SBrT 2019 1570551480

Reconhecimento de gestos em vídeos utilizando modelos ocultos de Markov e redes neurais convolucionais aplicado a Libras

Vinícius Morais Breda e Danilo Silva

***Resumo*— Este trabalho propõe um sistema para reconhecimento de uma sequência de gestos dinâmicos em vídeos utilizando uma combinação de redes neurais convolucionais para modelar as formas da mão e modelos ocultos de Markov para modelar os gestos. São utilizados um total de 15 sinais da Libras. A acurácia obtida atingiu 100% sob certas restrições, como um único gesticulador, a utilização de luvas e um ambiente controlado.**

***Palavras-Chave*— Reconhecimento de imagens, reconhecimento de gestos, Libras, CNN, HMM.**

***Abstract*— This work proposes a system for recognition of a sequence of dynamic gestures in videos using a combination of convolutional neural networks to model the hand shapes and hidden Markov models to model the gestures. A total of 15 signs from the Brazilian Sign Language are used. The accuracy achieved reached 100% under certain restrictions, such as a single gesticulator, the use of gloves and a controlled environment.**

***Keywords*— Image recognition, gesture recognition, Libras, CNN, HMM.**

I. Introdução

Os gestos são uma forma natural de comunicação e expressão de emoções, sendo muito importante a capacidade de máquinas poderem reconhecê-los para realizar diversas funções, que vão desde o controle de periféricos até a identificação de gestos suspeitos. Uma das principais áreas no reconhecimento de gestos é o reconhecimento dos gestos realizados pelas mãos. Há vários trabalhos na área de reconhecimento de gestos manuais, sendo a maior parte focado no reconhecimento de gestos estáticos (reconhecimento do formato da mão) ou no reconhecimento de gestos dinâmicos (com movimento da mão) isolados.

Em [1] são utilizadas redes neurais artificiais (ANNs) para reconhecer dez posturas estáticas da mão, obtendo uma acurácia de 90%. Outra técnica mais moderna é o uso de redes neurais convolucionais (CNNs) para reconhecer as posturas das mãos, como é feito em [2]. As CNNs se provaram ótimos reconhecedores de padrões [3] [4] gerando resultados superiores a outras técnicas mais clássicas como SVM ou ANNs. Apesar das CNNs serem normalmente utilizadas para o reconhecimento de imagens estáticas, como a forma da mão, em [5] elas são utilizadas para fazer o reconhecimento de gestos dinâmicos isolados.

Vinícius Morais Breda, Universidade Federal de Pelotas, Pelotas-RS, e-mail: vmbreda@ufpel.edu.br; Danilo Silva, Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, Florianópolis-SC, e-mail: danilo.silva@ufsc.br.

Os modelos ocultos de Markov (HMMs) já são consagrados no reconhecimento da fala, sendo ideais para modelar processos temporais. Por isso eles também são muito utilizados no reconhecimento de gestos dinâmicos. Em [6] são utilizados HMMs para fazer o reconhecimento de frases contínuas com 40 sinais da Língua Americana de Sinais. Já em [7] são utilizados HMMs para reconhecer 262 gestos isolados. Uma abordagem diferente é feita em [8] e [9], onde cada gesto é modelado por mais de um HMM.

Neste trabalho é feito o reconhecimento de 15 sinais da Língua Brasileira de Sinais (Libras) realizados de forma contínua, sendo o ambiente controlado e necessária a utilização de luvas coloridas. Ao contrário de outros trabalhos que utilizam somente HMMs para reconhecer gestos dinâmicos e CNN para gestos estáticos, este utiliza a combinação de ambos, sendo utilizada uma rede de HMMs como classificador da sequência de gestos e uma CNN responsável por gerar descritores relativos a forma das mãos, que são adicionados a outros descritores espaciais e geométricos. O sistema é capaz de fazer a aquisição, processamento e classificação em tempo real, obtendo resultados promissores.

