Mestrado Multiinstitucional de Pós-Graduação em Ciência da Computação - MMCC

RECONHECIMENTO DE SINAIS DA LIBRAS UTILIZANDO DESCRITORES DE FORMA E REDES NEURAIS ARTIFICIAIS

Por

Igor Leonardo Oliveira Bastos

Dissertação de Mestrado

SALVADOR

Maio/2015

IGOR LEONARDO OLIVEIRA BASTOS

RECONHECIMENTO DE SINAIS DA LIBRAS UTILIZANDO DESCRITORES DE FORMA E REDES NEURAIS ARTIFICIAIS

Dissertação de Mestrado apresentada ao Mestrado Multiinstitucional de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia e Universidade Estadual de Feira de Santana como requisito para obtenção do grau de Mestre em Ciência da Computação.

Orientador: *Michele Fúlvia Angelo*
Co-Orientador: *Angelo Conrado Loula*

SALVADOR
Maio/2015

Aos meus pais e irmão, Milton, Nalma e Miltinho e; ao meu amor, Larissa.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, o qual tem me orientado e me fornecido todas as ferramentas para a conclusão deste trabalho.

Agradeço aos meus pais e irmão pelo apoio dado e compreensão em momentos de dificuldade.

Agradeço a Larissa, fonte de minha paz e serenidade e responsável por me alegrar mesmo quando as coisas não saíam como esperado.

Agradeço aos professores intérpretes de Libras e alunos do Instituto Municipal de Educação (INSME) da cidade de Itaberaba-Bahia, essenciais para a aquisição de imagens e, consequentemente, para o desenvolvimento deste trabalho.

Agradeço também aos meus orientadores Michele e Angelo, os quais me auxiliaram e foram essenciais para garantir a qualidade e corretude de toda a metodologia empregada.

Por fim, agradeço a todos que contribuíram, direta ou indiretamente, na construção deste trabalho. Obrigado.

Grandes poderes trazem grandes responsabilidades.

—BEN PARKER

Resumo

Gestos são ações corporais não-verbais voltadas para a expressão de algum significado. Estes incluem movimentos de mãos, face, braços, dedos, entre outros, sendo abordados por trabalhos que visam reconhecê-los para promover interações humanas com sistemas computacionais. Devido à grande aplicabilidade do reconhecimento de gestos, tem-se notado que estes trabalhos estão se tornando mais comuns, utilizando técnicas e metodologias mais elaboradas e capazes de prover resultados cada vez melhores. A opção por quais técnicas aplicar para o reconhecimento de gestos varia de acordo com a estratégia empregada em cada trabalho e quais aspectos são utilizados para este reconhecimento. Tem-se, por exemplo, trabalhos baseados no uso de modelos estatísticos. Outros optam pela aquisição de características geométricas de mãos e partes do corpo, enquanto outros, dentre os quais se enquadra o presente trabalho, optam pelo uso de descritores e classificadores, responsáveis por extrair características das imagens relevantes para o seu reconhecimento e; por realizar a classificação efetiva dos gestos baseado nestas informações. Neste âmbito, o presente trabalho visa elaborar, aplicar e apresentar uma abordagem para o reconhecimento de gestos, embasando-se em uma revisão da literatura a respeito das principais técnicas e metodologias empregadas para este fim e escolhendo como campo prático, a Língua Brasileira de Sinais (Libras). Para a extração de informações das imagens, optou-se pelo uso de um vetor de características resultante da aplicação dos descritores Histograma de Gradientes Orientados (HOG) e Momentos Invariantes de Zernike (MIZ), os quais voltam-se para as formas e contornos presentes nas imagens. Para o reconhecimento, foi utilizado o classificador Perceptron Multicamada, sendo este disposto em uma arquitetura onde o processo de classificação é dividido em 2 estágios. Devido à inexistência de *datasets* públicos da Libras, fez-se necessária, com o auxílio de especialistas da língua e alunos surdos, a criação de um *dataset* de 9600 imagens, as quais referem-se a 40 sinais da Libras. Isso fez com que a presente abordagem partisse desta criação do *dataset* até a etapa final de classificação dos sinais. Por fim, testes foram realizados e obteve-se 96,77% de taxa de acerto, evidenciando um alto índice de acerto. Este resultado foi validado considerando possíveis ameaças à abordagem, como a realização de testes considerando um indivíduo não-presente no conjunto de treinamento do classificador e a aplicação da abordagem em um *dataset* público de gestos.

Palavras-chave: Reconhecimento de gestos, Histograma de Gradientes Orientados, Momentos Invariantes de Zernike, Redes Neurais Artificiais, Libras.

