q1.01.04 - Matemática / Matemática Aplicada

UM ESTUDO COMPARATIVO DA TRANSFORMADA DE DISTÂNCIA E DA IMAGEM DE PROFUNDIDADE PARA RECONHECIMENTO DE GESTOS ESTÁTICOS DA MÃO USANDO REDES NEURAIS CONVOLUCIONAIS

Givanildo Lima[1], Lucas Amaral[2], Tiago Vieira[3], Thales Vieira [4]

1. Estudante de Engenharia da Computação do Instituto de Computação da UFAL
2. Estudante de Ciência da Computação do Instituto de Computação da UFAL
3. Pesquisador do Instituto de Computação da UFAL
4. Instituto de Matemática - UFAL / Orientador

Resumo:
Neste artigo propomos um estudo comparativo da acurácia de dois conjuntos de imagens para reconhecimento de gestos estáticos da mão. O primeiro é composto por imagens de profundidade capturadas a partir do sensor RealSense. Em seguida, a mão é segmentada do plano de fundo e depois aplica-se a Transformada de Distância a cada imagem, resultando nas imagens que compõem o segundo conjunto. Ambos são utilizados para treinar uma rede neural convolucional (CNN) focada na classificação de múltiplas poses da mão. Para avaliar este método em um contexto prático, coletamos uma base de dados contendo 28000 imagens representando 14 classes de configurações de mão distintas representando configurações da Língua Brasileira de Sinais (Libras).

Palavras-chave: Transformada de Distância; Redes Neurais Convolucionais; Gestos de Libras;

Apoio financeiro: CNPq, FAPEAL e UFAL.

Trabalho selecionado para a JNIC pela instituição: UFAL

Introdução:

Em nosso cotidiano, é cada vez mais comum nos depararmos com línguas de sinais. Seja nas ruas ou em comerciais de televisão, essas línguas, que permitem um meio de interação cinésico-visual entre seus usuários, têm se mostrado cada vez mais importantes para inclusão social. Porém, para ampliar essa inclusão de deficientes auditivos, faz-se necessário o desenvolvimento de novas tecnologias que facilitem a comunicação com indivíduos que não são fluentes nestas línguas.

Por outro lado, interfaces naturais de usuário (NUI) têm se tornado, aos poucos, uma realidade, permitindo a interação entre homem e máquina através de gestos do corpo humano, especialmente com o advento de sensores de profundidade como o Kinect [1] e o Real Sense [2].

Neste trabalho, iremos comparar as taxas de acurácia do reconhecimento de gestos estáticos da mão sendo feito utilizando dois conjuntos de imagens distintos, um deles contendo somente imagens de profundidade, capturadas utilizando o sensor RealSense e outro montado a partir das imagens de profundidade, as quais passaram por um etapa de segmentação, binarização e aplicação da Transformada de Distância. Ambos os conjuntos de imagem são usados para treinar e classificar uma Rede Neural Convolucional (CNN). Como aplicação, demonstramos a eficácia do método para reconhecimento de alguns sinais da Língua Brasileira de Sinais (Libras) caracterizados por poses (ou gestos estáticos) da mão. Libras é usada principalmente por deficientes auditivos no Brasil e é considerada a língua de sinais oficial brasileira desde 2002.

Metodologia:

A. Visão Geral

O método deste trabalho é baseado em aprendizagem de máquina usando redes neurais convolucionais, tendo como entradas os dois conjuntos de imagens de profundidade previamente mencionados. Ao realizar uma configuração da mão na frente do RealSense, o usuário terá a imagem de profundidade da mão capturada. Todas as imagens obtidas nesta etapa irão compor o primeiro conjunto. Na primeira etapa do método, cada umas destas imagens é segmentada e binarizada para possibilitar a extração da região da mão. Em seguida, a imagem resultante é esqueletonizada aplicando-se o operador de Transformada de Distância [5]. Ambos os conjuntos de imagens serão utilizados tanto para treinamento, quanto para reconhecimento usando uma rede neural convolucional.

B. Imagens de Profundidade

Ao capturar uma fotografia, uma câmera tradicional armazena informações referentes a cor em cada pixel da imagem produzida. Por outro lado, sensores de profundidade também são capazes de obter dados relativos à distância do objeto ao sensor, ou profundidade. Desse modo, uma imagem de profundidade pode

ser representada por uma função $f : U \subset R^2 \rightarrow R$, onde z = f (x, y) representa a distância do sensor até o objeto visível na direção do pixel (x, y). Em posse dessas informações fornecidas por um sensor de profundidade, o objetivo é poder treinar e classificar poses da mão e, em particular, gestos estáticos da Língua Brasileira de Sinais (LIBRAS).

