutz, marilton}@inf.ufpel.edu.br

1. INTRODUÇÃO

De acordo com o último censo realizado no ano de 2010, cerca de 5,1% da população brasileira possui deficiência auditiva severa ou algum grau de deficiência auditiva (IBGE, 2010). Dessa forma, a preocupação acerca da necessidade por inclusão desses brasileiros é crescente, principalmente na questão da interpretação da linguagem usada por eles, que se difere em vários aspectos da linguagem natural. Portanto, abordagens para o reconhecimento de gestos de LIBRAS (Língua Brasileira de Sinais) fazendo uso de tecnologias assistivas são necessárias. Até então, os custos, tanto de equipamento quanto de processamento, tornavam essas tecnologias pouco viáveis e acessíveis no dia a dia.

Recentemente, novas tecnologias têm surgido para reconhecimento de movimentos, como por exemplo, o Microsoft Kinect™, que foi utilizado inicialmente para a interação com jogos. Esta tecnologia se mostrou útil na área de estudos que envolvem o reconhecimento de LIBRAS, pois além de coletar uma vasta gama de dados, como os pontos que mapeiam completamente o esqueleto humano, ele também é uma forma de tecnologia acessível e que não exige muito do usuário, já que basta se posicionar diante do Kinect™.

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 está inserido o projeto LIKI – *Libras e Kinect Aplicados às Tecnologias Assistivas*, atualmente em desenvolvimento e que é abordado neste trabalho. Este projeto busca propor modelos e ferramentas computacionais para auxiliar no processo de ensino-aprendizagem da LIBRAS para crianças com algum tipo de deficiência auditiva.

Desta forma, no contexto de reconhecimento de imagens, necessário para diferenciar os sinais apresentados pelo usuário, surge a necessidade da busca pelos melhores algoritmos de treino e reconhecimento onde, dados os parâmetros, um resultado satisfatório com a menor margem de erros possa ser apresentado. A plataforma utilizada para classificar os dados obtidos do Kinect™ foi o Weka (WEKA, 2014), que consiste em uma coletânea de algoritmos de aprendizado de máquina (SMOLA, 2008).

As análises foram feitas com diferentes abordagens, como por exemplo a de *K-Nearest Neighbours* (KNN), que dado um conjunto de características, encontra os K-vizinhos mais próximos que as possuem; e, também, a técnica de *Support Vector Machine* (SVM), que consiste em encontrar um par de hiperplanos paralelos que levam à máxima separação entre duas classes de características. Ambas estão presentes no trabalho de CORREIA (2013), que aborda a Linguagem Gestual Portuguesa, e se mostraram eficazes no reconhecimento de imagens através do Kinect™.

2. METODOLOGIA

Através da análise de registros históricos, percebe-se que durante muitos anos pessoas com algum tipo de deficiência eram, de certa forma, desligadas do convívio em sociedade (ARAÚJO, 2007). A Língua Brasileira de Sinais é a língua usada nas comunidades surdas, e se distingue por sua modalidade visual-gestual; na língua de sinais, o desenvolvimento se mostra lógico e aceitável para que o surdo possa se comunicar utilizando as mãos e também movimentos corporais (COUTO, 2014).

Tendo em vista a preocupação de desenvolver ferramentas computacionais de ensino-aprendizagem mais completas e abrangentes, se mostrou necessário o estudo de métodos que pudessem exercer a classificação de vídeos e apresentar resultados satisfatórios para o reconhecimento de padrões em gestos dinâmicos.

A *priori*, uma das abordagens para realizar a classificação de vídeos envolve dividi-lo em uma cadeia de *frames* e aplicar técnicas de reconhecimento de imagem estática. Porém, esse método dá margem para erros de interpretação, uma vez que as características analisadas em cada frame poderiam não fazer parte do tópico principal do vídeo (NG, 2015). Esta abordagem traz a necessidade de interpretar os dados de uma forma global, de modo que os acontecimentos descritos nos mesmos possam ser classificados de maneira coesa.

Assim, este trabalho utiliza duas arquiteturas de redes neurais convolucionais (*Convolutional Neural Network*, CNN), que vem ganhando destaque no reconhecimento de objetos (FERNANDES, 2013). A CNN trabalha com células simples e complexas, utilizadas para extração implícita de características dos padrões visuais apresentados como entrada e que posteriormente são integradas a uma rede completamente conectada.

Uma das perspectivas a ser analisada é a *Feature Pooling Architecture*, um recurso de agrupamento temporal que, pelo fato de ser utilizado em conjunto com redes neurais, cria uma camada de características que é atualizada sempre que características novas são criadas e não obrigatoriamente a cada frame (NG, 2015).