Abstract

Gestures are nonverbal bodily actions used for the expression of some meaning. These include movements of hand, face, arms, fingers, and others, and they are addressed by work that aim at recognizing them to promote human interactions with computer systems. Because of the wide applicability of the gesture recognition, it has been noticed that these works are becoming increasingly common, and the techniques and methodologies applied are getting even more elaborated, providing better and better results. The choice of techniques applied to the recognition of gestures varies according to the strategy employed in each work and which aspects are used for this recognition. For instance, there are studies based on use of statistical models. Other studies are based on the acquisition of geometrical characteristics of hands and body parts. On the other hand, studies, including the present one, uses descriptors and classifiers, responsible by extracting features of images that are relevant for recognition and classifying the gestures based on this information. In this context, this work aims to develop, implement and present an approach to recognize gestures, being based in a literature review about the main techniques and methodologies used for this purpose. The Brazilian Sign Language (Libras) was choosed as practical field. In order to extract the image information, we composed a feature vector resultant from the application of the descriptors Histogram Oriented Gradients (HOG) and Invariant Zernike Moments (MIZ), which gather information about shapes and edges in the images. For the recognition, the classifier Multilayer Perceptron was used, which is arranged in a architecture where the classification process is divided into two stages. Due to the absence of public Libras datasets, we created, with the help of experts and deaf students, a dataset containing 9600 images related to 40 Libras signs. In this way, the approach started with the creation of the image dataset and had the classification of the signs as final step. At the end, tests were conducted and yielded 96.77% recognition rate, revealing a high success rate. This result was validated considering potential threats to the approach, such as tests considering a non-present person at the training set of the classifier and tests applying the approach in a public gesture dataset.

Keywords: Gesture Recognition, Histogram of Oriented Gradients, Invariant Zernike Moments, Artificial Neural Networks, Libras.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

RGB Red, Green and Blue - Vermelho, Verde e Azul

HSV Hue, Saturation and Value - Matiz, Saturação e Valor

YUV/YCbCr Luminance, Blue-Chrominance and Red-Chrominance - Luminância, Crominância-Azul e Crominância-Vermelha

TSL Tint, Saturation and Luminance - Tinta, Saturação e Luminância

MIZ Momentos Invariantes de Zernike

HOG Histogram of Oriented Gradients - Histograma de Gradientes Orientados

SVM Support Vector Machine - Máquina de Vetores de Suporte

ANN Artificial Neural Network - Rede Neural Artificial

NN Nearest Neighbors - Vizinhança (Vizinhos) mais próxima

MMD Minimum Mean Distance - Menor Distância Média

1 Introdução

O reconhecimento de gestos, com o auxílio de recursos e técnicas de visão computacional, tem sido cada vez mais abordado e discutido na comunidade científica. Metodologias distintas têm sido empregadas para este fim (Pavlovic *et al.*, 1997), variando desde trabalhos que utilizam ferramentas auxiliares para os processos de rastreio das mãos (tracking) e reconhecimento, como no proposto por Parvini *et al.* (2009), até outros que realizam este reconhecimento sem o advento de qualquer ferramenta, como nos trabalho de Yang and Xu (1994), Liwicki and Everingham (2009) e Bowden *et al.* (2004).

Os gestos incluem ações e movimentos de mãos, dedos, face, braços, entre outros (Mitra and Acharya, 2007), sendo que a sua aplicabilidade é extensa. Como exemplo, tem-se sistemas de navegação virtual, elaboração de ferramentas para auxiliar deficientes auditivos, detecção de mentiras, sistemas de identificação forense e, também, reconhecimento de línguas de sinais (Mitra and Acharya, 2007).

As línguas de sinais representam o principal meio de comunicação entre deficientes auditivos. Estas não são universais, na medida em que cada país possui a sua própria língua de sinais, a qual é afetada por aspectos de sua cultura (Teodoro and Digiampietri, 2013). Estas têm sido utilizadas como um campo prático para a aplicação do reconhecimento de gestos. sendo que algumas merecem destaque pela quantidade de trabalhos encontrados que relacionam-se ao reconhecimento de gestos que as compõem, como: a *American Sign Language* (ASL), a *British Sign Language* (BSL) e; a Língua Brasileira de Sinais (Libras), sendo a última, oficialmente, adotada pelo governo como uma língua falada no Brasil (Felipe, 2007).

A comunicação através de línguas de sinais se dá a partir do canal visual-espacial (Sousa, 2010) e baseia-se em alguns parâmetros que relacionam-se ao significado dos gestos. No caso da Libras, estes parâmetros são: (a) a configuração das mãos, (b) o movimento realizado com as mãos, (c) lugar de articulação (lugar onde o gesto é executado), (d) orientação da palma

e; (e) expressões não-manuais (expressões corporais e faciais) (Felipe, 2007; Teodoro and Digiampietri, 2013).

A dificuldade inerente ao reconhecimento de uma grande gama de gestos e a consideração de todos os parâmetros representam um dos maiores obstáculos no tocante à elaboração de sistemas que visam operar com línguas de sinais. No entanto, técnicas cada vez mais elaboradas vêm sendo empregadas para este reconhecimento, as quais têm conseguido abranger cada vez mais sinais e com maiores taxas de acerto. Assim, a presente dissertação visa apresentar uma abordagem para reconhecimento de gestos combinando descritores de forma, técnicas de processamento de imagens digitais e classificação utilizando redes neurais artificias. Como campo prático, optou-se pela Libras, a qual é também abordada em trabalhos de reconhecimento de gestos como os propostos por Pizzolato *et al.* (2010), Carneiro *et al.* (2010) e Anjo *et al.* (2012).

