Um sistema de baixo custo para reconhecimento de gestos em LIBRAS utilizando visão computacional

Carlos Henrique de A. Monteiro, Luiz Felipe Inácio Pecoraro, Angélica Takamine Lacerda, Anna Regina Corbo, Gabriel Matos Araujo

***Resumo*— Este artigo apresenta um arcabouço para reconhecimento de palavras em LIBRAS (Língua Brasileira de Sinais) em sequências de vídeo. A extração de características é baseada em subamostragem de imagens residuais, obtidas pela subtração de quadros sucessivos e a classificação é feita com o auxílio do algoritmo k-NN. Outra contribuição deste trabalho é uma base de dados (em expansão) contendo 24 palavras em LIBRAS executadas por diversos voluntários com 3 tipos de planos de fundo. Todo o sistema de reconhecimento foi desenvolvido no ambiente Matlab. O sistema proposto neste trabalho possui uma taxa de acerto média de 75%, o que indica que as técnicas descritas aqui são bastante promissoras.**

***Palavras-Chave*— LIBRAS, linguagem de sinais, visão computacional, processamento de imagens.**

***Abstract*— This paper present a framework to recognize words in LIBRAS (Brazilian Sign Language) in video sequences. The features extraction is based on sub-sampling residual images (obtained by subtracting consecutive frames) and the classification is performed by a k-NN algorithm. Another contribution of this work is a database (in expansion) containing 12 words in LIBRAS performed by several volunteers and 3 types of background. The system was developed in Matlab. The proposed system has a hit rate of 75% which indicates that the techniques described here are quite promising.**

***Keywords*— Sign Language, Computer Vision, Pattern Recognition, Image Processing.**

I. Introdução

É possível encontrar na literatura diversos trabalhos envolvendo algum aspecto do problema de reconhecimento de gestos [1], [2], [3], [4], [5]. É possível dividir estes trabalhos em dois grupos, os que envolvem os uso de sensores acoplados ao indivíduo e os que usam visão computacional [1]. Os métodos que envolvem uso de sensores acoplados ao corpo facilitam a digitalização do movimento e tendem a produzir resultados bem confiáveis. No entanto, os sensores empregados são muito caros e geram desconforto ao usuário [1]. Os métodos baseados em visão, por outro lado, somente necessitam de uma câmera. A desvantagem é que este tipo de sistema é muito sensível à variação de iluminação/iluminação não uniforme, mudança do indivíduo e da câmera [1].

Reconhecimento de linguagem de sinais é um caso particular de reconhecimento de gestos. Existem diversos tipos de linguagens de sinais diferentes: A Linguagem de Sinais Americana (ASL) para os Estados Unidos, (BSL) para os Britânicos, (SPL) para os espanhóis, além de linguagens para os Japoneses, Coreanos, Mexicanos, etc [6]. Os Brasileiros também possuem a sua própria Linguagem Brasileira de Sinais (LIBRAS). O desafio encontrado para produzir um algoritmo computacional capaz de interpretar gestos em linguagem de sinais não se restringe em mapear a trajetória realizada pelas mãos. É necessário elaborar sistemas que não só diferenciem as informações presentes em pontos de alta movimentação da face, mas que também classifiquem os sinais considerando os sotaques presentes na cultura surda.

Também é possível encontrar na literatura muitos trabalhos relacionados com reconhecimento de linguagem de sinais [6], [7], [8], [9], [10]. Em [6], o reconhecimento das palavras em ASL é feito através do casamento entre os quadros do vídeo de entrada e um dicionário. O melhor casamento, obtido por *Conditional Template Matching*, é usado para estabelecer a frase escrita em linguagem natural. Em [7] é possível encontrar um estudo sobre medidas de quantidade de movimento do falante de linguagem de sinais. As medidas de movimento são usadas para detectar a postura característica de cada falante e assim determinar as palavras. Em [8], há um método de baixo custo que utiliza *Hu-moments* para reconhecer palavras em linguagem de sinais e gerar uma transcrição para texto. Uma revisão das principais técnicas de reconhecimento em linguagem de sinais Árabes pode ser encontrado em [9]. Já em [10], o sistema de reconhecimento de linguagem de sinais é baseado em uma combinação entre k - *Nearest Neighbors (K-NN)* e *Hidden Markov Models (HMM)*.

