# RECONHECIMENTO DE SINAIS DA LIBRAS POR VISÃO COMPUTACIONAL

Giovanna Ono Koroishi, Bruna Vieira Louzada Silva

**Resumo -** O reconhecimento automático linguagens de sinais promete facilitar a comunicação entre surdos e o restante da população. As tecnologias atuais esbarram em obstáculos como dificuldade de rastreio das mãos e reconhecimento de poses. Este projeto estuda uma abordagem probabilística para a identificação de sinais da LIBRAS com um sensor RGBD. Essa identificação é baseada na classificação SematosÊmica da linguagem. Durante o projeto, foi implementado um identificador automático de sinais da LIBRAS que captura o vídeo através do sensor Kinect, segmenta o objeto de interesse (a mão direita) e calcula a probabilidade dos sinais. O principal desafio do trabalho foi reconhecer as configurações de mão e para solucioná-lo utilizou-se modelos estruturados em nuvens de pontos e o algoritmo de ICP. O projeto mostrou que essa abordagem torna viável o reconhecimento automático, pois obteve-se 65% de acerto entre 48 testes envolvendo 12 sinais diferentes, mesmo com as limitações de recursos e tempo existentes.

Palavras-chave: LIBRAS. Reconhecimento automático. Visão computacional. Kinect. Nuvem de pontos.

## 1 Introdução

De acordo com a Organização Mundial da Saúde (WORLD HEALTH ORGANIZATION, 2014), cerca de 360 milhões de pessoas possuem surdez incapacitante. Isso significa que mais de 5% da população mundial possui, no ouvido mais aguçado, perda auditiva maior que 40dB em adultos e 30dB em crianças, englobando tipos de suave a severo. Entre essa minoria, há aqueles cuja principal consequência da surdez é o impacto na comunicação verbal. Nesses casos, a comunicação pode ser feita de diferentes formas, tais como a leitura labial, a escrita, a leitura e a linguagem de sinais. Entretanto, as três primeiras formas são intrinsecamente ligadas à linguagem falada, enquanto a última possui suas próprias regras e estrutura gramatical (GUIMARÃES et al., 2010), mostrando que tanto a linguagem de sinais quanto a falada são independentes e passíveis de tradução.

Há um grande desenvolvimento de tecnologias de tradução no sentido da linguagem falada para linguagem de sinais, contudo, o sentido contrário possui o desafio técnico da visualização e do reconhecimento dos sinais. No caso brasileiro, esse desafio se mostra ainda mais complexo, já que a LIBRAS (Linguagem Brasileira de Sinais) possui rica diversidade dos SematosEmas (conceito detalhado a seguir) de articulação de mão, sendo muito importante distinguir a configuração das mãos e dos dedos, o que torna mais complexo o reconhecimento dos MorfEmas (conceito detalhado a seguir) (CAPOVILLA; RAPHAEL; MAURICIO, 2013).

## 2 Linguagem Brasileira de Sinais

Dois conceitos muito importantes da língua utilizados no trabalho são: SematosEma e MorfEma (CAPOVILLA; RAPHAEL; MAURICIO, 2013). MorfEma é a menor unidade sublexical que codifica significado; enquanto SematosEma é a menor unidade sublexical da sinalização capaz de distinguir um sinal do outro, ou seja, é o detalhamento de como articular o sinal de acordo com a classificação: articulação da mão, local da articulação, movimento e expressão facial (quando

aplicável), a qual pode ser utilizada como base para o sistema de reconhecimento de MorfEmas a partir dos SematosEmas, indicando quais variáveis devem ser analisadas.

Um exemplo desse conceito são os sinais estudar e universidade. O sinal estudar é, por sí só, um MorfEma composto pelos SematosEmas: mãos abertas, palmas para cima, bater duas vezes o dorso dos dedos direitos sobre a palma dos dedos esquerdos (CAPOVILLA; RAPHAEL; MAURICIO, 2013). Percebe-se, por esse exemplo, que a descrição SematosÊmica é suficiente para a realização de um sinal. O sinal universidade, por sua vez, é composto pelo sinal estudar, seguido dos SematosEmas: mão em U, palma para frente, movê-la em um círculo vertical para a esquerda no sentido anti-horário (CAPOVILLA; RAPHAEL; MAURICIO, 2013).