No entanto, a presente abordagem avança, quando comparada às de Pizzolato *et al.* (2010), Carneiro *et al.* (2010) e Anjo *et al.* (2012), por abranger uma maior quantidade de sinais e por realizar o processo de reconhecimento utilizando uma combinação de descritores de forma, adicionando robustez ao processo e elevando as taxas de acerto. Além disso, a disponibilização do *dataset* de imagens construído, passo final da abordagem, constitui um outro diferencial em relação aos projetos mencionados e pode auxiliar no tocante à comparação de resultados de abordagens voltadas ao reconhecimento de sinais da Libras.

1.1 A Problemática

Apesar de recentes avanços no tocante ao desenvolvimento de práticas inclusivas no Brasil (Felipe, 2007), nota-se que há ainda uma grande dificuldade por parte dos surdos na utilização da língua Portuguesa (escrita), conforme registrado por Capovilla (2008). Isto representa uma grande barreira social aos deficientes, muitas vezes, os privando de direitos básicos como saúde, lazer e educação (Carneiro, 2010). Outro ponto abordado por Capovilla (2008) é a carência de recursos que atendam às necessidades próprias da comunidade surda brasileira, sendo que muitos dos softwares de auxílio aos deficientes auditivos são importados e não são adequados aos usuários da Libras (Faria *et al.*, 2001).

Assim, o presente trabalho foi proposto de forma a se enquadrar na carência encontrada pela comunidade surda brasileira, sendo um auxiliar no processo de comunicação dos surdos ao apresentar uma abordagem desenvolvida para o reconhecimento de gestos manuais que correspondam a sinais da Libras, traduzindo-os para a língua Portuguesa. Diante da gama de

sinais existentes na Libras e da quantidade de parâmetros que esta utiliza, a presente abordagem representa o reconhecimento de uma parte da língua. Porém, esta pode ser tomada como ponto inicial para futuros trabalhos, na medida em que as técnicas empregadas, além da formalização de um *dataset* de sinais de Libras, podem ser estendidos e utilizados em aplicações maiores.

1.2 Objetivos do trabalho

O presente trabalho tem como objetivo geral apresentar uma abordagem baseada no uso dos descritores de forma Histograma de Gradientes Orientados (HOG) e Momentos Invariantes de Zernike (MIZ), juntamente com um classificador neural disposto em 2 etapas. Este classificador é voltado para o reconhecimento de gestos manuais que fazem parte da Libras.

Esta abordagem apresenta objetivos específicos que vão desde a construção do *dataset* de imagens utilizado até a realização de testes e validação dos resultados encontrados. Assim, os objetivos específicos são os seguintes:

- Revisão dos principais métodos e trabalhos que visam reconhecer gestos em imagens digitais, com ênfase em trabalhos voltados para o reconhecimento de gestos pertencentes às línguas de sinais.
- Elaboração e avaliação de abordagem para segmentação de pele em imagens digitais.
- Análise e seleção de sinais pertencentes à Libras a serem reconhecidos.
- Criação de um *dataset* de imagens contendo os gestos selecionados para os processos de treinamento e teste do classificador.
- Avaliação dos descritores de forma usados no reconhecimento de gestos e seleção dos que mais se adequam à presente abordagem.
- Ajuste de parâmetros, treinamento e teste do classificador de 2 estágios utilizado.
- Aquisição e análise dos resultados obtidos.
- Validação dos resultados encontrados através de testes que visam avaliar a robustez da abordagem.

1.3 Escopo do Trabalho

O presente trabalho relaciona-se ao reconhecimento de gestos aplicado à Libras. No entanto, o reconhecimento de gestos representa uma grande área da visão computacional e compreende uma grande quantidade de técnicas e métodos. Por isso, esta dissertação realizou uma revisão de uma parcela destes trabalhos, com ênfase naqueles voltados para as línguas de sinais.

Quanto a forma que a abordagem atua, esta se dá reconhecendo gestos (sinais) que fazem parte da Libras. Todos estes sinais são estáticos e, devido ao fato de ter-se utilizado descritores ligados às formas presentes nas imagens, tomou-se, como único parâmetro, a configuração das mãos. Gestos dinâmicos e outros parâmetros da Libras, como expressões faciais/corporais, não são considerados.

Outro ponto a ser ressaltado é que a Libras também compreende um grande campo de estudo. Como o presente trabalho visa apresentar uma abordagem para o reconhecimento de gestos utilizando técnicas computacionais, este restringiu-se a apresentar aspectos da Libras relevantes para este reconhecimento. Os estudos realizados a respeito da língua restringiram-se ao conhecimento de sinais da Libras, seus parâmetros e ao entendimento de como o presente trabalho poderia ser aplicado a esta língua.

1.4 Contribuições do trabalho

O presente trabalho apresenta as seguintes contribuições:

- **Dataset de imagens da Libras.** Este *dataset* é composto por 9600 imagens, as quais referem-se a 40 diferentes gestos. Todos as imagens correspondem a posturas de mão estáticas, adquiridas com o auxílio de especialistas e estudantes surdos fluentes na língua. Pretende-se disponibilizar este *dataset* na Internet, de forma que o mesmo possa ser usado para a comparação de técnicas de reconhecimento de sinais da Libras.

