ENGENHARIA ELÉTRICA - COMPUTADORES

SISTEMA DE RECONHECIMENTO DE LÍNGUA BRASILEIRA DE SINAIS APLICADO A ROBÔS DE SERVIÇO

Aluno: Beatriz Barsocchi Testa (uniebtesta@fei.edu.br); Lucas Toledo Pastori (unielpastori@fei.edu.br); Mariana Aguirre De Oliveira (uniemarioliveira@fei.edu.br); Patrick De Sousa Pessoa (unieppessoa@fei.edu.br); Thiago Borges Coronado (unietcoronado@fei.edu.br)

Orientador: Reinaldo A. C. Bianchi (rbianchi@fei.edu.br)

Resumo

Com o avanço da inteligência artificial, das tecnologias de reconhecimento de voz e de visão computacional, os robôs de serviço estão ocupando espaços sociais com cada vez mais intensidade. Estes robôs pretendem auxiliar nas tarefas cotidianas e para isso precisam interagir com humanos de forma natural. Grande parte destes robôs utilizam a voz como forma principal de viabilizar esta comunicação. A comunidade surda se comunica por meio de uma forma não verbal da linguagem, portanto a interação somente por fala limita a capacidade dos robôs se comunicarem com essas pessoas. O projeto desenvolveu um sistema de reconhecimento de gestos em Libras (Língua Brasileira de Sinais) utilizando recursos de Inteligência Artificial para promover e incentivar a inclusão desta forma de interação nas novas tecnologias e aplicações.

Introdução

O HERA (*Home Environment Robot Assistant*), mostrado na Figura1, é o robô de serviço desenvolvido pela equipe RoboFEI@Home com o intuito de realizar tarefas em cooperação com humanos, interagindo com estes de forma autônoma. O time participa da competição RoboCup@Home e utiliza as tarefas desta como forma de validação das capacidades da plataforma robótica desenvolvida. A interação humano-robô é um dos focos de pesquisa da equipe, propiciar uma comunicação natural e fluída é um interesse, visto que este robô visa participar de atividades sociais. A principal forma de se comunicar com os humanos utilizada pelo HERA, assim como pela maioria dos robôs de serviço, é por meio da voz.

Figura 1- HERA

Conforme os dados do IBGE (Instituto Brasileiro de Geografia e Estatística) , de 2010, aproximadamente 9,7 milhões de brasileiros possuem algum grau de deficiência auditiva, representando cerca de 5,1% da população. Grande parte dessas pessoas se comunica por Libras. Essa língua é constituída por gestos que contêm parâmetros primários: configuração de mão (CM), ponto ou local de articulação (PM) e movimento (M); e secundários: Orientação/direcionalidade e Expressão facial e/ou corporal.

Os crescentes avanços na Inteligência Artificial, no processamento de imagens e da visão computacional oferecem ferramentas para que tornemos possível a realização de tarefas difíceis para uma máquina, como o reconhecimento de padrões por meio de imagens. Utilizando essas técnicas o projeto incluiu o reconhecimento de palavras em Libras para que o HERA realizasse o atendimento em um restaurante. Adicionar essa forma de comunicação às capacidades do robô pretende promover e incentivar a inclusão das formas de comunicação não verbal nas tecnologias.

Desenvolvimento

Tendo em vista a aplicação nos robôs de serviço foi levado em consideração que as palavras compreendidas na classificação deveriam ser facilmente ampliadas, uma vez que, os robôs atuam em diversos ambientes que exigem o reconhecimento de palavras utilizadas em diferentes contextos. Além disso a inferência dos gestos realizados deve ser em tempo real.

O sistema desenvolvido trata as imagens com um pacote de visão computacional, chamado CVZone, para simplificar as imagens deixando nelas somente as características mais relevantes para a identificação dos gestos realizados em Libras. Por meio deste pacote foi criada uma imagem em fundo preto com os pontos chave da mão e da face como é mostrado na Figura 2. A partir dessa imagens foi treinada uma rede neural convolucional para classificar as palavras escolhidas. Essa simplificação permite realizar treinamentos mais curtos, com base de dados menores e utilizando redes com menos parâmetros e de inferência mias rápida.