Existem diversos trabalhos sobre reconhecimento de palavras em linguagem de sinais de outras línguas, mas é muito difícil encontrar algum aprofundado sobre reconhecimento de LIBRAS. Por outro lado, pesquisas deste tipo são fundamentais, principalmente por seu caráter inclusivo. Provavelmente uma das principais razões disso é a falta de uma base de dados grande o suficiente para treinar algoritmos de reconhecimento. O Governo do Brasil, por meio do Instituto Nacional de Educação de Surdos (INES), mantém um dicionário *online* contendo 3853 sinais/itens lexicais [11]. Contudo, o dicionário do INES possui somente uma sequência de vídeo para cada verbete, executado por somente uma pessoa em ambiente controlado e plano de fundo branco. Uma das contribuições deste trabalho é a criação de uma base de dados contendo mais amostras de cada verbete. A base de dados proposta, em expansão, atualmente conta com 24 palavras em LIBRAS, executadas por diversos voluntários, em 3 tipos de planos de fundo.

Além do banco de dados, este trabalho também propõe um método simples para reconhecer gestos em LIBRAS. Este método é composto por três etapas: geração de uma

Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), campus de Nova Iguaçu, Rio de Janeiro, Brasil. E-mails: {focus.carloshenriquemonteiro, luizfelipeinacio, angelicatakamine}@gmail.com {anna.costa, gabriel.araujo}@cefet-rj.br.

sequência de resíduo, subtraindo quadros adjacentes; Obtenção da matriz de características através da acumulação do movimento em células acumuladoras; Classificação usando k - *Nearest Neighbors (k-NN)*. Vale ressaltar que, embora a base de dados proposta contenha 24 verbetes, somente 12 estavam disponíveis no momento da condução dos experimentos

O restante deste trabalho está organizado da seguinte forma. Na Seção II a base de dados e o método proposto são descritos. Os resultados são descritos e apresentados na Seção III. A conclusão e algumas perspectivas futuras estão na Seção IV

II. Metodologia

A base de dados utilizada e o método proposto são descritos detalhadamente nesta Seção.

A. Base de dados

O Instituto Nacional de Educação de Surdos (INES), mantém um dicionário *online* contendo 3853 sinais/itens lexicais [11]. Cada um destes verbetes é apresentado em uma sequência de vídeo com resolução 240×180 *pixels*, executado por uma única mulher em ambiente controlado de fundo branco. Esse dicionário tem por finalidade ensinar LIBRAS. Para tarefas de reconhecimento automático de palavras em LIBRAS, é necessária uma base de dados contendo mais amostras de cada palavra, em diversos cenários. Por este motivo, este trabalho propõe uma nova base de dados mais adequada para treinar e testar sistemas de reconhecimento de palavras em LIBRAS.

A base de dados proposta está em expansão, mas atualmente conta 548 sequências de vídeo, distribuídas em 24 verbetes que podem ser subdivididos em dois grupos: um cujo o deslocamento das mãos segue uma trajetória horizontal, sendo elas *amante, bebê, compromisso, dado, elástico, impossível, milagre, notícia, orquestra, surfe, televisão e veloz*, e outra composta de palavras de deslocamento vertical, *andaime, depressão, fantasma, martelo, nacionalidade, neve, paciência, parede, relâmpago, trampolim e vapor*. Cada amostra é composta por uma sequências de vídeo de resolução 640×480 *pixels* e uma taxa de quadros de 30 *Hz*. As gravações foram feitas em ambiente com três tipos de plano de fundo estáticos: simples, intermediário e complexo. Elas foram executadas por 20 voluntários não surdos, sendo 13 mulheres e 7 homens. Convém observar que nem todos os voluntários executaram todas as palavras em todos os tipos de plano de fundo. A execução de palavras por uma pessoa não fluente em LIBRAS pode resultar em um banco não confiável. Para garantir uma confiabilidade mínima, a base de dados do INES foi utilizada como referência durante as gravações. A Figura 1 ilustra algumas sequências da base de dados proposta.