- **Abordagem para o reconhecimento de pele.** Com base em estudos relacionados ao reconhecimento de pele em imagens digitais, foi proposta neste trabalho uma abordagem combinando elementos destes trabalhos (informações quanto aos espaços de cores utilizados) e classificação utilizando redes neurais artificiais, chegando a um método para o realce de pele que apresenta altas taxas de acerto.

- **Abordagem para reconhecimento de gestos.** A abordagem proposta neste trabalho baseia-se na combinação de 2 descritores de forma e no uso de classificadores neurais

Perceptron Multicamada. Esta se mostra inovadora por combinar estes descritores e por realizar o reconhecimento dos gestos através de uma arquitetura de classificadores em 2 estágios.

1.5 Organização da dissertação

Excluindo-se o capítulo de Introdução (Capítulo 1), o presente trabalho está dividido em 4 capítulos, os quais são descritos a seguir.

O Capítulo 2 destina-se a revisão bibliográfica realizada para a construção da abordagem. Esta aborda as principais técnicas aplicadas no presente trabalho, como as técnicas para realce de pele, os descritores HOG e MIZ; e o classificador Perceptron Multicamada. Além disso, esta revisão aborda os principais trabalhos voltados para o reconhecimento de gestos, com destaque para aqueles relacionados às línguas de sinais.

No Capítulo 3, a metodologia empregada é apresentada. Neste capítulo são apresentadas as etapas realizadas para a construção do presente trabalho, justificando a opção por cada técnica empregada e apresentando o fluxo de dados que ocorre desde o processo de aquisição das imagens até a classificação.

No Capítulo 4, os resultados e discussões são apresentados. Estes resultados decorrem do processo de classificação e dos testes empregados para validar a abordagem.

Por fim, no Capítulo 5, são apresentadas as conclusões do trabalho, levando-se em consideração os resultados encontrados. Além disso, as contribuições do presente trabalho são apresentadas, assim como propostas para trabalhos futuros.

1.6 Produção Científica

Durante o desenvolvimento deste trabalho, três artigos foram publicados em eventos:

- Bastos, Igor; Angelo, Michele ; Loula, Angelo. Gesture Recognition using Shape Characteristics. Proceedings of XXVII Sibgrapi Conference on Graphics, Patterns and Images - WIP, 2014. Rio de Janeiro (Premiado como melhor trabalho de Computação Gráfica/Visual do WIP).

- Bastos, Igor L. O.; Angelo, Michele F. Reconhecimento de pele em imagens digitais utilizando redes neurais artificiais. Proceedings of X Workshop de Visão Computacional, 2014. Uberlândia. Páginas 125-130.

- Bastos, Igor Leonardo Oliveira; Angelo, Michele Fulvia. Classificação de gestos da Libras utilizando Redes Neurais. Anais da XIV Escola Regional de Computação Bahia-Alagoas-Sergipe - ERBASE - WPos, 2014. Feira de Santana.

2

Revisão de Literatura

A construção de sistemas e abordagens que visam reconhecer gestos em imagens tem sido bastante frequente nos últimos anos (Khan and Ibraheem, 2012), sendo que as línguas de sinais têm sido utilizadas como um campo prático para estes trabalhos. Seguindo esta linha enquadra-se o presente trabalho, o qual visa apresentar uma abordagem desenvolvida para o reconhecimento, em imagens digitais, de gestos (sinais) da Língua Brasileira de Sinais (Libras).

De forma a embasar a construção da metodologia a ser empregada neste trabalho, foi feito inicialmente um esforço no tocante ao estudo de trabalhos relacionados ao reconhecimento de gestos, técnicas empregadas para este fim e, por fim, a respeito da Libras, a qual foi escolhida como campo prático para a aplicação da abordagem.

Devido a opção pelo uso da Libras, fez-se necessário o entendimento de aspectos relacionados à mesma, com ênfase na construção dos seus sinais e nos aspectos relevantes para a identificação dos mesmos. Este entendimento foi embasado em trabalhos a respeito da língua, os quais tratam desde o surgimento desta até os elementos que constituem os parâmetros para a identificação de cada sinal. Assim, a primeira parte desta revisão objetiva dar um breve apanhado sobre a Libras.

Com o intuito de segmentar regiões de interesse nas imagens dos gestos (mãos), chegou-se a trabalhos e técnicas que auxiliassem neste processo. Devido à simplicidade da aplicação e adequabilidade à presente abordagem, optou-se pela segmentação das zonas de pele como estratégia para a segmentação das mãos. Assim, levantou-se trabalhos a respeito desta estratégia de segmentação, os quais permitiram a elaboração de uma técnica própria para este fim. Os trabalhos utilizados para embasar a construção da estratégia desenvolvida para a segmentação de pele são abordados nesta revisão.

A presente revisão também aborda, após os trabalhos de segmentação de pele, os descritores utilizados: o Histograma de Gradientes Orientados (HOG) e os Momentos Invariantes de Zernike

(MIZ). Esta traz trabalhos que empregaram estes dois descritores para o fim de reconhecimento, os quais incluem trabalhos de reconhecimento de gestos. Além disso, aspectos téoricos a respeito de ambos são apresentados e discutidos, sendo estes essenciais para a compreensão de como estes atuam sobre as imagens digitais e como devem ser empregados para a obtenção de melhores resultados. Um raciocínio análogo é feito para a inclusão, nesta revisão, de uma fundamentação a respeito do classificador Perceptron Multicamada, trazendo os seus aspectos, características e embasando a escolha pelo mesmo.