II. Modelos Ocultos de Markov

Um HMM [10] [11] é um tipo de modelagem estocástica apropriada para sequências estocásticas não estacionárias, de forma a modelar uma sequência de observações como um conjunto de processos estacionários. O HMM é basicamente uma máquina de estados estocástica, a qual é constituída por um conjunto de estados, onde cada estado é responsável pela emissão de um símbolo de acordo com uma distribuição de probabilidade. A transição entre os estados ocorre de acordo com um conjunto de probabilidades chamado de probabilidades de transição. Os modelos são chamados de ocultos devido ao fato de termos acesso apenas as observações emitidas pelos estados, e não aos estados que emitiram as observações. De acordo com a sequência de observações pode-se estimar qual a sequência de estados mais provável de tê-las gerado.

A figura 1 mostra um HMM com quatro estados, onde cada estado possui uma probabilidade de emitir determinado símbolo s. Pode-se ver que as transições entre os estados ocorrem de forma que, no instante inicial o estado ativo seja o primeiro, e a cada instante o estado ativo deve ser ele mesmo ou o próximo. Este é um exemplo de um HMM do tipo esquerda-direita, porém estas transições podem ocorrer de qualquer maneira possível, dependendo do que se deseja modelar.

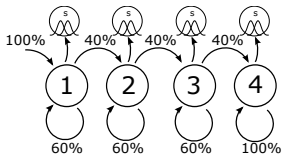

Fig. 1. Exemplo de um HMM.

O conjunto de parâmetros que os estados podem emitir são chamados de símbolos, e os símbolos emitidos pelos estados em determinada realização do processo, de observações. Logo uma realização do processo gera uma sequência de observações $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_T}$ proveniente da sequência de estados $\mathbf{y} = y_1, y_2, \ldots, y_T$.

Um HMM pode ser definido pelos seguintes parâmetros: o conjunto de distribuições de probabilidade de emissão de símbolos de cada estado $\mathbf{b}$; as probabilidades de transições entre os estados $\mathbf{A}$; a probabilidade inicial de cada estado $\boldsymbol{\pi}$. Então um HMM pode ser representado por $\lambda = (\mathbf{A}, \mathbf{b}, \boldsymbol{\pi})$. Dado um modelo $\lambda = (\mathbf{A}, \mathbf{b}, \boldsymbol{\pi})$ e uma sequência de observações $\mathbf{X}$, podemos estimar $p\{\mathbf{X}|\lambda\}$ ou encontrar a sequência de estados $\mathbf{y}$ ótima. O treinamento de um modelo consiste em otimizar os parâmetros λ do modelo de modo a maximizar $p\{\mathbf{X}|\lambda\}$, sendo $\mathbf{X}$ uma sequência de treinamento.

A. HMMs Aplicados ao Reconhecimento de Gestos

No caso da mão realizando um gesto, podemos dividir esse gesto em uma sequência de processos estacionários, e dessa forma modelá-lo por um HMM. Sendo assim, os HMMs podem modelar e fazer o reconhecimento de um único gesto isolado. Uma das formas para fazer o reconhecimento de um conjunto de gestos realizados em sequência em um vídeo, por exemplo, consiste na conexão dos modelos isolados de forma a gerar uma rede de HMMs. Essa rede gerada consiste, na verdade, de um novo HMM.

A figura 2 representa uma rede chamada de rede de gestos, onde cada HMM modela um gesto, como os gestos correspondentes aos sinais barato, bonito e verdade. Cada HMM contém cinco estados emissores de símbolos (em branco) e dois estados não emissores (em cinza). Os estados não emissores funcionam como pontos de conexão entre os HMMs. Nessa rede os HMMs estão conectados em paralelo, o que significa que partindo do estado inicial até o final, somente um dos caminhos pode ser percorrido, indicando a ocorrência de apenas um gesto.

