Reconhecimento de gestos estáticos da mão usando a Transformada de Distância e aplicações em Libras

Lucas Amaral, Givanildo Lima, Tiago Vieira
Instituto de Computação
Universidade Federal de Alagoas

Thales Vieira
Instituto de Matemática
Universidade Federal de Alagoas

***Resumo*—Neste artigo propomos um método para reconhecer gestos estáticos da mão representados por imagens de profundidade. Inicialmente, nós segmentamos a mão do plano de fundo e depois calculamos a Transformada de Distância para treinar uma rede neural convolucional (CNN) que é usada para classificar poses da mão. Para avaliar nosso método em um contexto prático, coletamos uma base de dados contendo 1400 imagens representando 14 classes de configurações de mão distintas representando sinais da Língua Brasileira de Sinais (Libras). Nosso método alcançou uma taxa de reconhecimento de 96.42% na média.**

***Abstract*—In this paper we propose a method to recognize static hand gestures from depth images. We first segment the hand from the background, and then compute the Distance Transform to train a Convolutional Neural Network (CNN) that is later used to classify hand poses. In order to evaluate our method in a practical context, we collected a dataset containing 1400 images representing 14 different hand configurations representing signs of the Brazilian Sign Language (Libras). Our method achieved an average recognition rate of 96.42%.**

***Keywords*—Transformada de Distância; Redes Neurais Convolucionais; Gestos de Libras;**

I. Introdução

Em nosso cotidiano, é cada vez mais comum nos depararmos com línguas de sinais. Essas línguas, que permitem um meio de interação cinésico-visual entre seus usuários, têm se mostrado cada vez mais importantes para inclusão social. Isso vem possibilitando uma melhor qualidade de vida para aqueles que possuem as línguas de sinais como sua única forma de comunicação. Porém, para ampliar a inclusão social 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, apresentamos um método para reconhecer gestos estáticos da mão que pode ser usado para desenvolver interfaces naturais de usuário e, em particular, aplicações voltadas à língua brasileira de sinais (LIBRAS), usando imagens obtidas a partir do sensor de profundidade RealSense. LIBRAS é usada principalmente por deficientes auditivos no Brasil e é considerada língua oficial brasileira desde 2002. Apresentamos, em especial, a Transformada de Distância da imagem de profundidade como representação usada 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. Destacamos que este trabalho pode ser um ponto de partida para o desenvolvimento de métodos mais sofisticados que permitam o reconhecimento de gestos dinâmicos, inclusive de Libras.

II. Trabalhos Relacionados

Nos últimos anos, observou-se um crescente surgimento de trabalhos relacionados ao reconhecimento de gestos, sendo alguns focados na criação de interfaces naturais de usuário [3]–[6], enquanto outros, como o presente trabalho, têm como aplicação principal o reconhecimento de línguas de sinais, onde destacamos os trabalhos propostos por Rahman e Afrin [7], Uebersax et al. [8] e Fábio Domínio [9] que reconhecem sinais de língua americana de sinais (ASL); além dos trabalhos de Bowden et al. [10] e Liwicki e Everingham [11], que lidam com o reconhecimento de gestos da língua britânica de sinais (BSL).

O reconhecimento de sinais direcionado a Libras também é abordado em algumas obras como aquela proposta por Anjo, Pizzolato e Pedroso [12]. Eles obtiveram uma acurácia de 90,7% no reconhecimento de 27 sinais diferentes de libras, sendo 8 gestos dinâmicos. Outros trabalhos também merecem destaque, como o reconhecimento utilizando descritores de forma em uma abordagem feita por Bastos, Angelo e Loula [13], além do trabalho de Escobedo e Câmara [14], no qual são utilizados cossenos de direção e um histograma de magnitudes cumulativas a fim de realizar o reconhecimento;

Com relação aos métodos de aprendizagem baseados em redes neurais convolucionais, técnica também usada no presente trabalho, destacamos o método desenvolvido por Lopes, Aguiar e Santos [15] para reconhecer expressões faciais. Nossa arquitetura apresenta algumas semelhanças com esse método.

III. Metodologia

A. Visão Geral

Nosso método é baseado em aprendizagem de máquina usando redes neurais convolucionais, tendo como entrada imagens de profundidade extraídas do sensor RealSense. Ao realizar uma configuração da mão na frente deste sensor,

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

o usuário terá sua imagem de profundidade capturada. Esta imagem será, na primeira etapa do método, segmentada para extração da região da mão. Em seguida, a imagem resultante será esqueletonizada aplicando-se o operador de Transformada de Distância [16]. A imagem resultante será usada 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 $f : U \subset \mathbb{R}^2 \to \mathbb{R}$, onde $z = f(x, y)$ representa a distância do sensor até o objeto visível na direção do pixel (x, y). Dotados dessas informações fornecidas por um sensor de profundidade, nosso objetivo é poder treinar e classificar poses da mão e, em particular, gestos estáticos da Língua Brasileira de Sinais (Libras).