ENGENHARIA ELÉTRICA - COMPUTADORES

SISTEMA DE RECONHECIMENTO VISUAL DE LINGUAGEM BRASILEIRA DE SINAIS

Aluno: Beatriz Barsocchi Testa (uniebtesta@fei.edu.br); Lucas Toledo Pastori (unielpastori@fei.edu.br); Mariana Aguirre De Oliveira (uniemarioliveira@fei.edu.br); Patrick De Sousa Pessoa (unieppessoa@fei.edu.br); Thiago Borges Coronado (unietcoronado@fei.edu.br)

Orientador: Reinaldo A. C. Bianchi (rbianchi@fei.edu.br)

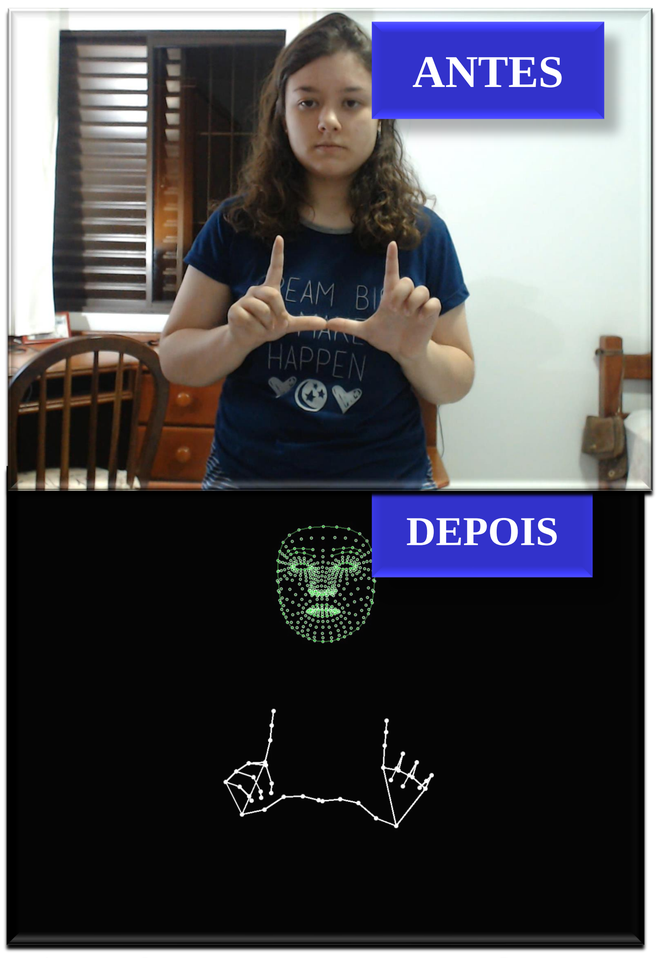

Figura 2- Tratamento com o CVZone

Para testar o sistema desenvolvido foi escolhida uma das tarefas realizadas pelo HERA na competição em que participa. A tarefa escolhida foi a restaurante, na qual o robô deve realizar o atendimento de pessoas e servi-las de acordo com os pedidos realizados por estas. Tendo definido o domínio em que a aplicação ocorreria, a realização do projeto foi separada em cinco principais etapas. A Figura 3 mostra um fluxograma das partes do projeto na ordem em que foram realizadas. Cada etapa esta descrita com mais detalhes abaixo:

1. Definição dos sinais em Libras para classificação:
Foi necessário definir os sinais em Libras que deveriam ser reconhecidos pelo sistema. Foram escolhidas 5 palavras de opção de pedido em um restaurante. Os gestos escolhidos foram: pastel, coxinha, sanduíche, água e refrigerante.