Já a figura 3 utiliza a rede da figura 2 como uma sub-rede de uma rede maior. Além dos modelos da rede de gestos, há a adição de um HMM de um estado emissor no início e no final, modelando a posição de repouso. Essa rede implica em um HMM capaz de reconhecer uma sequência 4 gestos quaisquer que estejam modelados na rede de gestos da figura 2.

III. Redes Neurais Convolucionais

Uma rede neural artificial é uma ferramenta matemática muito utilizada no reconhecimento de padrões. Consiste de

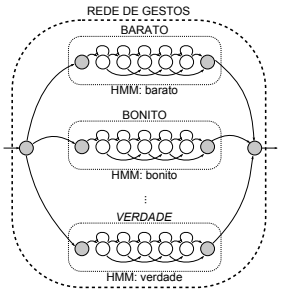

Fig. 2. Rede de gestos.

Fig. 3. Rede de HMMs para quatro gestos.

múltiplas unidades, chamadas de neurônios, organizadas em camadas. A saída dos neurônios de uma camada consiste da soma de suas entradas ponderadas por pesos que é então aplicada a uma função de ativação. A saída de cada neurônio é conectada às entradas dos neurônios da camada posterior. A aprendizagem consiste em obter o conjunto de pesos que, quando aplicada determinada entrada na rede, gerem a saída desejada.

As redes neurais convolucionais (CNN) [12] possuem uma arquitetura especialmente adaptada para o reconhecimento de imagens. Nas suas camadas, os neurônios estão distribuídos de forma tridimensional, onde cada fatia bidimensional é chamada de canal. Cada canal é composto por neurônios que compartilham os mesmos valores para seus pesos, e isso pode ser interpretado como se cada canal se especializa em reconhecer apenas um tipo tipo de padrão na imagem. Estas camadas são chamadas de camadas convolucionais. Após uma camada convolucional é comum existir uma camada de pooling, que são camadas que fazem uma subamostragem de forma a compactar a informação passada para a camada convolucional posterior.

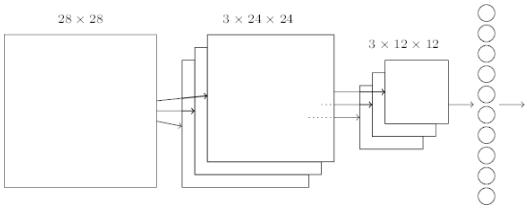

Fig. 4. Uma entrada conectada a uma CNN com uma camada convolucional, uma de pooling e uma camada totalmente conectada na saida.

IV. Metodologia

O sistema de reconhecimento de gestos desenvolvido foi dividido em três partes principais: rastreamento, extração dos descritores e classificação. Informações detalhadas sobre os algoritmos, descritores, treinamento e testes podem ser obtidas em [13].

A. *Rastreamento*

É na etapa de rastreamento que é feita a detecção, segmentação e rastreamento das mãos e da face em cada quadro do vídeo. Para facilitar a detecção são utilizadas luvas amarelas, e controla-se o fundo da imagem para que não contenha cores similares à cor da pele ou das luvas. O gesticulador também deve usar camiseta de manga comprida e cor diferente da cor da luva e da pele.

Primeiramente, os pixels da imagem dentro de uma região de interesse são classificados de acordo com a sua cor (crominâncias Cr e Cb) como pertencentes a uma das três classes possíveis: luva, pele ou fundo, onde a probabilidade da classe dada a cor observada é modelada por uma mistura de cinco gaussianas. Existe uma região de interesse para cada objeto que deve ser rastreado (chamado de objeto de interesse), ou seja, a região da mão esquerda, da mão direita e da face. Os pixels classificados dentro de uma região de interesse formam uma imagem binária, indicando se o pixel pertence ao fundo ou ao objeto que a respectiva região rastreia. Essa imagem é filtrada e, em seguida, todos os pixels conectados que não pertencem ao fundo são agrupados em objetos candidatos ao objeto da região de interesse. Dentre estes objetos, um deles é selecionado como objeto de interesse através um conjunto de regras que analisam sua área, posição, velocidade e comparam com limiares e valores anteriores.