C. Segmentação e binarização

Nessa etapa, o foco é segmentar 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, nosso objetivo é criar a imagem f' dada pela seguinte função:

$$f'(x,y) = \begin{cases} f(x,y), \ 1 \leq f(x,y) \leq D_{\min} + T \\ 0, \ f(x.y) > D_{\min} + T \end{cases}$$

onde $D_{\min} = \min_{\mathbb{U}}(f(x,y))$ e T é um limiar de profundidade usado para extrair apenas a região da mão, como ilustra a Figura 1. Em seguida, antecedendo o cálculo da Transformada de Distância, construímos uma imagem binarizada

$$b(x,y) = \begin{cases} 1, \ f'(x,y) > 0 \\ 0, \ f'(x,y) \leq 0 \end{cases}$$

que representa a binarização de f'. Note que estas duas operações podem ser realizadas em um único passo.

D. Transformada de Distância

A partir da imagem binarizada b obtida na etapa de segmentação da imagem, temos como próximo passo a aplicação da transformada de distância [16] 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 [16].

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

A rede tem como entrada uma imagem de distância de dimensões 50×50 em tons de cinza. A primeira camada

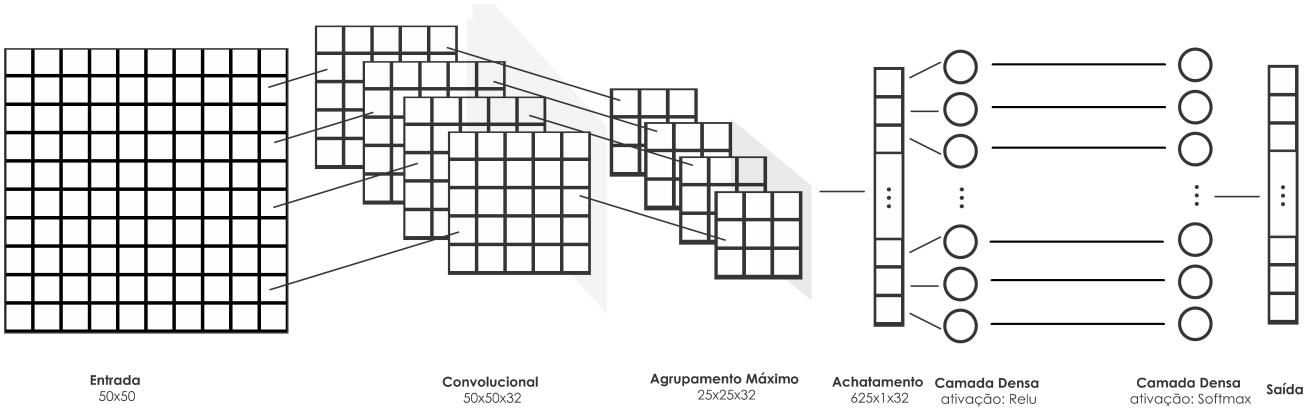

Fig. 2. Arquitetura da rede neural convolucional: uma imagem monocromática com resolução de 50×50 pixels é dada como entrada e um vetor de probabilidades de dimensão 14 é retornado no final.

(Convolucional) aplica um núcleo de convolução 3×3, e tem como saída uma imagem com as mesmas dimensões da de entrada. Em seguida, a imagem é escalonada para uma imagem com resolução 25×25 pela camada de Agrupamento Máximo que possui um núcleo 2×2. 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.

IV. Experimentos

Para validar nosso método, usamos a Língua Brasileira de Sinais (Libras) como aplicação. Com esse objetivo, coletamos uma base de dados de imagens de profundidade de Libras que descrevemos 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[1] e Tensorflow [2] 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 de nosso método, algumas das configurações selecionadas são parecidas, tornando o trabalho do classificador mais desafiador.

Para produzir nosso banco de imagens, foi utilizado um sensor de profundidade RealSense [2]. Foram capturados 100 exemplos de cada configuração, onde cada exemplo é composto por imagens de profundidade e imagens coloridas (rgb), com resolução de 640×480 pixels, totalizando 1400 exemplos. Os exemplos foram capturados de modo a representar variações da mesma configuração de mão com pequenas translações e rotações. Apesar de fazer parte da base de dados, as imagens rgb não são utilizadas em nosso método.

[1]https://keras.io/

[2]http://tensorflow.org/

B. Resultados

Para validar nosso método, selecionamos aleatoriamente 90% dos exemplos de cada classe para treinamento, e os 10% restantes para teste. Para o treinamento, o algoritmo de propagação reversa convergiu rapidamente com apenas 4 épocas de treino (com 40 lotes de amostra por época), sendo necessário menos de 1 minuto para completar o treinamento da rede.