C. Binarização

Nesta etapa, o foco é segmentar e binarizar a mão da imagem de profundidade, considerando que a mão sempre será o objeto mais próximo ao sensor. Sendo f uma imagem de profundidade capturada pelo sensor, o objetivo é criar a imagem b, que representa a binarização de f. A imagem b é dada pela seguinte função:

$$b(x,y) = \begin{cases} 1, & 1 \leq f(x,y) \leq D_{\min} + T \\ 0, & f(x.y) > D_{\min} + T \end{cases}$$

onde $D_{\min} = \min_{U}(f(x,y))$ e T é um limiar de profundidade usado para extrair apenas a região da mão, como ilustra a Figura 1.

D. Transformada de Distância

A partir da imagem binarizada b obtida na etapa de segmentação e binarização da imagem, o próximo passo é aplicar a transformada de distância [5] com a intenção de esqueletonizar a imagem, possibilitando assim o uso de uma representação mais concisa e discriminativa na camada de entrada da rede neural convolucional.

A Transformada de Distância de uma imagem é, basicamente, uma outra imagem calculada de forma que cada pixel interior a uma região, a qual representa um objeto da imagem, tenha seu valor dado pela distância do pixel à borda da região. Mais especificamente, vamos considerar a região:

$$V = \{(x,y) \in \mathbb{U} \mid b(x,y) = 1\}.$$

A imagem da Transformada de Distância é representada por:

$$h(x,y) = \min_{(a,b) \in \partial V} ||(x,y) - (a,b)||$$

A Figura 1 exibe um exemplo completo, desde a imagem de profundidade proveniente do sensor, até sua Transformada de distância. Para mais detalhes, ver [5].

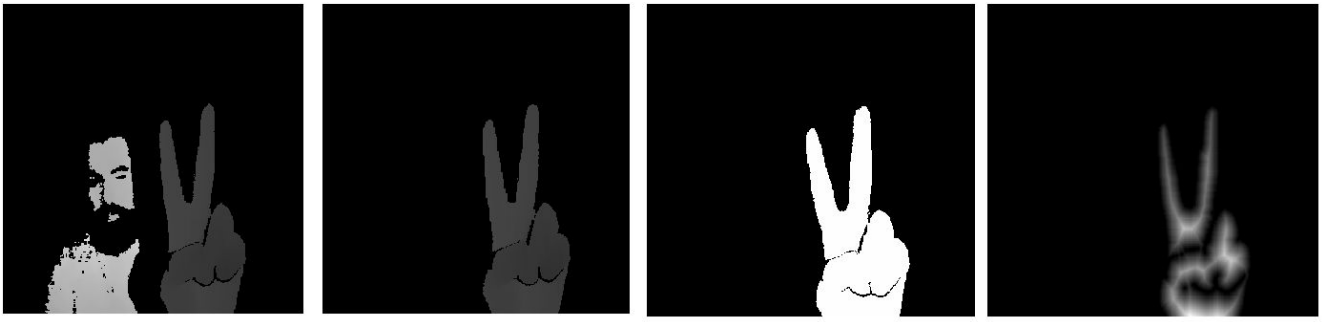

Figura 1. Visão geral do método: Geração da imagem usada como entrada para a rede neural convolucional.

E. Rede Neural Convolucional

a) Treinamento e Classificação:

Na etapa de treinamento, diversos exemplos de classes de poses de mão, representados pelas suas respectivas transformadas de distâncias, são dados de entrada para treinar uma rede neural convolucional, usando o método de propagação reversa (*back-propagation*). Adicionalmente, aumentamos a base de dados aplicando aleatoriamente cisalhamentos e aproximações nas imagens originais. Na etapa de classificação, transformadas de distâncias também são dadas de entrada para a rede neural convolucional, cuja saída corresponde a um vetor contendo as probabilidades de que um dado exemplo de entrada seja de cada classe treinada.

b) Arquitetura:

Nossa rede neural convolucional é composta das seguintes camadas na ordem que segue:

Convolucional, Agrupamento Máximo, Convolucional, Agrupamento Máximo, Achatamento, além de duas camadas densas (MLP), como exibido na Figura 2.

A rede tem como entrada uma imagem de distância de dimensões 50x50 em tons de cinza. A primeira camada (Convolucional) aplica um núcleo de convolução 3x3, e tem como saída uma imagem com as mesmas dimensões da de entrada. Em seguida, a imagem é escalonada para uma imagem com resolução 25x25 pela camada de Agrupamento Máximo que possui um núcleo 2x2. A camada de achatamento concatena os pixels da imagem de saída da camada anterior transformando-a em um único vetor de dimensão 625 e tem como saída 100 neurônios.