Baseado na área do objeto de interesse, em sua posição, velocidade e direção de movimento, é calculada uma nova região de interesse para fazer a detecção no quadro posterior. Duas situações especiais podem ocorrer: uma quando as mãos se sobrepõem ou se ocluem e a outra quando a mão oclui a face. No primeiro caso as duas mãos são tratadas como um único objeto, já no segundo caso são preservadas as características da face detectadas nos quadros antes de ocorrer a oclusão. A figura 5 mostra um exemplo contendo as regiões de interesse e as mãos e face detectadas.

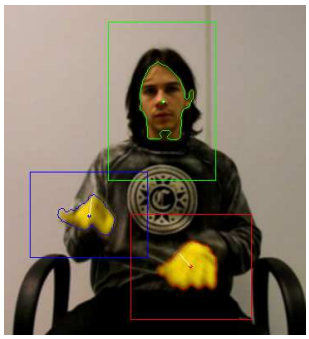

Fig. 5. Objetos detectados e sua região de interesse.

B. *Extração dos Descritores*

Esta parte é responsável por extrair, a cada quadro, os descritores das imagens segmentadas na etapa de rastreamento. Os descritores utilizados são divididos em três tipos:

- Descritores relativos ao movimento das mãos: distância em relação ao centro da face, ângulo em relação ao centro da face, direção do movimento, velocidade e aceleração.
- Descritores relativos a forma das mãos: alongamento, retangularidade, compactação, ângulo do maior eixo da mão, momentos centrais não escalados, momentos de HU [14], momentos invariantes a transformações afins, momentos do contorno de uma região, descritores de Fourier do contorno de uma região.
- Descritores gerados por uma CNN, relativos a forma da mão.

Todos estes descritores são testados em combinações, e é selecionado o conjunto com melhor desempenho.

C. *Utilizando a CNN para Gerar Descritores*

Para gerar o terceiro tipo de descritores, utilizou-se uma CNN que recebe em sua entrada a imagem da mão principal, em escala de cinza, e gera em sua saída a probabilidade daquela imagem ser cada um dos formatos que as mãos podem assumir. Como os gestos utilizados são sinais da Libras, e esta define um conjunto de formas que as mãos podem assumir na realização dos sinais (cerca de 61 segundo segundo [15]), selecionou-se um conjunto de formas das mãos necessárias para a realização dos gestos, além de novos formatos relativos a distorção causada pela rotação e oclusão das mãos, totalizando as 25 formas representadas na figura 6.

Fig. 6. Formas da mão utilizadas.

Foram testadas varias CNNs com estruturas inspiradas em CNNs conhecidas, como a LeNet-5, AlexNet, VGG-16 e GoogLeNet. Dentre as estruturas testadas, selecionou-se a que gerou o melhor desempenho no conjunto de validação. A estrutura selecionada foi inspirada na AlexNet e é representada na figura 7 e descrita em detalhes na tabela I. Esta possui 11 camadas, sendo que a primeira é a camada de entrada, que recebe uma versão em escala de cinza de 64x64 pixels da forma da mão que deve ser classificada. Todas as camadas utilizam a função de ativação RELU, com exceção da camada de saída, que utiliza softmax. As camadas de subamostragem utilizam max-pooling.

Os valores dos neurônios da camada de saída da CNN são utilizados como descritores relativos ao formato da mão.

D. *Classificação*

O classificador é composto por uma rede de HMMs, onde o algoritmo Token Passing é utilizado para fazer a classificação. Cada gesto foi modelado por um HMM com

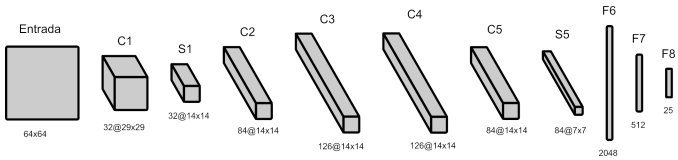

Fig. 7. CNN utilizada (C para camada convolucional; S para camada de subamostragem/max-pooling; F para camada totalmente conectada).