B. Sistema de reconhecimento de palavras em LIBRAS

O sistema de reconhecimento de LIBRAS proposto neste trabalho está ilustrado no diagrama da Figura 2. No primeiro bloco, intitulado "pré-processamento", as entradas são as sequências de vídeo contendo palavras e as saídas são as sequências de resíduo, obtidas a partir da subtração de quadros adjacentes. A segunda etapa é, chamada de "extração de características", consiste na obtenção da matriz de características através da acumulação do movimento em células acumuladoras; A etapa de "Classificação" é executada por um classificador baseado em k-*NN*.

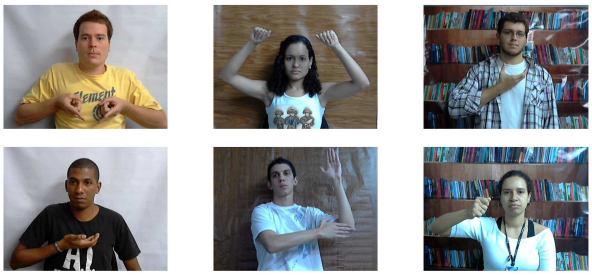

Fig. 1. Exemplos da base de dados proposta. Amostras com plano de fundo simples estão na coluna da esquerda. Planos de fundo intermediários na coluna central e complexos na coluna da direita.

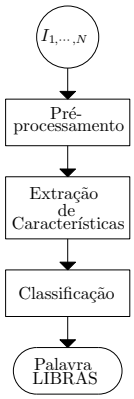

Fig. 2. Diagrama em blocos do sistema proposto.

1) Pré-processamento: A etapa de pré-processamento, consiste basicamente na obtenção da sequência residual. Para dois quadros adjacentes I_n e I_{n+1}, a imagem residual é R_n dada pela Equação (1). A geração de quadros residuais está ilustrada na Figura 3.

$$R_n = I_{n+1} - I_n. \tag{1}$$

Fig. 3. Geração de quadros residuais.

2) Extração de características: É razoável supor que o movimento é uma das características que mais discrimina as palavras em LIBRAS. Por outro lado, as imagens residuais ressaltam o movimento. Assim, a extração de características é construída sobre as imagens residuais justamente para incorporar as características de movimento no espaço de características. A primeira etapa consiste em somar 10 imagens residuais R_n igualmente espaçadas para compor a matriz de resíduos R de acordo com a Equação (2).

$$R = \sum_i R_i, \tag{2}$$

onde $i = \{1, N/10, N/10+10, N/10+20, \cdots, N\}$, sendo N o número total de imagens residuais R_n.

Em seguida, cada matriz residual R é subdividida em K^2 células $C_{i,j}$, $0 < (i,j) < K$. A título de exemplo, a etapa de quantização da matriz residual em 36 células ($K = 6$) está ilustrada Figura 4. Cada célula $C_{i,j}$ é mapeada em um elemento (i,j) da matriz de características F, de dimensão $K \times K$, de acordo com a Equação 3.

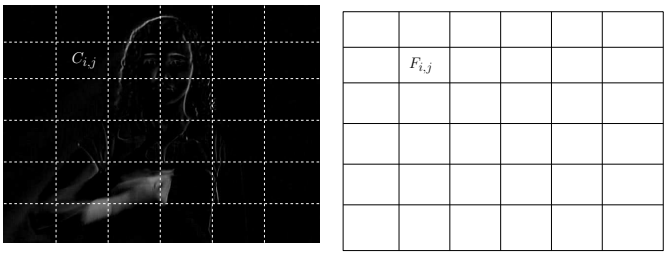

Fig. 4. Formação da matriz de características.

$$F_{i,j} = \sum_{\forall p,q} C_{i,j}(p,q), \tag{3}$$

onde $F_{i,j}$ é o elemento (i,j) da matriz de características F e $C_{i,j}(p,q)$ é um pixel (p,q) da célula $C_{i,j}$. Como resultado, a matriz de resíduo foi mapeada em um espaço de características de dimensão K^2, menor que a dimensão das imagens originais. O principal objetivo deste mapeamento é reduzir a complexidade computacional do sistema.