O reconhecimento de sinais compostos como universidade, entretanto, é um problema relacionado ao processamento de linguagem natural, o qual não seria possível de ser abordado dado o prazo de um ano do projeto. Por isso, o protótipo foi limitado ao reconhecimento de sinais constituídos de apenas um MorfEma.

# 3 Requisitos do Projeto

## 3.1 Requisitos Funcionais

O processamento do reconhecimento dos sinais deve ser realizado em um desktop com configurações padrões e os dados utilizados para análise do sinal devem ser captados por um sensor que utilize imagens, de forma a tornar o uso do protótipo o mais natural possível.

O protótipo deve identificar os SematosEmas que compõem o MorfEma e entregar as probabilidades de cada SematosEma para, então, calcular a probabilidade do MorfEma. Dentre os SematosEmas que podem compor sinais da LIBRAS (CAPOVILLA; RAPHAEL; MAURICIO, 2013), foram selecionados 39 para serem identificados.

## 3.2 Requisitos Não-Funcionais

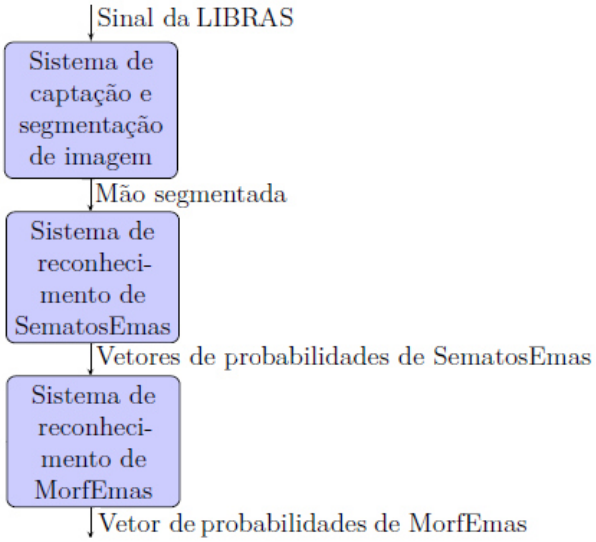
O protótipo deve ser facilmente operado por um não ouvinte, possível de ser utilizado em um desktop com configurações padrão e possuir sistema de segmentação suficientemente eficiente para, uma vez identificadas as probabilidades dos SematosEmas, permitir que o sistema de reconhecimento identifique o MorfEma correto. Além disso, o sistema deve ser uma plataforma de pesquisa para avaliar a viabilidade do reconhecimento de sinais da LIBRAS por visão computacional.

# 4 Arquitetura

O estudo dos trabalhos já realizados e de sinais da LIBRAS mostrou que o reconhecimento de determinados SematosEmas é complexo, pois existem SematosEmas bastante parecidos e podem haver imprecisões praticadas pelo próprio usuário. A sensibilidade dos sensores e as técnicas de segmentação conhecidas podem levar à não correspondência ou correspondência errada entre os SematosEmas reconhecidos e o MorfEma realizado. Para minimizar essas distorções, propôs-se uma abordagem probabilística para os sistemas.

O protótipo foi, então, dividido em três sistemas independentes, de modo que o reconhecimento do sinal pudesse ser feito como mostrado na Figura 1.

Figura 1 - Arquitetura do sistema.




O primeiro sistema capta o vídeo e segmenta a mão direita. O segundo sistema recebe a mão segmentada e, a cada frame, tem como saída três vetores: probabilidade de similaridade com cada SematosEma, probabilidade de movimentos, probabilidade de localização da mão. Por fim, essas informações são reconhecidas e avaliadas pelo terceiro sistema, o qual classificará os vetores de probabilidades de acordo com as estruturas descritas por (CAPOVILLA; RAPHAEL; MAURICIO, 2013), determinando assim, qual o possível sinal realizado.