TABELA I
PARÂMETROS DA CNN.

	Canais	Tamanho	Filtro	Stride	Padding
Entrada	1	64x64	-	-	-
C1	32	29x29	8x8	2x2	-
S1	32	14x14	3x3	2x2	-
C2	84	14x14	5x5	1x1	2
C3	126	14x14	1x1	1x1	-
C4	126	14x14	3x3	1x1	1
C5	84	14x14	3x3	1x1	1
S5	84	7x7	2x2	2x2	-
F6	-	2048	-	-	-
F7	-	512	-	-	-
F8	-	25	-	-	-

estrutura esquerda-direita sem transições de escape, porém os HMMs dos gestos com movimentos repetitivos possuem uma transição de retorno para o estado inicial. A estrutura esquerda-direita foi escolhida porque os gestos utilizados se desenvolvem progressivamente no tempo, sempre avançando para um estado posterior. Além dos gestos, também foi criado um HMM de um único estado para modelar a posição de repouso das mãos e outro de três estados para modelar os movimentos de transição que ocorrem entre um gesto e outro, ambos a probabilidade de emissão de símbolos dos estados modelados por uma única distribuição gaussiana.

Para realizar o treinamento foram utilizados vários vídeos onde os gestos são realizados continuamente (em sequência e sem pausa), e para cada vídeo foi feita uma transcrição indicando o nome do sinal correspondente ao gesto e o seu tempo de início e fim no vídeo. Inicialmente é feita uma estimativa inicial para os HMMs através do alinhamento de Viterbi, e posteriormente é feita a re-estimação de Baum-Welch [10] [11]. Os modelos são treinados e validados em um conjunto de validação várias vezes, variando-se diversos parâmetros como: o número de gaussinas utilizadas para modelar a probabilidade de emissão dos estados; o número de estados; o conjunto de descritores utilizados. Essa variação visou selecionar o conjunto de parâmetros que gerou o melhor resultado.

V. CARACTERÍSTICAS DOS VÍDEOS

Foram utilizados conjuntos de vídeos compostos por uma sequência de quatro gestos realizados continuamente em cada vídeo, sendo que as mãos iniciam e terminam em uma posição de repouso. Utilizou-se quinze sinais selecionados aleatoriamente entre o vocabulário da Libras: barato, bonito, feliz, filho, pessoa (utilizam apenas uma mão), branco, carro, casa, coisa, fazer, hoje, muito, praia, preto e verdade (utilizam as duas mãos). Os vídeos possuem 640 pixels de largura por 480 de altura, adquiridos a 30 quadros por segundo.

Os vídeos foram gravados em dois períodos de tempos e locais distintos, com câmeras diferentes, obtendo-se assim variações na gesticulação, qualidade da imagem e iluminação.

VI. TESTES E RESULTADOS

Diferentes CNNs e HMMs, variando-se seus parâmetros (como a estrutura da CNN, o número de estados e descritores dos HMMs), foram treinados com um conjunto de treinamento e validados em um outro conjunto de validação. Foi selecionada a CNN e os parâmetros dos HMMs que geraram o melhor resultado no conjunto de validação. A estrutura selecionada para a CNN foi explanada na seção anterior, enquanto a estrutura selecionada para os HMMs é a seguinte: oito estados para os HMMs que modelam os gestos com movimentos repetitivos e onze para os demais; a probabilidade de emissão de símbolos dos estados foi modelada por uma única distribuição gaussiana; os HMMs são conectados, formando assim a rede de HMMs que compõem o classificador, apresentada na figura 8, onde REDE DE GESTOS corresponde a rede da figura 2. Esta rede indica que o classificador é projetado para reconhecer uma sequência indefinida de gestos com movimentos de transição entre eles, desde que o vídeo comece e termine com as mãos na posição de repouso.