2. Gravação dos vídeos com os sinais definidos:
Com a definição dos sinais utilizados na detecção, foram gravados vídeos curtos (aproximadamente 10 segundos) para cada palavra escolhida. Os vídeos foram gravados por voluntários, totalizando 20 vídeos por gesto.

3. Tratamento no CVZone:
O pacote de visão computacional CVZone foi utilizado para a detecção e obtenção dos principais pontos das mãos e da face, colocados em um vídeo com fundo preto.

4. Gerar o Dataset para os sinais escolhidos:
Os vídeos após tratados com o CVZone foram cortados frame a frame para gerar as imagens para o treinamento. Para capturar a essência temporal dos gestos estes vídeos foram separados em suas partes inicial e final para serem classificados como parte um ou dois do movimento, assim tornando possível inferir se o movimento foi realizado por completo e com a direcionalidade correta. Por exemplo para a palavra água: o vídeo gravado realizando o gesto que corresponde a palavra contendo 10 segundos teria seus 5 primeiros segundos colocados como correspondentes a classe água1 e os 5 últimos na classe água2. Para movimentos mais complexos é possível separar em mais classes por palavra a fim de melhorar a classificação destes.

5. Treinar a rede neural EfficientNet:
Com o dataset pronto, foram realizados os treinamentos da rede neural EfficientNet B0 e B3. A rede B0 tem inferência mais rápida porém apresenta acurácia menor. Como foi possível realizar o processamento das imagens externamente ao robô foi treinada a rede B3, com mais parâmetros, mas ainda sim menor que redes como Inception e ResNet, pois para uma aplicação em tempo real a inferência deve ocorrer rapidamente e mesmo com o processamento externo durante as competições podem ocorrer oscilações na rede. Portanto é preferível ter uma rede de inferência rápida. As redes foram treinadas para reconhecer no total 5 palavras, com duas classes para cada.