Por fim, apresenta-se nesta revisão trabalhos relacionados ao reconhecimento de gestos, os quais são, em boa parte, relacionados a línguas de sinais. Estes trabalhos fundamentaram a construção de toda a metodologia da presente abordagem, alicerçando a escolha pelos descritores HOG e MIZ, pelo classificador Perceptron e pela forma como o fluxo de informações, desde o processo de aquisição de imagens até o reconhecimento efetivo do sinal, foi elaborado.

Desta maneira, esta revisão de literatura foi dividida em 5 partes: a primeira relacionada à Libras e construção dos sinais da língua; a segunda relacionada à abordagem de reconhecimento de pele em imagens digitais; seguida pelo levantamento de trabalhos ligados aos descritores utilizados. Em seguida, o classificador Perceptron Multicamada é abordado, com a apresentação de elementos referentes à teoria que rege o funcionamento do mesmo. Por fim, tem-se uma seção relacionada aos trabalhos que visam realizar o reconhecimento de gestos em imagens digitais, com ênfase nas línguas de sinais e na Libras.

2.1 Libras e os parâmetros para a construção dos sinais

O conhecimento a respeito da Libras fez-se necessário para alicerçar o presente trabalho. Na tentativa de entender a forma como os sinais desta língua são formados e os elementos relevantes para o reconhecimento destes, fez-se um estudo partindo da história da Libras até se chegar aos elementos que compõem os sinais: os parâmetros.

2.1.1 Histórico

As línguas de sinais são conhecidas como línguas naturais devido à sua espontaneidade no seu surgimento, o qual aconteceu de forma não-prevista ou não-planejada. Assim como línguas orais, estas promovem a interação entre pessoas e as suas estruturas permitem a expressão de qualquer conceito, seja de cunho descritivo, emotivo, racional, literal, metafórico ou abstrato (Brito, 2010).

Apesar da capacidade de expressão tão grande quanto a de línguas orais, as línguas de

sinais se diferenciam na medida em que utilizam, para a transmissão de mensagens, o canal visual-espacial ao invés do canal oral-auditivo (Sousa, 2010). No contexto de línguas de sinais, enquadram-se a *American Sign Language* (ASL), a *British Sign Language* (BSL), entre outras, além da Libras, a qual foi adotada como uma língua oficial do Brasil desde 2002 (Felipe, 2007).

Apesar de só adotada como língua oficial recentemente, o surgimento da Libras e sua difusão remetem ao império de D. Pedro II e à fundação, no Rio de Janeiro, do Instituto Nacional de Surdos-Mudos (INSM) em 1857, sendo este considerado o marco para o início oficial desta língua (Almeida and Almeida, 2012). Este surgimento é também associado à chegada ao Brasil do professor francês Eduard Huet, fato este importante para justificar a forte influência da Língua Francesa de Sinais na formação da Libras (Ramos, 2013). A fundação do INSM (hoje é conhecido como INES - Instituto Nacional de Educação dos Surdos) estimulou a migração de surdos das mais diversas partes do Brasil para o Rio de Janeiro, sendo as décadas seguintes à sua fundação marcadas por eventos importantes na formalização da língua, como em 1873, quando foi realizada a iconografia dos sinais, de autoria do aluno do INSM Flausino José de Gama (Monteiro, 2006).

Em 1881, o uso da Libras no INES foi proibido, assim como em todo o Brasil, representando um grande retrocesso na formalização e difusão da língua (Monteiro, 2006), sendo esta língua oficialmente proibida dos anos de 1957 até meados da década de 80 (Ramos, 2013). Apesar disso, a sua utilização nunca deixou de acontecer dentro do INES e o seu uso voltou a ganhar força nas décadas seguintes, chegando à sua adoção como língua oficial brasileira em 2002 (Felipe, 2007).

2.1.2 Parâmetros da Libras

A formalização das línguas de Sinais foi primeiramente abordada por Stokoe (1960), o qual estabeleceu a existência de três parâmetros principais (configuração da mão, locação da mão e movimento). Os estudos de Stokoe foram embasados na *American Sign Language*. No caso da Libras, alguns outros parâmetros foram adicionados. São eles: orientação da mão, expressões faciais e expressões corporais (Sousa, 2010).

Segundo Strobel and Fernandes (1998), a configuração das mãos corresponde à forma assumida pela mão durante a articulação de um sinal. Já a locação da mão define o ponto do corpo onde o sinal será realizado. O movimento, por sua vez, demonstra o deslocamento da mão durante a execução do sinal, possuindo diferentes formas e direções. Apesar do movimento ser considerado um parâmetro, alguns sinais da Libras são estáticos (Sousa, 2010).

Já Quadros and Karnopp (2004) definem o parâmetro de orientação das mãos como sendo a

direção em que a palma da mão indica na realização do sinal. Segundo ele, as expressões faciais e corporais, chamadas de componentes não-manuais, são usadas para a distinção do significado de alguns sinais, além de serem importantes na representação de sentimentos (Quadros and Karnopp, 2004).