A primeira camada densa recebe como entrada a saída da camada de achatamento e tem como ativação uma função linear de retificação (ReLu). A segunda camada densa está ligada a 40 neurônios e tem a SoftMax como sua função de ativação, que retorna justamente o vetor de probabilidades.

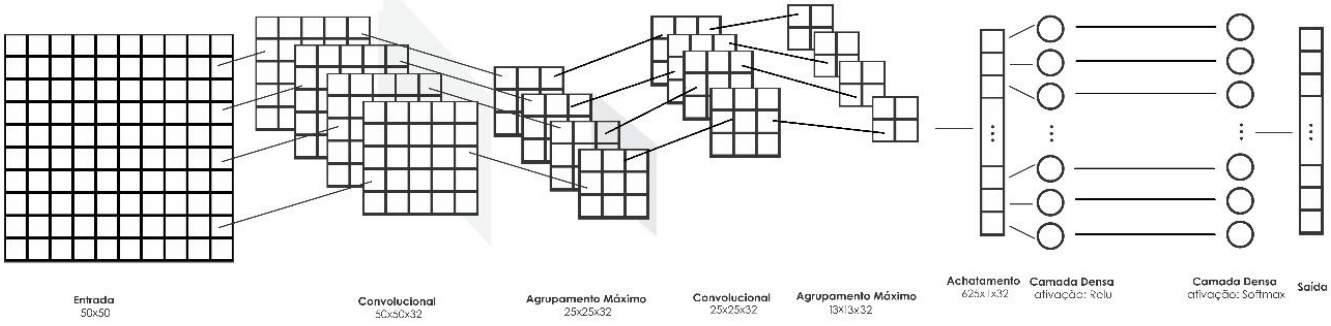

Figura 2. Arquitetura da rede neural convolucional: Uma imagem monocromática com resolução de 50x50 pixels é dada como entrada, e um vetor de probabilidades com dimensão 14 é retornado no final

Resultados e Discussão:

Para validar o método, foi utilizada a Língua Brasileira de Sinais (Libras) como aplicação. Com esse objetivo, foi coletada uma base de dados de imagens de profundidade de Libras que será descrita a seguir. A implementação do método foi realizada usando a linguagem C++, OpenCV e a SDK do sensor RealSense para a etapa de aquisição e pré-processamento das imagens; e a linguagem Python com as bibliotecas Keras e Tensorflow para treinamento e classificação com CNN.

A. Base de Dados:

Inicialmente, a fim de produzir um banco de imagens, foram selecionadas 14 configurações (poses) de mão que representam sinais de Libras. Para testar a eficiência do método algumas das configurações selecionadas são parecidas, tornando o trabalho do classificador mais desafiador.

Para produzir o banco de imagens foi utilizado um sensor de profundidade RealSense [2]. Foram capturados 1000 imagens de profundidade de cada configuração, com resolução de 640x480 pixels, totalizando 14000 exemplos que compõem o primeiro conjunto. Os exemplos foram capturados de modo a representar variações da mesma configuração de mão com pequenas translações e rotações. Em seguida, para produzir o segundo conjunto, cada uma dessas imagens passou pelo processo de binarização e aplicação da Transformada de Distância. Ao final do processo, o banco possui 28000 imagens divididas em 2 conjuntos de 14000 imagens cada.

B. Comparação dos Resultados:

Para validar o método, foram selecionados aleatoriamente 90% dos exemplos de cada classe para treinamento, e os 10% restantes para teste. Para o treinamento, o algoritmo de propagação reversa convergiu com cerca de 15 épocas de treino (com 40 lotes de amostra por época). As matrizes de confusão com os resultados são exibidas na Tabela I e II. Em média, 97.64% dos exemplos das imagens de profundidade (conjunto 1) foram corretamente classificados, enquanto que essa taxa foi de 99.14% para imagens nas quais foi aplicada a transformada de distância (conjunto 2). Cada elemento a_{ij} representa a quantidade de exemplos da classe i que foram classificados como pertencentes à classe j.