# 5 Metodologia

## 5.1 Captura e segmentação

O sensor escolhido para o projeto é o Kinect (XBOX, 2014), um sensor RGBD da Microsoft, cuja taxa de captura é de até 30 frames/s. Porém, como esta taxa produz diferenças mínimas entre os frames e o tempo de processamento do sinal é diretamente relacionado à quantidade de frames, este projeto trabalha com taxa de 8 frames/s. Os frames são processados com a utilização do próprio SDK da Microsoft para o Kinect.

## 5.2 Segmentação da mão

O processo de segmentação consiste em definir uma região de interesse do frame completo e selecionar apenas os dados dos pontos dessa região. O SDK do Kinect já implementa funções de localização de juntas de um corpo humano, com isso, obtém-se as coordenadas da junta "mão direita" no espaço, segmenta-se a região da mão definindo um paralelepípedo ao redor da junta e seleciona-se os pontos nessa região.

## 5.3 Reconhecimento SematosÊmico

### *A SematosEma de Movimento*

O SematosEma de movimento é avaliado através da diferença da posição da mão direita entre um frame e o seu anterior. Calcula-se o cosseno entre o vetor de deslocamento e os seis semi-eixos que representam os movimentos para cima, baixo, direita, esquerda, frente e trás, sendo o menor ângulo corresponde ao movimento o mais provável. A probabilidade de cada movimento ter sido realizado é calculada para o conjunto anterior de frames com mesmo movimento no momento em que se detecta a mudança de direção e/ou sentido entre dois frames consecutivos. Essa probabilidade é inversamente proporcional ao ângulo formado entre o vetor descolamento e cada

semi-eixo. Já a probabilidade da mão não ter se movido é calculada pela função sigmoide:

$$1/(1 + e^{(d-t)/k}), \quad (1)$$

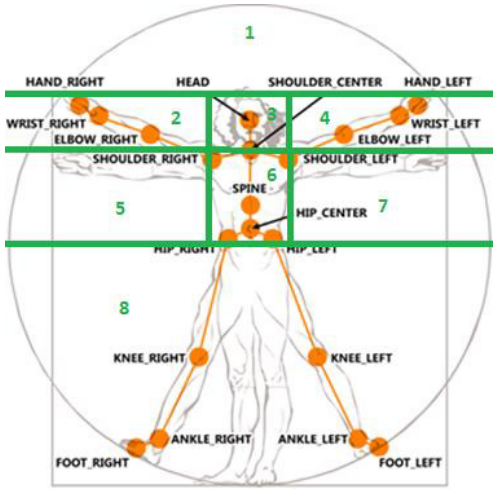
onde d é a amplitude detectada do movimento, t é a distância mínima para que seja considerado movimento e k é a distância a mais em relação a t em que ainda há 25% de probabilidade de não ter ocorrido o movimento.

*B SematosEma de Local de Articulação*

O SematosEma de local de articulação é reconhecido através das coordenadas das juntas. O espaço de reconhecimento divide-se em: acima da cabeça, à direita da cabeça, em frente ao rosto, à esquerda da cabeça, à direita da cintura, em frente à cintura, à esquerda da cintura e abaixo da cintura e as coordenadas da mão direita são avaliadas em referência a outras juntas, conforme a Figura 2.

Testes mostraram que o sensor e a função de identificação de juntas são precisos, por isso, a probabilidade de que a mão esteja na região avaliada é considerada 0.8; para as regiões fronteiriças, 0.15/(nº regiões fronteiriças); e para as regiões mais distantes, 0.05/(nº regiões distantes).Se realmente for necessário usar um nível a mais, divida a sub-seção em sub-itens.

Figura 2 - Regiões dos SematosEmas de local de articulação. Adaptado de Microsoft Developer Network.