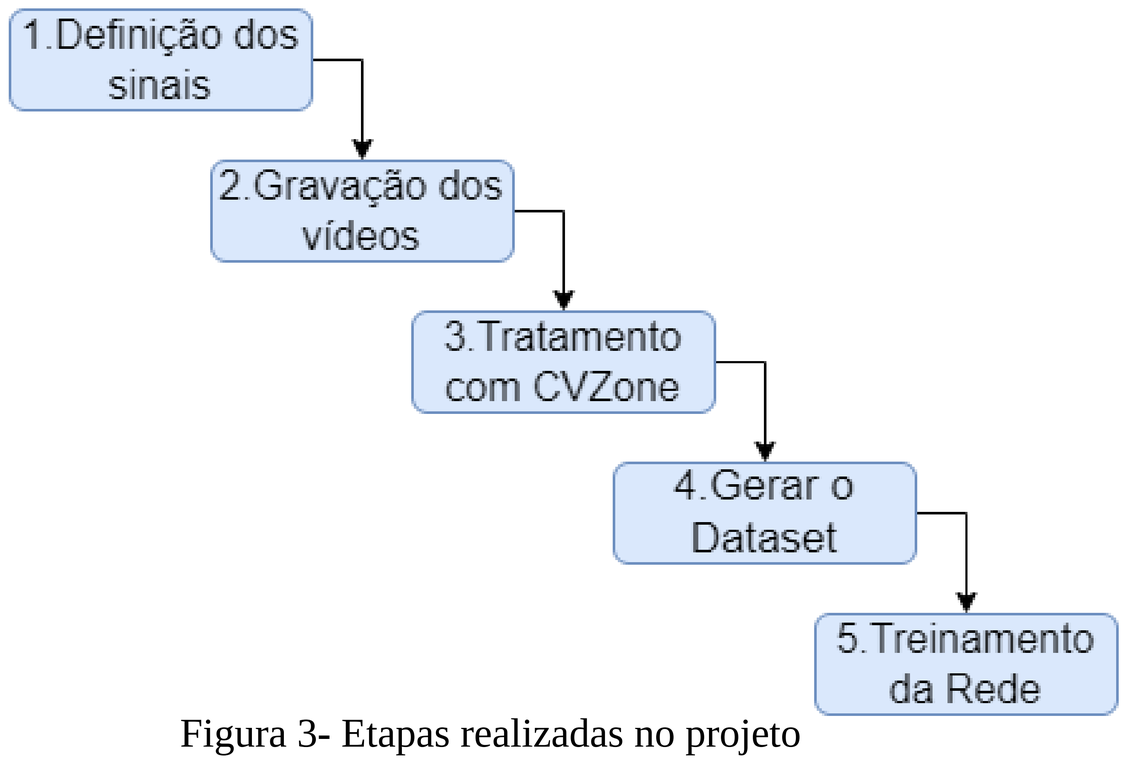

Figura 3- Etapas realizadas no projeto

Os resultados das redes EfficientNet B0 e B3 estão apresentados nas figuras 4 e 5, respectivamente. A acurácia da rede B0 foi de aproximadamente 71% e da B3 75%. Como a rede B3 não tem tempo de inferência muito superior ao da B0 e teve acurácia maior, esta foi a rede escolhida para implementar no robô. Essa escolha também foi influenciada pelo fato de ser possível realizar o processamento externo das imagens obtidas no robô. Como a plataforma robótica utiliza ROS (Robot Operating System) é possível enviar as imagens através da rede para serem processadas externamente ao robô.

ENGENHARIA ELÉTRICA - COMPUTADORES

SISTEMA DE RECONHECIMENTO VISUAL DE LINGUAGEM BRASILEIRA DE SINAIS

Aluno: Beatriz Barsocchi Testa (uniebtesta@fei.edu.br); Lucas Toledo Pastori (unielpastori@fei.edu.br); Mariana Aguirre De Oliveira (uniemarioliveira@fei.edu.br); Patrick De Sousa Pessoa (unieppessoa@fei.edu.br); Thiago Borges Coronado (unietcoronado@fei.edu.br)

Orientador: Reinaldo A. C. Bianchi (rbianchi@fei.edu.br)