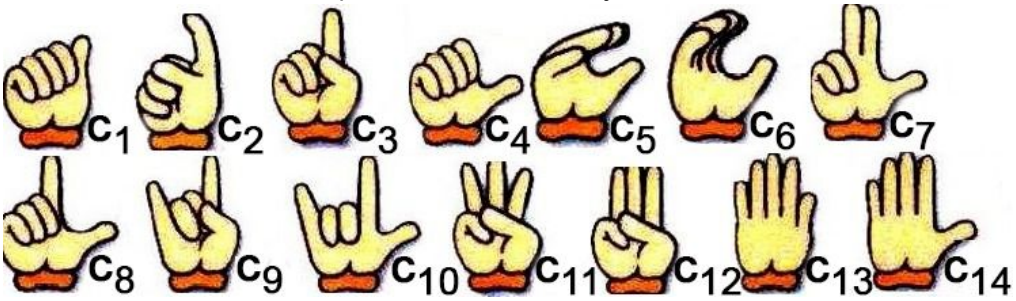

Figura 3. Ilustração com as configurações de mão selecionadas para o treinamento

Tabela I
MATRIZ DE CONFUSÃO DAS IMAGENS DE PROFUNDIDADE

classe	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}
c_1	98	2	0	0	0	0	0	0	0	0	0	0	0	0
c_2	0	97	0	0	0	0	0	1	1	0	0	0	0	1
c_3	0	0	100	0	0	0	0	0	0	0	0	0	0	0
c_4	0	1	0	99	0	0	0	0	0	0	0	0	0	0
c_5	0	0	0	0	99	0	0	0	0	0	0	0	0	1
c_6	0	0	0	0	1	99	0	0	0	0	0	0	0	0
c_7	0	0	0	0	0	0	95	1	2	0	0	0	0	2
c_8	0	0	0	0	0	0	0	99	0	0	1	0	0	0
c_9	0	1	3	0	1	0	0	0	92	1	0	1	1	0
c_{10}	0	0	0	0	0	0	0	1	0	98	0	0	0	1
c_{11}	0	0	0	0	0	0	0	0	0	0	100	0	0	0
c_{12}	1	1	0	0	0	0	0	0	0	0	0	95	1	2
c_{13}	0	3	0	0	0	0	0	1	0	0	0	0	96	0
c_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Tabela II
MATRIZ DE CONFUSÃO DAS IMAGENS DA TRANSFORMADA DE DISTÂNCIA

classe	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}
c_1	99	0	0	1	0	0	0	0	0	0	0	0	0	0
c_2	0	100	0	0	0	0	0	0	0	0	0	0	0	0
c_3	0	1	99	0	0	0	0	0	0	0	0	0	0	0
c_4	0	0	0	99	0	0	0	0	1	0	0	0	0	0
c_5	0	0	0	0	98	0	0	0	2	0	0	0	0	0
c_6	0	0	0	0	0	100	0	0	0	0	0	0	0	0
c_7	0	0	0	0	0	0	100	0	0	0	0	0	0	0
c_8	0	0	0	0	0	0	0	100	0	0	0	0	0	0
c_9	0	0	2	0	0	0	0	0	98	0	0	0	0	0
c_{10}	0	0	0	0	0	0	0	0	0	100	0	0	0	0
c_{11}	0	0	0	0	0	0	0	0	0	0	100	0	0	0
c_{12}	0	0	0	0	0	0	0	0	0	0	0	99	1	0
c_{13}	0	0	0	0	0	0	0	0	1	0	0	0	99	0
c_{14}	0	0	0	0	0	0	0	0	1	0	0	0	2	97

A partir dos resultados apresentados nas tabelas acima, podemos concluir que o método de reconhecimento é mais eficiente para as imagens do conjunto 2, as quais foi aplicada a Transformada de Distância, sendo mais eficiente no reconhecimento de 9 dos 14 gestos e com a mesma eficiência para outros 2 gestos.

Conclusões:

Este trabalho apresentou um estudo comparativo para realizar o reconhecimento de gestos estáticos da Língua Brasileira de Sinais (Libras). Nosso método foi baseado nas seguintes etapas: aquisição de imagens; segmentação e binarização; cálculo da Transformada de Distância; treinamento e classificação de redes neurais convolucionais; comparação dos resultados obtidos. A partir dos resultados obtidos, pode-se concluir que, para a rede adotada, a acurácia do conjunto de imagens que passou por um tratamento é maior que a do conjunto de imagens de profundidade, embora os resultados sejam semelhantes.

Referências bibliográficas:

[1] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in CVPR, 2011, pp. 1297–1304.

[2] Intel, “Intel realsense technology,” 2017. [Online]. Available: https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

[3] R. Faugeroux, T. Vieira, D. Martinez, and T. Lewiner, “Simplified training for gesture recognition,” in Sibgrapi, 2014, pp. 133–140.

[4] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. M. Campos, “Real-time gesture recognition from depth data through key poses learning and decision forests,” in Sibgrapi, 2012, pp. 268–275.

[5] A. Peixoto and L. C. Velho, Transformadas de distância. PUC, 2000.