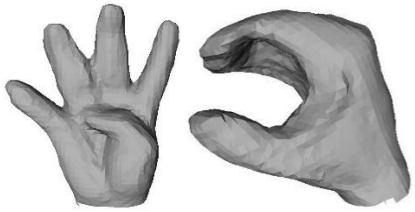



*C SematosEma de Configuração de Mão*

O reconhecimento da articulação da mão e dos dedos é o processo mais difícil no reconhecimento dos sinais. Após estudar vários métodos, optou-se pelo reconhecimento por modelo, ou seja, o frame capturado é comparado com modelos dos SematosEmas pré-definidos.

Essa abordagem necessitava de um banco de dados de SematosEmas de articulação de mãos e para criá-lo foi utilizado o próprio Kinect e o programa KinectFusion (NEWCOMBE et al., 2011) (IZADI et al., 2011) para escanear mãos nas configurações escolhidas, como pode ser observado na Figura 3.

Figura 3 - Modelos de mão em 4 e em C, respectivamente



Esses modelos e o frame são armazenados na forma de nuvens de pontos, que são coordenadas espaciais de vários pontos que compõem o modelo ou a mão.

A nuvem de pontos capturada é comparada com os modelos utilizando o algoritmo Iterative Closest Point (ICP) implementado pela biblioteca Point Cloud Library (PCL) (RUSU e COUSINS, 2011). O ICP é um algoritmo iterativo de otimização, que minimiza a distância entre os pontos de duas nuvens por meio de movimentos de corpo rígido, rotações e translações.

Para reconhecer o MorfEma é necessário saber qual a probabilidade de um SematosEma de articulação de mão ter sido realmente realizado, tendo os resultados do ICP para os dados capturados com o sensor. Tal problema é bastante complexo, podendo ser considerado como um refinamento do reconhecimento de membros do corpo realizado pelo Kinect, o qual foi solucionado pela Microsoft por meio de machine learning, utilizando muitos recursos computacionais e um extenso e diversificado banco de dados (SHOTTON et al., 2013). Entretanto, como esse trabalho, não dispôs de tanto tempo e nem de recursos tão abrangentes, o cálculo das probabilidades dos SematosEmas de articulação de mão se baseou em um procedimento bem mais simples de aprendizado de máquina.

Realizaram-se três testes para cada articulação de mão disponível no banco de modelos e com estes dados, observou-se a correspondência entre SematosEma realizado e as respostas dadas pelo ICP. Pressupôs-se que o usuário realizou, necessariamente, algum dos SematosEmas presentes no banco e para limitar o escopo da resolução do problema. Para cada frame, os resultados do ICP foram ordenados do melhor modelo alinhado para o pior e utilizou-se apenas os cinco primeiros lugares de forma qualitativa (a colocação do modelo é utilizada e não o valor bruto do alinhamento). Além disso, para incorporar os erros não-modelados do sensor e do algoritmo de alinhamento, foi considerado um termo de distribuição homogênea de probabilidades para todos os SematosEmas.

Quando o usuário realiza um sinal com a mão na configuração C1, o frame será comparado com todos os modelos e os cinco mais bem alinhados serão M1, M2, M3, M4 e M5, onde o Mi é o i-ésimo mais bem alinhado.

Para a configuração de mão S1, P1 é a probabilidade do usuário ter realizado S1, dado que M1 foi o melhor alinhamento e é calculada por:

$$\frac{\text{total de } M1 \text{ em } 1º \text{ lugar para } S1}{\text{total de } M1 \text{ em } 1º \text{ para todos SematosEmas}}. \quad (2)$$

P2 é a probabilidade do usuário ter realizado S1, dado que M2 foi o segundo melhor alinhamento e é calculada por:

$$\frac{\text{total de } M2 \text{ em } 2º \text{ lugar para } S1}{\text{total de } M2 \text{ em } 2º \text{ para todos SematosEmas}}. \quad (3)$$

P3 é a probabilidade do usuário ter realizado S1, dado que M1, M2, M3, M4 os cinco melhores alinhamentos e é calculada por:

$$\frac{1}{5}\sum_{i=1}^{5}\frac{\text{total de } Mi \text{ entre os 5 primeiros para } S1}{\text{total de } Mi \text{ entre os 5 para todos SematosEmas}}. \quad (4)$$

P4 é a distribuição homogênea de probabilidades, a qual corresponde a 1/(total de

SematosEmas), no caso do projeto, 1/24.