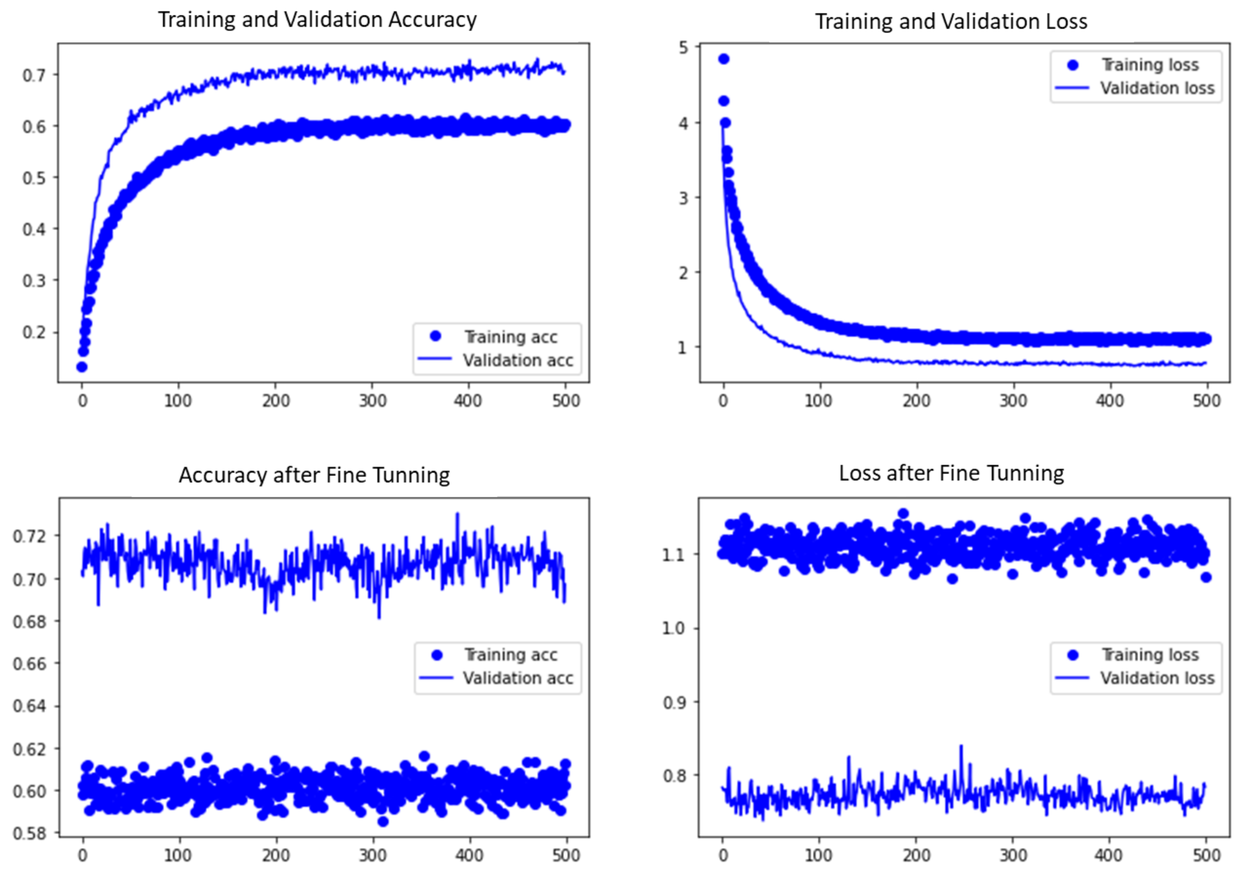

Figura 4- Resultados EfficientNet B0

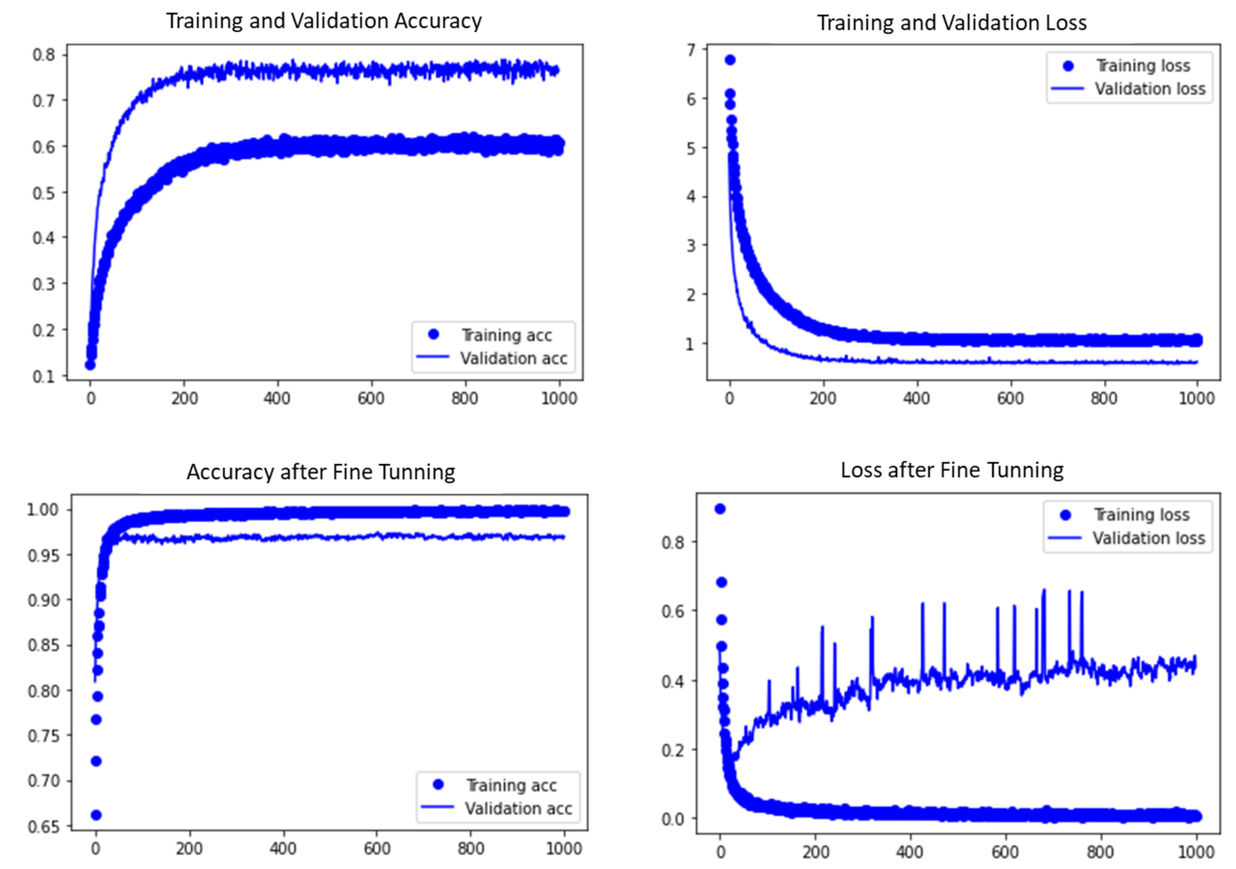

Figura 5- Resultados EfficientNet B3

Implementação no HERA

Para implementar o sistema para a realização da tarefa, outros aspectos de interação com humanos e com o ambiente devem ser levados em consideração. Para as interações com o ambiente, como navegar e pegar os pedidos foram utilizados os sistemas já desenvolvidos para tais tarefas pela equipe. Para as interações por meio da voz, utilizada na tarefa para solicitar o pedido para a balconista, também foi usado o sistema de voz que o robô já possui. Porém para a interação com o surdo, além do reconhecimento é necessário responder o usuário para tornar a comunicação mais natural. Foram gravados vídeos em Libras com legendas em texto que são apresentados na tela do HERA quando a resposta é necessária, a fim de mostrar uma das formas com que essa comunicação poderia ser realizada.

Os testes foram realizados simulando um serviço de atendimento ao cliente em um restaurante, este teste em um ambiente externo ao laboratório visou observar se os tempos de resposta do sistema eram compatíveis com a aplicação em tempo real, além disso notar possíveis falhas do sistema.

Para realizar essa tarefa, o robô deve identificar o cliente que está solicitando o atendimento e se dirigir até sua mesa, por meio do sistema de navegação autônoma, mantendo distâncias seguras das mesas e pessoas. Por meio de