3) Classificação: A etapa de classificação é baseada no k-NN mais simples que existe, o 1-NN. Deseja-se distinguir C palavras em LIBRAS. Neste contexto, cada palavra é uma classe e $\boldsymbol{\mu}_c$, $c = \{1, 2, \cdots C\}$ é a média do conjunto de treinamento de cada classe. A distância euclidiana $d(X, \boldsymbol{\mu}_c)$ entre uma amostra X de classe desconhecida e a media de uma classe $\boldsymbol{\mu}_c$, em um espaço métrico $\mathcal{X}$, pode ser definida de acordo com a Equação (4):

$$d(X, \boldsymbol{\mu}_c) = \|X - \boldsymbol{\mu}_c\| = \sqrt{(X - \boldsymbol{\mu}_c) \cdot (X - \boldsymbol{\mu}_c)}. \tag{4}$$

Com isso, é possível classificar uma amostra desconhecida X utilizando a Equação (5).

$$D_c = \{X \in \mathcal{X} | d(X, \boldsymbol{\mu}_c) \leq d(X, \boldsymbol{\mu}_l),\ \forall l \neq c\}. \tag{5}$$

A Região D_c é chamada de região de Voronoi, e corresponde a uma porção do espaço métrico que corresponde a uma classe c. A união de todas as regiões de Voronoi forma o diagrama de Voronoi.

III. Resultados e discussões

Nesta seção são apresentadas as condições sob as quais os experimentos foram conduzidos, os resultados obtidos, bem como uma discussão acerca destes resultados. O sistema proposto neste trabalho foi desenvolvido em Matlab.

A. Treinamento e teste

Embora a base de dados proposta seja composta por 24 verbetes, somente 12 estavam disponíveis no momento da condução dos experimentos descritos nesta seção: Andaime, Depressão, Fantasma, Martelo, Nacionalidade, Neve, Paciência, Parede, Redação, Relâmpago, Trampolim e Vapor. Em outras palavras, das 548 sequências de vídeo da base de dados disponibilizada, apenas 450 foram utilizadas nos experimentos. Este subconjunto foi dividido de modo que 70% das amostras de cada palavra foram reservadas para treinamento e 30% para teste.

O treinamento consiste em determinar as médias μ_c de cada classe, de acordo com a Equação 6.

$$\mu_c = \frac{1}{M} \sum_{m=1}^{M} X_m^c, \tag{6}$$

onde X_m^c é uma amostra do conjunto de treinamento da classe c e M é o total de amostras pertencentes ao conjunto de treinamento da classe c.

B. Escolha da dimensão da matriz de características

A complexidade computacional do sistema proposto é diretamente proporcional ao tamanho K da matriz de características. Portanto, foram elaborados experimentos envolvendo diversos tamanhos desta matriz com o objetivo de obter um um sistema que utiliza o menor tamanho K possível. Os resultados deste experimento estão na Tabela I. Cada entrada desta tabela contém a taxa de acerto no conjunto de teste para uma palavra e um tamanho K. Foram testados valores de $K = 3$ até 39.

Com o objetivo de determinar o menor tamanho possível da matriz de características, foi elaborado o gráfico da Figura 5. Cada ponto deste gráfico representa o valor médio da das taxas de acerto de uma linha da Tabela I. Em outras palavras, cada ponto é a taxa de acerto médio de todas as palavras para um dado tamanho K da matriz de características. Analisando os resultados na Figura 5, é possível observar que, do ponto de vista da taxa de acerto, não há vantagem significativa em usar tamanhos da matriz de características maiores que $K = 7$. Como resultado, foi desenvolvido um sistema de reconhecimento de gestos em LIBRAS com baixo custo e com taxa média de acerto de 75%, que é considerado satisfatório.