Calcula-se então para cada configuração S de mão:

$$P(S) = w1 \cdot P1 + w2 \cdot P2 + w3 \cdot P3 + w4 \cdot P4. \quad (5)$$

Os pesos w1, w2, w3 e w4 foram escolhidos respeitando o critério de que a primeira resposta tem maior influência no resultado, seguida da segunda e assim sucessivamente. Alguns pesos foram escolhidas arbitrariamente e testados, sendo que os que apresentou melhores resultados foram 0.6, 0.25, 0.1 e 0.05 para w1, w2, w3 e w4, respectivamente.

### 5.4 Reconhecimento do MorfEma

O reconhecimento do MorfEma é realizado após a captação do sinal. Primeiramente, analisam-se as probabilidades dos movimentos e segmenta-se o sinal através do movimento mais provável entre cada frame, isto é, frames com probabilidades de movimentos iguais são considerados como pertencentes a um único segmento do MorfEma. Paralelamente a isto, o sinal é filtrado, retirando-se frames em que há transição da direção do movimento para diminuir o ruído causado pela dificuldade em sincronizar a troca de movimento com a taxa de captura do sistema.

Segmentado o sinal pelo movimento, a probabilidade de cada MorfEma do bando de dados (composto por 12 MorfEmas) ter sido realizado é a multiplicação da probabilidade para cada um dos seus segmentos. Esta, por sua vez, é feita com outras quatro probabilidades: movimento do segmento, posição da mão no início do segmento, posição da mão no final do segmento e média das probabilidades do SematosEma de configuração de mão ao longo dos frames do segmento. A probabilidade de MorfEmas com número de segmentos diferente do capturado é zero, pois a probabilidade dos segmentos excedentes ou faltantes terem sido feitos é zero.

### 5.5 Paralelização

O cálculo do ICP é o processo de maior custo computacional, correspondendo a até 98,6% do tempo total do reconhecimento do sinal. Todavia, como cada alinhamento depende apenas do frame atual e de um modelo, é possível paralelizar o processo, de forma que cada alinhamento seja feito em uma thread independente da thread que contém o programa principal de análise dos sinais.

É importante notar, porém, que o processamento das threads dependem diretamente de todos os resultados do ICP para um frame, por isso, é preciso assegurar que todas terminaram antes de calcular as probabilidades.

Com a paralelização, conseguiu-se que tempo médio de alinhamento de um frame com todos os modelos fosse 6.5 vezes mais rápido que o obtido inicialmente.

## 6 Protótipo

O programa de reconhecimento de sinais foi desenvolvido em linguagem C++ no ambiente de desenvolvimento VisualStudio, utilizando as bibliotecas do Kinect e PCL. Os testes foram realizados em um desktop com características descritas na Tabela 1.

Tabela I - Configurações do computador utilizado no projeto

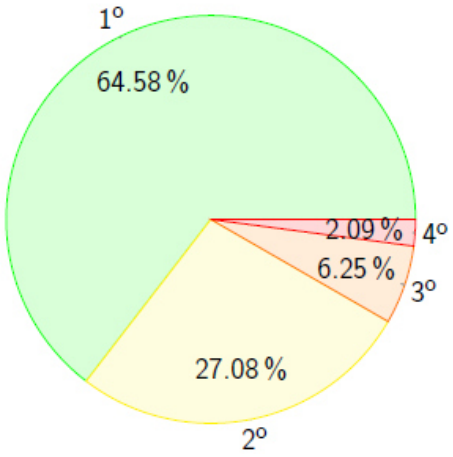
| | |
|---|---|
| **Processador** | Intel® Core™ i7-3970X CPU @ 3.50 GHz |
| **Sistema Operacional** | Windows 7 Professional 64-bit |
| **Memória** | 24.0 GB |
| **Espaço em disco** | 1 TB |

Através de testes foi confirmado que a taxa de 8 frames/s para a captura do sinal era suficiente. Além disso, foi estipulado o tempo de realização do sinal em 25 frames pois os sinais de teste podem ser feitos nesse período.