Fig. 8. Rede de HMMs do classificador.

Após selecionada a estrutura da CNN e dos HMMs, estes foram novamente treinados utilizando tanto o conjunto de treinamento quanto o de validação. Os testes foram feitos em um novo conjunto de vídeos totalmente independente dos vídeos utilizados para o treinamento e validação, e foram divididos em duas partes: os testes relativos ao desempenho da CNN como classificador da forma da mão, e os testes relativos aos HMMs como classificador dos gestos nos vídeos.

A. Resultados da CNN

A CNN foi treinada com um conjunto de 9038 imagens das formas da mão (contendo um número variável de amostras para cada uma das 25 formas), extraídas dos vídeos dos conjuntos de treinamento e validação. Para os testes foram utilizadas 2434 formas da mão extraídas dos vídeos do conjunto de teste. A CNN obteve uma taxa de classificação correta de 96,59%, sendo que 31 dos 83 erros são devidos a semelhança entre duas formas muito parecidas, onde uma delas teve muitas amostras de treinamento e a outra poucas.

B. Resultados dos HMMs

Para o treinamento dos HMMs foram utilizados 107 vídeos, totalizando 432 amostras de gestos para o treinamento. Já os testes foram realizados em um conjunto de 24 vídeos, totalizando 96 amostras de gestos. Dentre as várias combinações de descritores utilizadas no processo de validação, selecionou-se

três conjuntos de descritores que geraram um bom desempenho na validação para serem utilizados nos testes:

- Conjunto básico: distância e ângulo das mãos em relação ao centro da face, direção do movimento, velocidade, aceleração, alongamento e compactação das mãos; ângulo do maior eixo de cada mão.
- Conjunto CNN: as 25 saídas da CNN, relativas ao formato da mão principal.
- Conjunto misto: mistura-se os dois conjuntos acima, excluindo-se os descritores de velocidade e aceleração devido a grande quantidade de ruído observada nestes.

A partir destes três conjuntos foram treinadas três redes de HMMs, sendo que o resultado dos testes é dado na tabela II. Os resultados são dados para cada um dos três conjuntos de descritores utilizados, onde ND representa o número de descritores; Gestos e Sequência representam a taxa de classificação correta dos gestos e das sequências de 4 gestos em cada vídeo.

TABELA II
RESULTADOS DOS CLASSIFICADORES.

Descritores	ND	Gestos	Sequência
Conjunto básico	16	85.42%	58.33%
Conjunto CNN	25	98.96%	95.83%
Conjunto misto	37	100%	100%

C. Tempo de Processamento

O funcionamento do sistema consiste em adquirir um vídeo com uma sequência de gestos e logo em seguida gerar a descrição. Para isso é feita a extração dos descritores quadro a quadro, e ao final do vídeo, todos os descritores extraídos estão a disposição da rede de HMMs para que seja realizado o reconhecimento. Foi utilizado um processador Intel Core i3-7100 e uma placa de vídeo NVIDIA GeForce GTX 1050 Ti.

O tempo de aquisição dos quadros dos vídeos é de 30fps $= 33.33$ms por quadro; o tempo de processamento e extração dos descritores (conjunto misto) de cada quadro é, em média, 5.59ms por quadro; já o tempo para o reconhecimento é de aproximadamente 531ms para cada sequência de gestos. Como pode ser visto, o tempo de processamento é consideravelmente menor que o tempo de aquisição, o que permite a execução em tempo real. Já o reconhecimento ocorre após o processamento do vídeo e em um intervalo de tempo muito pequeno.