IV. Conclusões e trabalhos futuros

Este trabalho possui duas contribuições principais. A primeira delas é a disponibilização de uma nova base de dados de palavras em LIBRAS contendo 548 sequências de vídeo, distribuídas em 24 verbetes. Os verbetes podem ser subdivididos em dois grupos: um cujo o deslocamento das mãos segue uma trajetória horizontal. Cada amostra é composta por uma sequências de vídeo de resolução 640×480 *pixels*. As gravações foram feitas em ambiente com três tipos de plano de fundo estático: simples, intermediário e complexo. Elas foram executadas por 20 voluntários não surdos, sendo 13 mulheres e 7 homens.

A segunda contribuição deste trabalho é o desenvolvimento de um sistema de reconhecimento de palavras em LIBRAS

TABELA I

TAXAS DE ACERTO PARA 12 PALAVRAS AVALIADAS NESTE TRABALHO E TAMANHOS DA MATRIZ DE CARACTERÍSTICAS DE $K = 3$ ATÉ 39.

$K \times K$	Gestos											
	Andaime	Depressão	Fantasma	Martelo	Nacionalidade	Neve	Paciência	Parede	Redação	Relâmpago	Trampolim	Vapor
3x3	25,00	50,00	41,67	57,14	37,50	0,00	55,56	66,67	0,00	60,00	40,00	60,00
5x5	50,00	80,00	50,00	50,00	87,50	37,50	66,67	77,78	33,33	40,00	50,00	70,00
7x7	66,67	100,00	66,67	71,43	87,50	62,50	100,00	66,67	88,89	90,00	50,00	50,00
9x9	66,67	90,00	66,67	78,57	87,50	43,75	77,78	66,67	77,78	60,00	80,00	60,00
11x11	66,67	60,00	50,00	78,57	62,50	75,00	77,78	33,33	77,78	60,00	60,00	60,00
13x13	58,33	80,00	91,67	64,29	100,00	81,25	88,89	55,56	44,44	80,00	80,00	60,00
15x15	58,33	80,00	66,67	71,43	87,50	68,75	77,78	55,56	66,67	60,00	80,00	60,00
17x17	41,67	90,00	91,67	85,71	50,00	75,00	88,89	44,44	66,67	70,00	70,00	70,00
19x19	58,33	100,00	58,33	64,29	87,50	68,75	88,89	66,67	77,78	50,00	80,00	80,00
21x21	58,33	90,00	58,33	71,43	87,50	56,25	88,89	55,56	88,89	100,00	70,00	70,00
23x23	50,00	100,00	50,00	78,57	75,00	68,75	88,89	44,44	66,67	80,00	70,00	90,00
25x25	66,67	90,00	58,33	92,78	75,00	50,00	100,00	22,22	77,78	80,00	90,00	90,00
27x27	83,33	80,00	58,33	85,71	75,00	0,00	100,00	44,44	66,67	80,00	30,00	80,00
29x29	66,67	100,00	91,67	64,29	75,00	75,00	88,89	66,67	55,56	60,00	80,00	70,00
31x31	41,67	90,00	83,33	64,29	87,50	62,50	100,00	66,67	66,67	90,00	70,00	90,00
33x33	83,33	90,00	58,33	50,00	100,00	0,00	88,89	77,78	55,56	90,00	70,00	90,00
35x35	66,67	90,00	91,67	71,43	87,50	0,00	66,67	66,67	44,44	90,00	40,00	70,00
37x37	75,00	100,00	66,67	78,57	75,00	68,75	88,89	55,56	88,89	50,00	80,00	70,00
39x39	41,67	90,00	91,67	71,43	87,50	68,75	88,89	66,67	77,78	60,00	50,00	80,00