## 7 Resultados

O sistema foi testado 48 vezes (quatro vezes para cada um dos MorfEmas no banco de dados), sendo que o usuário em todos os testes é a mesma pessoa cuja mão compõe o banco de dados dos modelos de SematosEmas. Dentro dessa amostra, o sistema reconheceu corretamente 31 sinais, correspondendo a uma taxa de acerto de 65%. Por outro lado, entre os 17 sinais não reconhecidos, o sinal correto foi classificado 13 vezes como o segundo mais provável, 3 vezes como terceiro e somente uma vez como quarto. A distribuição dos resultados pode ser vista no Gráfico 1.

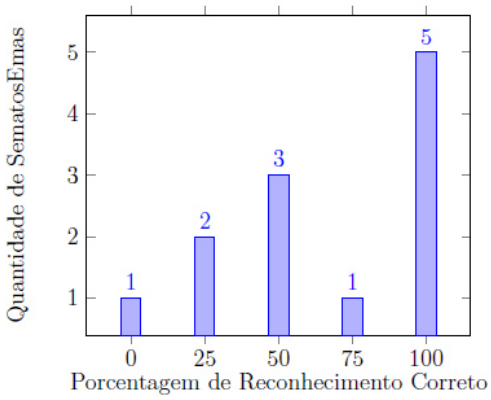
Gráfico 2 - Resultados dos testes: posições da resposta correta na resposta do sistema




Verificando a taxa de acerto em que o sinal correto é classificado em primeiro e segundo lugar (91,67%) têm-se um indício de que, apesar de não trabalhar com o processamento de linguagem natural, o sistema pode proporcionar reconhecimentos próximos o suficiente para possibilitar o entendimento de uma sequência de sinais. Entretanto, nos casos em que o sinal real aparece como segundo mais provável, caberia ao usuário atestar o significado através do contexto.

Analisando-se, também, a porcentagem de acerto individual de cada SematosEma nos quatro testes realizados para cada, obteve-se os resultados apresentados no Gráfico 2.

Gráfico 2 - porcentagem de acerto de SematosEmas em primeiro lugar pela quantidade.



## 8 Conclusão

Os testes realizados com o protótipo apresentaram resultados satisfatórios. A viabilidade da proposta original fica assim demonstrada. O projeto também mostrou que a abordagem de reconhecimento do MorfEmas através da avaliação de seus SematosEmas é bastante promissora, já que, mesmo que um SematosEma seja avaliado erroneamente, tanto por presença de ruídos ou por limitação do sistema, a utilização de todos os SematosEmas faz com que a resposta final esteja próxima do esperado. Além disso, a existência da classificação dos sinais da LIBRAS por SematosEmas permite dividir o problema de reconhecimento dos sinais de forma automática por meio de visão computacional em problemas menores.

O sistema, entretanto, possui limitações, tais como reconhecimento de um número reduzido de SematosEmas em relação a todos os existentes; tempo de resposta longo e variável, não permitindo o processamento on-line; e tempo de captura fixo em, aproximadamente, 3 segundos para cada sinal. Tais limitações, porém, são contornáveis.

## 9 Trabalhos Futuros

Como sugestões de trabalhos futuros há a implementação da identificação de mais SematosEmas, como de articulação de mão não utilizados neste trabalho, de orientação da palma da mão, de configuração do braço e de expressão facial. Tais implementações aumentarão o custo computacional e o tempo do reconhecimento ainda mais. Para melhorar isso seria aconselhável otimizar a implementação do algoritmo ICP, tanto em tempo quanto em precisão. Também seria ideal que o banco de dados utilizasse mais pessoas e mais testes na sua composição, para que o aprendizado de máquina seja mais efetivo.