2.2 Reconhecimento de pele

A detecção de pele em imagens digitais tem sido utilizada cada vez mais nos últimos anos, sendo que esta se mostra frequente em aplicações como as de vídeo-vigilância, reconhecimento de face e/ou gestos, além de aplicações que realizam interação homem-computador através do reconhecimento de partes do corpo (Prema and Manimegalai, 2012). Muitos são os métodos utilizados para o reconhecimento de pele em imagens digitais. Dentre estes, os que apresentam abordagens mais simples são aqueles que utilizam as componentes de diferentes espaços de cores e que defendem a estipulação de intervalos dentro dos quais os pixels são considerados pele (Vezhnevets *et al.*, 2003). Para uma melhor compreensão destas técnicas, é necessário que se entenda quais são os principais espaços de cores utilizados nestes algoritmos e a forma como os seus canais são dispostos.

2.2.1 Espaços de cores

Segundo Foley *et al.* (1990), espaços de cores são sistemas tridimensionais de coordenadas, onde cada eixo, ou canal, representa um aspecto primário. A reprodução de cada cor é dada pela combinação de valores dos canais. Esta seção visa apresentar 3 diferentes espaços de cores, comumente utilizados em trabalhos de reconhecimento de pele: um espaço baseado no modelo RGB, os espaços HSV e YUV.

2.2.1.1 Espaço baseado no modelo RGB

O espaço de cores baseado no modelo RGB é um dos mais utilizados, sendo também bastante aplicado em trabalhos de detecção de pele. Este espaço é composto por 3 canais, os quais representam suas cores primárias: *Red* (vermelho), *Green* (verde) e *Blue* (azul). Estes canais, quando misturados, produzem as demais cores (Ribeiro, 2006).

O RGB é um modelo aditivo, na medida em que suas cores são formadas pela adição das cores primárias, as quais, quando somadas com valor máximo, resultam no branco. A representação deste modelo é dada por um cubo, onde os vértices denotam as cores primárias, as secundárias, o branco e o preto, como mostrado na Figura 2.1.

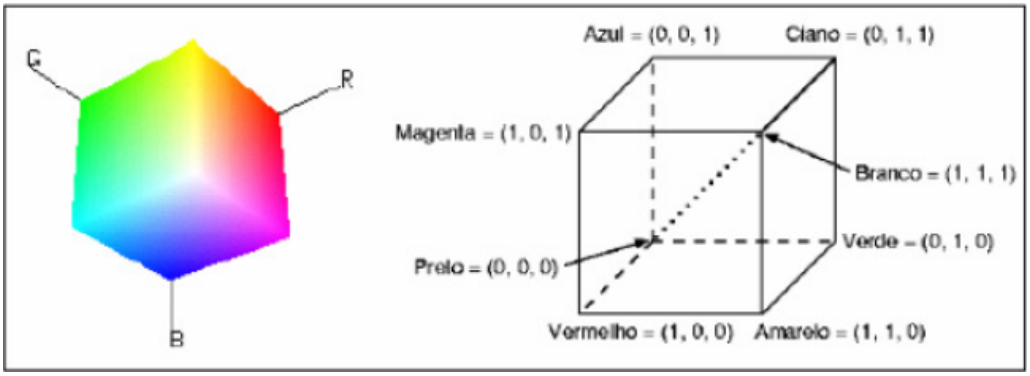


Figura 2.1: Representação em cubo do modelo RGB (Gonzalez and Woods, 2000).

Apesar de bastante utilizado em trabalhos de reconhecimento de pele, como nos propostos por Kovac *et al.* (2003) e Cheddad *et al.* (2009), o espaço baseado no modelo RGB apresenta uma desvantagem para este fim: a formação de cores neste espaço não possui correlação com a percepção humana. Ou seja, cores próximas neste espaço (pouca variação das cores primárias) não estão necessariamente próximos em termos de percepção (Ribeiro, 2006). Desta forma, a estipulação de intervalos baseando-se exclusivamente neste espaço de cores sofre com essa desvantagem.

2.2.1.2 Espaço HSV

O espaço HSV baseia-se, para a formação das cores, nas componentes de *Hue* (matiz) e *Saturation* (saturação). A formação de cores neste espaço se dá de forma intuitiva, utilizando as componentes mencionadas e uma componente chamada *Value* (valor).

O matiz representa a cor em si, definindo a cor dominante. Ele diferencia, por exemplo, o azul do vermelho. Já a saturação mede a pureza da cor, representando o quanto de cor branca está misturada à cor. Ela permite a distinção, por exemplo, do vermelho para o rosa. Por fim, o valor representa a luminância da cor, diferenciando o claro do escuro em cada cor da matiz (Ribeiro, 2006).

O espaço HSV apresenta características desejáveis no tocante à segmentação de pele, tais como o fato das cores serem formadas de maneira intuitiva, além da separação das componentes de crominância e luminância, bastante utilizada nestes algoritmos. Trabalhos como o proposto por Gomez *et al.* (2002) utilizam este espaço, o qual também é abordado e comparado a outros por Zarit *et al.* (1999).

A representação do espaço HSV é dada por um hexacone no qual um ângulo H com relação ao eixo horizontal determina a matiz da cor desejada. Já a distância perpendicular do centro até

a borda determina a saturação, denotada por S. Por fim, o valor V é dado pela distância no eixo vertical. A Figura 2.2 mostra o hexacone representando este espaço.