A matriz de confusão exibindo os resultados é exibida na Tabela I. Em média, 96.42% dos exemplos foram corretamente classificados. O único resultado abaixo de 90% foi observado na classe c_4, que corresponde a uma configuração de mão que de fato é muito parecida com a classe c_1, como podemos ver nas Figuras 3 e 4, dificultando o trabalho do classificador.

V. Conclusão

Este trabalho apresentou um método 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.

Tabela I
Matriz de confusão: cada elemento a_{ij} representa a quantidade de exemplos da classe i que foram classificados como pertencentes à classe j.

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	10	0	0	0	0	0	0	0	0	0	0	0	0	0
c_2	0	10	0	0	0	0	0	0	0	0	0	0	0	0
c_3	0	0	10	0	0	0	0	0	0	0	0	0	0	0
c_4	3	0	0	7	0	0	0	0	0	0	0	0	0	0
c_5	0	0	0	0	10	0	0	0	0	0	0	0	0	0
c_6	0	0	0	0	0	10	0	0	0	0	0	0	0	0
c_7	0	0	0	0	0	0	10	0	0	0	0	0	0	0
c_8	0	0	0	0	0	0	0	10	0	0	0	0	0	0
c_9	1	0	0	0	0	0	0	0	9	0	0	0	0	0
c_{10}	0	0	0	0	0	0	0	0	1	9	0	0	0	0
c_{11}	0	0	0	0	0	0	0	0	0	0	10	0	0	0
c_{12}	0	0	0	0	0	0	0	0	0	0	0	10	0	0
c_{13}	0	0	0	0	0	0	0	0	0	0	0	0	10	0
c_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Fig. 3. Transformadas de distâncias de exemplos de cada classe de nossa base de dados. Cada classe c_i é representada na i-ésima linha com 3 exemplos, de acordo com a ordenação da Tabela I.

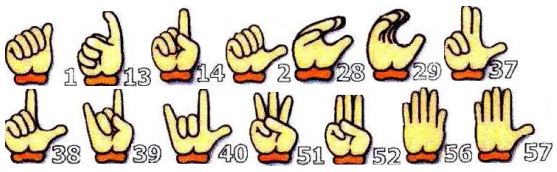

Fig. 4. Ilustração com as configurações de mão selecionadas para o treinamento. As classes seguem a mesma ordenação da Tabela I e da Figura 3.

Os ótimos resultados alcançados nos motivam a prosseguir em duas direções: realizar experimentos mais avançados, tendo como objetivo treinar redes capazes de reconhecer um vocabulário mais extenso de Libras; e desenvolver métodos para reconhecimento de gestos dinâmicos, tendo como base o método proposto.

AGRADECIMENTOS

Os autores gostariam de agradecer ao CNPq, à FAPEAL e à UFAL pelo auxílio financeiro e a estrutura física disponível para a realização deste trabalho.

REFERŁNCIAS

[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] T. Vieira, R. Faugeroux, D. Martínez, and T. Lewiner, "Online human moves recognition through discriminative key poses and speed-aware action graphs," *Machine Vision and Applications*, vol. 28, no. 1, pp. 185–200, Feb 2017. [Online]. Available: https://doi.org/10.1007/s00138-016-0818-y

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

[5] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, A. W. Vieira, and M. F. M. Campos, "Online gesture recognition from pose kernel learning and decision forests," *Pattern Recognition Letters*, vol. 39, pp. 65–73, 2014.

[6] L. Miranda, T. Vieira, D. Martínez, 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.

[7] M. H. Rahman and J. Afrin, "Hand gesture recognition using multiclass support vector machine," *International Journal of Computer Applications*, vol. 74, no. 1, 2013.

[8] D. Uebersax, J. Gall, M. Van den Bergh, and L. Van Gool, "Real-time sign language letter and word recognition from depth data," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 383–390.

[9] F. Dominio, M. Donadeo, and P. Zanuttigh, "Combining multiple depth-based descriptors for hand gesture recognition," *Pattern Recognition Letters*, vol. 50, pp. 101–111, 2014.

[10] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," *Computer Vision-ECCV 2004*, pp. 390–401, 2004.

[11] S. Liwicki and M. Everingham, "Automatic recognition of fingerspelled words in british sign language," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 2009, pp. 50–57.

[12] E. B. Pizzolato, M. dos Santos Anjo, and G. C. Pedroso, "Automatic recognition of finger spelling for libras based on a two-layer architecture," in *Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, 2010, pp. 969–973.

[13] I. L. Bastos, M. F. Angelo, and A. C. Loula, "Recognition of static gestures applied to brazilian sign language (libras)," in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*. IEEE, 2015, pp. 305–312.

[14] E. J. E. Cardenas and G. C. Chávez, "Finger spelling recognition from depth data using direction cosines and histogram of cumulative magnitudes," in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*. IEEE, 2015, pp. 173–179.

[15] A. T. Lopes, E. de Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*. IEEE, 2015, pp. 273–280.

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