Libras e texto, o robô solicita o pedido ao cliente e infere, através da rede neural e por meio do processamento por uma máquina externa, qual a sua solicitação. Com os dados retornados pela rede e filtrados, confirma ao cliente a compreensão (ou não) do pedido e, caso tenha compreendido, vai até o balconista e solicita, por voz, o pedido do cliente. Após isso, busca o pedido, o entrega ao cliente e o informa para pegar seu pedido e chamar o robô novamente caso necessite de um novo atendimento.

Conclusão

O projeto implementado no robô HERA, testado em uma das tarefas realizada por este na competição, foi capaz de identificar as palavras escolhidas e responder ao usuário em tempos compatíveis com a aplicação durantes os testes realizados. Atuando em conjunto com os outros sistemas do robô, o reconhecimento dos gestos permitiu que o robô realizasse um atendimento que inclui uma nova forma de se comunicar com os humanos.

O sistema tem potencial para ser aplicado em novos contextos. A inclusão de novas palavras não exige grande base de dados, o que facilita a expansão do sistema. A rede foi treinada com imagens obtidas a partir de 20 vídeos de cada gesto, aproximadamente 1000 imagens por classe. Isso foi possível pela simplificação das imagens realizada a partir do tratamento pelo CVZone. Com a expansão do número de palavras identificadas o sistema pode, em trabalhos futuros, ser utilizado em conjunto com a interpretação semântica para o reconhecimento de frases.