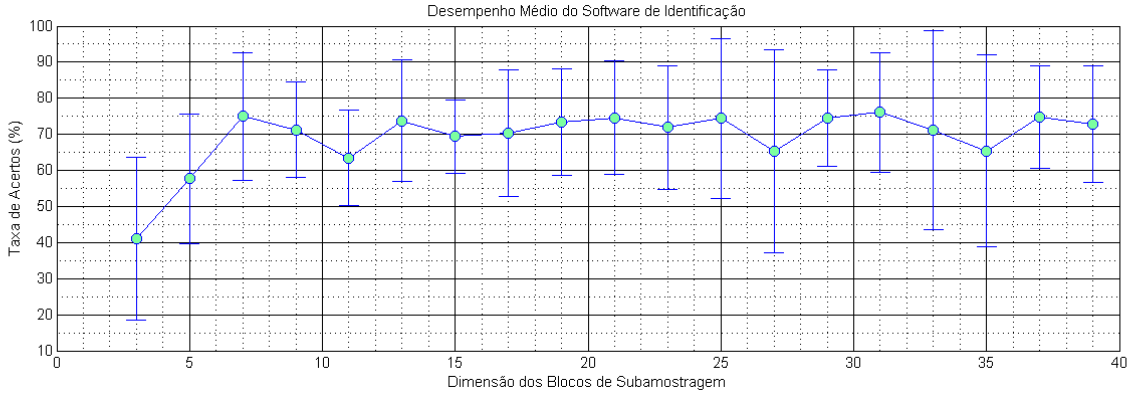

Fig. 5. Desempenho médio do sistema de identificação

de baixo custo usando visão computacional. O sistema de reconhecimento é baseado na subamostragem de imagens residuais, obtidas pela subtração de quadros sucessivos e na classificação usando um algoritmo 1-NN. O desempenho médio do sistema é de 75%. Embora haja margem para diversas melhorias, este resultado é considerado satisfatório.

Existem diversas vertentes para trabalhos futuros. Como a base de dados está em expansão, espera-se disponibilizar mais palavras em breve. Com relação ao sistema proposto, espera-se melhorar o desempenho global atacando melhorando o sistema de extração de características, bem como empregando classificadores mais robustos. Com relação à etapa experimental, pretende-se rodar os experimentos usando a base de dados completa.

REFERÊNCIAS

[1] C. Harshith, Karthik R. Shastry, Manoj Ravindran, M. V. V. N. S. Srikanth, and Naveen Lakshmikhanth, "Survey on various gesture recognition techniques for interfacing machines based on ambient intelligence," *CoRR*, vol. abs/1012.0084, 2010.

[2] M. Panwar, "Hand gesture recognition based on shape parameters," in *2012 International Conference on Computing, Communication and Applications*, Feb 2012, pp. 1–6.

[3] X. Zhao, A. M. Naguib, and S. Lee, "Kinect based calling gesture recognition for taking order service of elderly care robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, Aug 2014, pp. 525–530.

[4] M. K. Ahuja and A. Singh, "Static vision based hand gesture recognition using principal component analysis," in *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on*, Oct 2015, pp. 402–406.

[5] R. Xie, X. Sun, X. Xia, and J. Cao, "Similarity matching-based extensible hand gesture recognition," *IEEE Sensors Journal*, vol. 15, no. 6, pp. 3475–3483, June 2015.

[6] A. Kar and P. S. Chatterjee, "An approach for minimizing the time taken by video processing for translating sign language to simple sentence in english," in *Computational Intelligence and Networks (CINE), 2015 International Conference on*, Jan 2015, pp. 172–177.

[7] M. Takai, "Measurement of motion quantity from human movement and detection of the sign language word," in *The 2012 International Conference on Advanced Mechatronic Systems*, Sept 2012, pp. 298–303.

[8] M. Fernando and J. Wijayanayaka, "Low cost approach for real time sign language recognition," in *2013 IEEE 8th International Conference on Industrial and Information Systems*, Dec 2013, pp. 637–642.

[9] M. Mohandes, Junzhao Liu, and M. Deriche, "A survey of image-based arabic sign language recognition," in *Systems, Signals Devices (SSD), 2014 11th International Multi-Conference on*, Feb 2014, pp. 1–4.

[10] S. Liu and Q. Xiao, "A signer-independent sign language recognition system based on the weighted knn/hmm," in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on*, Aug 2015, vol. 2, pp. 186–189.

[11] INES, "Dicionário da língua brasileira de sinais," 2005.