## REFERÊNCIAS

CAPOVILLA, F. C.; RAPHAEL, W. D.; MAURICIO, A. C. L. Novo Deit-LIBRAS: dicionário enciclopédico ilustrado trilíngue da Língua de Sinais Brasileira baseado em linguística e neurociências cognitivas. 3. Ed. Rev. e ampl. São Paulo, INEP/CNPq/EDUSP, 2013. 2v

GUIMARÃES, C. et al. Technological artifacts for social inclusion: structure of the Brazilian sign language (LIBRAS), gestures for citizenship. In IADIS INTERNACION CONFERENCE WWW/INTERNET. Proceedings. [s.L.]: Timisorar, 2010, p. 267-271.

IZADI, S. et al. Kinect Fusion: real-time 3D reconstruction and interaction using a moving depth camera. In: ANNUAL ACM SYMPOSIUM ON USER INTERFACE SOFTWARE AND TECHNOLOGY, 24., Santa Barbara,CA. Proceedings. New York: Association for Computing Machinery, 2011. p. 559.

NEWCOMBE, R. A. et al. KinectFusion: real-time dense surface mapping and tracking. In: IEEE INTERNATIONAL SYMPOSIUM ON MIXED AND AUGMENTED REALITY, 10., Basel, Switzerland, 2011. ISMAR, [s.L.]: IEEE Computer Society, 2011. p. 127–136.

RUSU, R. B.; COUSINS, S. 3D is here: Point Cloud Library (PCL). In: IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION, Shanghai, 2011. ICRA. [S.I.]: IEEE, 2011.

SHOTTON, J. et al. Real-time human pose recognition in parts from single depth images. Communications of the ACM, v. 56 n. 1, p. 116-124 Jan. 2013. Disponível em: < http://dl.acm.org/ft_gateway.cfm?id=2398381&type=html >. Acesso em: 8 Abr. 2014.

XBOX. Kinect. Disponível em: <, http://www.xbox.com/en-US/kinect >. Acesso em: 8 Abr. 2014.

WORLD HEALTH ORGANIZATION. Deafness and hearing loss. Disponível em: <http://www.who.int/mediacentre/factsheets/fs300/en/>. Acesso em 5 Abr. 2014.

# LIBRAS SIGN RECOGNITION BY COMPUTER VISION

**Abstract -** Development of technologies towards the accessibility of the disabled is a subject that stimulates researchers all over the world. Specifically in the field of communication between deaf and listeners, is notable the development of tools to translate from the spoken language to the sign language, while the opposite direction is more technically challenging, due to the visualization and sign recognition. It is even more complex to automatically recognize the Brazilian signs, since LIBRAS widely uses the fingers to express the signs. On the other hand, the LIBRAS signs have a cheremic classification that allowed to split the problem into simpler ones. This project is a study of the feasibility of using 3D sensors to automatically recognize a set of LIBRAS signs, based on this classification to a probabilistic approach. During the project, a proof of concept of an automatic LIBRAS sign recognizer was implemented. The developed prototype for this verification records the signs by the Kinect sensor, segments the object of interest (the right hand) and calculates the sign probabilities. The main challenge of the work was to recognize the hand's configuration, models structured in cloud points and the ICP algorithm were used to solve them. The project showed that this approach makes automatic recognition feasible, as it reached the level of 65% of correct signs in 48 tests with 12 different signs, even with the limited resources and time.

Keywords: LIBRAS. Automatic recognition. Computer vision. Kinect. Cloud points

**Giovanna Ono Koroishi**, formada em Engenharia Mecatrônica pela Escola Politécnica da USP em janeiro de 2015 e mestranda em Engenharia Mecânica pela mesma instituição, trabalha como desenvolvedora e atua no desenvolvimento de soluções que utilizam modelos de dados aeronáuticos e de voo e de soluções de comando e controle.

**Bruna Vieira Louzada Silva**, formada em Engenharia Mecatrônica pela Escola Politécnica da USP em janeiro de 2015. Como analista de software, desenvolve e presta suporte para automação financeira utilizando a arquitetura CEN/XFS.