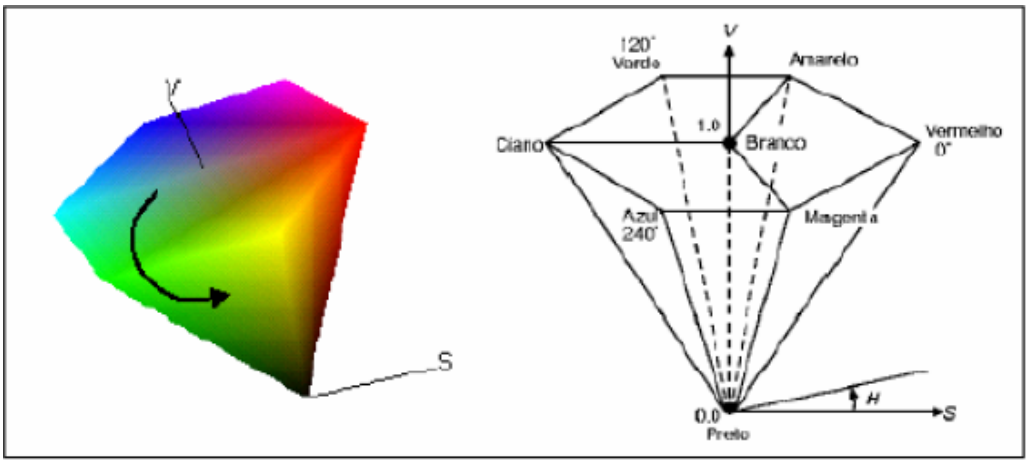


Figura 2.2: Representação em hexacone do espaço HSV (Gonzalez and Woods, 2000).

2.2.1.3 Espaço YCbCr

O espaço YCbCr é um sinal codificado do RGB, comumente utilizado por estúdios de TV europeus e para trabalhos de compressão de imagens (Ribeiro, 2006). Neste espaço, a cor é representada pela luma (Y) e por 2 componentes que correspondem à subtração desta luma das componentes azul e vermelha: Cb e Cr. A partir do espaço RGB, pode-se equacionar este espaço de cores segundo as equações 2.1, 2.2 e 2.3.

$$\boxed{Y = 0.299R + 0.587G + 0.114B} \tag{2.1}$$

$$\boxed{Cr = (R - Y) * 0.713 + 128} \tag{2.2}$$

$$\boxed{Cb = (B - Y) * 0.564 + 128} \tag{2.3}$$

A Figura 2.3 mostra uma representação do espaço YCbCr. Nesta, pode-se observar a variação de luma no eixo Y. Para o valor mínimo de Y, a cor é preta. Já para o valor máximo, esta é branca, independente dos valores de Cb e Cr. Assim como o espaço HSV, este espaço separa as componentes de luminância e crominância, representando uma vantagem no uso em algoritmos de reconhecimento de pele.

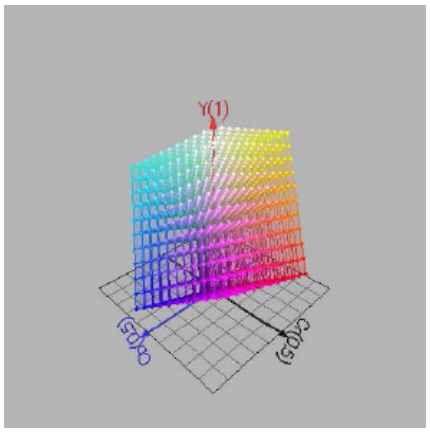


Figura 2.3: Representação em cubo do espaço YCbCr (Ribeiro, 2006).

2.2.2 Algoritmos para o reconhecimento de pele baseados em intervalos

Com o entendimento a respeito dos espaços de cores, pode-se voltar a discussão para os algoritmos de reconhecimento de pele baseados na definição de intervalos. Estes intervalos, dados por valores referentes aos canais dos espaços de cores, foram estipulados com base em diferentes imagens de variados tons de pele e análises a respeito destas. A partir destas análises, pôde-se avaliar quais componentes ou espaços de cores permitem a melhor separação dos pixels que são de pele ou não são, além de permitir a criação de regras que sejam genéricas, capazes de reconhecer corretamente os diferentes tons de pele, como os mostrados na Figura 2.4.

Figura 2.4: Diferentes tons de pele (Cheddad *et al.*, 2009).

O trabalho proposto por Cheddad *et al.* (2009), por exemplo, utiliza as componentes verde e azul do espaço de cores Vermelho, Verde e Azul (RGB) para o cálculo de um valor de luminância. Este valor é utilizado como parâmetro para a definição de um pixel como pele ou não. Para tal,

uma margem foi estipulada a partir de modelos estatísticos e a luminância calculada foi então comparada a ela.