O segundo método a ser abordado é a *Long Short Term Memory Architecture* (*LSTM Architecture*). Posto que em vídeos os dados representados são dinâmicos, pode-se assumir que cada *frame* pode conter uma informação valiosa a respeito de um conjunto de características, podendo melhorar a precisão da classificação (NG, 2015).

Nesse método, que também trabalha com camadas, uma dada sequência de parâmetros é computada por uma rede neural de recorrência, e, a cada *frame*, a *LSTM Architecture* decide se determinada característica já identificada é relevante o suficiente para permanecer na camada e se novas características devem ser adicionadas.

3. RESULTADOS E DISCUSSÃO

Nesta seção, serão apresentados os conceitos mais importantes oriundos do estudo dos referenciais teóricos e tecnológicos no estágio atual de desenvolvimento do trabalho.

As arquiteturas FP e LSTM, apresentadas na seção anterior, mostram um grande potencial para a aplicação de classificação de vídeos relacionados à LIBRAS. Um dos grandes adicionais das técnicas apresentadas é a possibilidade

de processar vídeos relativamente longos, de vários segundos, e a capacidade de fazer isso em apenas um processamento (NG, 2015).

Ainda nesse âmbito, outra grande vantagem dessas arquiteturas é a de classificar gestos que envolvem movimentos corporais, dando um maior poder de expressão ao usuário das ferramentas computacionais em desenvolvimento, visando diminuir ao máximo qualquer tipo de limitação no exercício da língua gestual.

Além do uso nas ferramentas para o auxílio de ensino-aprendizagem, as técnicas apresentadas podem gerar um grande avanço no sentido de fornecer uma maior documentação para vídeos que retratam a LIBRAS, facilitando a busca dos mesmos e refinando os resultados para torná-los mais satisfatórios para o usuário.

4. CONCLUSÕES

A relativa exclusão de deficientes auditivos ainda é uma realidade no país, e a dificuldade de levar tecnologias assistivas até crianças e adolescentes também se faz muito presente. Por esses motivos, o desenvolvimento de ferramentas educacionais utilizando o Kinect™ se mostra eficaz, posto que, além do baixo custo, permite alcançar grande eficiência na inclusão e no aprendizado dos usuários da Língua Brasileira de Sinais.

Os métodos apresentados, que num futuro próximo serão utilizados na classificação de gestos dinâmicos da LIBRAS, poderão inclusive ser usados amplamente na área acadêmica para outros fins. Ainda há muito a ser analisado, mas os referenciais teóricos indicam que a exploração dessas novas arquiteturas pode trazer grandes inovações na área de reconhecimento e classificação de vídeos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO, D. M. S.. **A influência da LIBRAS no processo educacional de estudantes surdos em escola regular**. 2007. Trabalho de Conclusão de Curso (Graduação em Pedagogia) - Universidade Federal de Pernambuco.

CORREIA, M. M. **Reconhecimento de Elementos da Língua Gestual Portuguesa com Kinect**. 2013. Tese (Mestrado Integrado em Engenharia Eletrotécnica e de Computadores) – Universidade do Porto.

COUTO, L. F. Libras: uma análise histórica na perspectiva da educação inclusiva. **Revista Eletrônica Saberes da Educação**, v.5, n.1, p. 1 - 16, 2014.

FERNANDES, B. J. T. **Redes Neurais com Extração Implícita de Características para Reconhecimento de Padrões Visuais**. 2013. Tese (Doutorado em Ciência da Computação) - Universidade Federal de Pernambuco.

IBGE. **Resultados do Censo de 2010**. Acessado em 20 jul. 2015. Online. Disponível em: http://censo2010.ibge.gov.br/resultados.

NG, J. Y. H. Beyond Short Snippets: Deep Networks for Video Classification. **Cornell University Library**, v.2, p. 1 - 9, 2015.

RAMOS, C. R. **LIBRAS: a língua de sinais dos surdos brasileiros**. Editora Arara Azul, Petrópolis. Acessado em 24 jul. 2015. Online. Disponível em: http://www.editora-arara-azul.com.br/pdf/artigo2.pdf.

SMOLA, A., VISHWANATHAN, S. V. N. **Introduction to Machine Learning**. Cambridge: Cambridge University Press, 2008, 1v.

WEKA. **WEKA Manual**. Universidade de Waikato. Acessado em 24 jul. 2015. Online. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/documentation.html.