VII. LIMITAÇÕES

Apesar do classificador ter obtido uma acurácia de 100% no conjunto de testes, é importante ressaltar que esse resultado não se generaliza para um gesticulador diferente do que foi utilizado nos vídeos de treinamento ou para diferentes condições de iluminação. Além disso não é permitido mais de uma pessoa no vídeo e objetos de cor semelhante a da pele ou das luvas, sendo necessário o uso de camiseta de manga comprida e luvas coloridas. Outro fator que deve ser levado em conta é o tamanho do vocabulário que contém apenas 15 gestos, além de cada vídeo conter uma sequência de apenas quatro gestos. Porém, como o conjunto de testes é independente do de treinamento, acreditamos que o sistema apresentará resultados similares para novos gestos realizados sob as mesmas condições.

VIII. CONCLUSÕES

Este trabalho utilizou um conjunto de descritores, extraídos a cada quadro, composto por descritores espaciais e geométricos relativos as mãos em conjunto com as saídas de uma CNN treinada para reconhecer as formas da mão. O classificador é composto por uma rede de HMMs responsável por identificar a sequência de gestos nos vídeos. O sistema apresentou um resultado promissor, reconhecendo sequências de gestos realizados continuamente e sem pausa, em tempo real. Apesar da acurácia de 100% obtida, não é esperado que esse resultado se generalize para condições diferentes das limitações apresentadas. Alguns pontos importantes para trabalhos futuros seria a utilização de um algoritmo de detecção e rastreamento das mãos que fosse capaz de eliminar a necessidade do uso de luvas e fosse menos sensível a variação da iluminação; utilizar um vocabulário maior e tornar o sistema multi-usuário.

REFERÊNCIAS

[1] V. Bobić e P. Tadić e G. Kvaščev, Hand gesture recognition using neural network based techniques. *2016 13th Symposium on Neural Networks and Applications (NEUREL)*, pp. 1–4, Novembro 2016.

[2] O. Oyedotun e A. Khashman, Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, Abril 2016.

[3] J. Nagi e F. Ducatelle, Max-pooling convolutional neural networks for vision-based hand gesture recognition. *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 342–347, Novembro 2011.

[4] A. Tang e K. Lu, A Real-Time Hand Posture Recognition System Using Deep Neural Networks. *ACM Transactions on Intelligent Systems and Technology*, v. 6, n. 2, pp. 1–23, Março 2015.

[5] L. Pigou e S. Dieleman, Sign Language Recognition Using Convolutional Neural Networks. *Computer Vision - ECCV 2014 Workshops*. Springer International Publishing, pp. 572–578, 2015.

[6] T. Starner e A. Pentland, Real-time American Sign Language recognition from video using hidden Markov models. *International Symposium on Computer Vision*, pp. 265–270, 1995.

[7] K. Grobel e M. Assan, Isolated sign language recognition using hidden Markov models. *Computational Cybernetics and Simulation 1997 IEEE International Conference on Systems, Man, and Cybernetics*, v. 1, pp. 162–167, Outubro 1997.

[8] R. Liang e M. Ouhyoung, A real-time continuous gesture recognition system for sign language. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 558–567, Abril 1998.

[9] M. Zaki e S. Shaheen, Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, v. 32, n. 4, pp. 572–577, 2011.

[10] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pp. 257–286, 1989.

[11] P. Dymarski, Hidden Markov Models, Theory And Applications. InTech, 2011.

[12] A. Nielsen, Neural networks and deep learning. Disponível em http://neuralnetworksanddeeplearning.com/index.html, 2015. Acessado em junho de 2017.

[13] V. Breda, Reconhecimento de gestos em vídeos utilizando modelos ocultos de Markov e redes neurais convolucionais aplicado a Libras. Dissertação (Mestrado em Engenharia Elétrica) - UFSC, Florianópolis. 2018. Disponível em http://tede.ufsc.br/teses/PEEL1838-D.pdf. Acessado em agosto de 2019.

[14] M. Hu, Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions*, v. 8, n. 2, pp. 179–187, 1962.

[15] F. Ramos, Libras. ULBRA, Canoas. 2016.