De forma similar ao modelo proposto por Cheddad *et al.* (2009), as abordagens de Kovac *et al.* (2003) e Bhuiyan *et al.* (2003) apresentam margens baseadas em um único espaço de cores. No primeiro, o espaço RGB é usado e as componentes vermelha, verde e azul são associadas a regras definidas empiricamente, as quais variam mediante as condições de iluminação da imagem. Já no último, o espaço YIQ é usado e as componentes de fase e quadratura são usadas para a determinação de um pixel como pele ou não. Na Figura 2.5, imagens referentes ao método de Bhuiyan *et al.* (2003) podem ser vistas, evidenciando que a detecção se deu de forma eficaz.

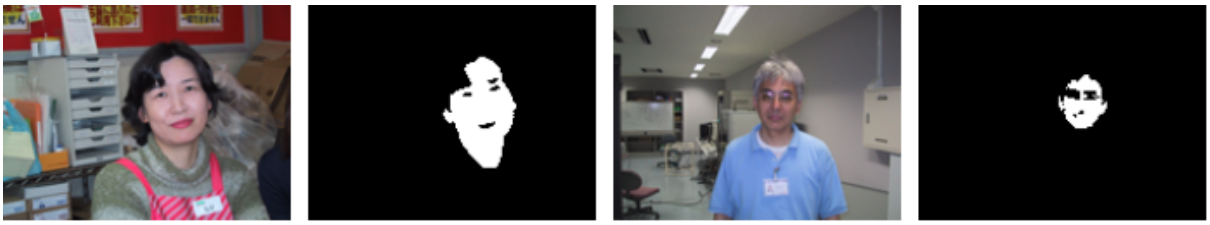


Figura 2.5: Reconhecimento de pele pelo método de (Bhuiyan *et al.*, 2003).

Já os trabalhos de Gomez *et al.* (2002), Tomaschitz and Facon (2009) e Jati and Dominic (2008) utilizam componentes de diferentes espaços de cores para o reconhecimento de pele em imagens. No primeiro, são utilizadas as componentes dos espaços de cores HSV, RGB e YUV, além de um coeficiente determinado empiricamente pelos próprios pesquisadores para a definição de intervalos que classificam os pixels como pele. No segundo, os espaços RGB, HSV e TSL (Tinta, Saturação e Luminância) são usados na formulação de margens para a estipulação de quais pixels correspondem à pele ou não. Por fim, no último, somente as componentes Cb e Cr do espaço YCbCr e a componente H da HSV são aplicadas na elaboração destes valores de margem. Na Figura 2.6, um resultado da aplicação da abordagem de Gomez *et al.* (2002) pode ser visto. Nota-se que o realce dos pixels de pele se deu de forma fiel à imagem.

Figura 2.6: Reconhecimento de pele pelo método de Gomez, Sanchez e Sucar (Gomez *et al.*, 2002).

Por sua vez, o trabalho proposto por Aibinu *et al.* (2012) compara diferentes estudos a respeito do reconhecimento de pele em imagens digitais e propõe a utilização de uma rede neural artificial para realizar a classificação baseando-se nos espaços de cores RGB e YCbCr. Esta abordagem proposta apresentou resultados superiores às demais testadas, com destaque para a rede neural associada ao espaço YCbCr.

2.3 Descritores

Descritores de características, ou somente descritores, correspondem a métodos e/ou técnicas para a obtenção de informação em imagens digitais que permitam a sua representação em algum domínio segundo alguma característica, como formas presentes nestas imagens, cores, texturas, entre outros (Maji, 2005). Além disso, estes apresentam atributos que podem fazer o seu uso desejável em determinados contextos, como por exemplo: invariância a rotação, escala e translação; insensibilidade à iluminação e orientação de objetos nas imagens, entre outras (Maji, 2005).

Os dois descritores utilizados no presente trabalho foram o Histograma de Gradientes Orientados e os Momentos Invariantes de Zernike. A escolha por estes descritores se deu, primeiramente, devido à sua capacidade de prover informação relevante às formas e contornos presentes nas imagens, sendo assim, aplicáveis na presente abordagem. Além disso, estes descritores foram e vêm sendo empregados, com sucesso, em diversos trabalhos voltados ao reconhecimento de objetos em imagens estáticas, como nos de Tian *et al.* (2013), Tsolakidis *et al.* (2014) e Qader *et al.* (2006), fundamentando a opção pelo uso dos mesmos.

2.3.1 Histograma de Gradientes Orientados - HOG

O histograma de gradientes orientados (HOG) é um descritor utilizado em trabalhos onde deseja-se reconhecer objetos em imagens. A ideia que rege o descritor HOG é a de que a aparência local de um objeto, assim como sua forma, podem ser descritas pela distribuição da intensidade dos gradientes e direção das bordas (Gritti *et al.*, 2008).

O método empregado para o uso do HOG é o de divisão da imagem a qual se deseja operar este descritor em regiões menores, chamadas células (Dalal and Triggs, 2005; Gritti *et al.*, 2008). Para cada uma destas células, um histograma 1-D é computado levando-se em consideração as direções e intensidades dos gradientes. Ao fim do processo, estes histogramas 1-D são combinados para que se tenha uma representação de toda a imagem (Gritti *et al.*, 2008).

Com o intuito de melhorar os resultados ao aumentar a invariância da técnica a condições de

iluminação e contraste, um método de normalização é utilizado em muitos trabalhos que empregam o HOG (Dalal and Triggs, 2005; Gritti *et al.*, 2008; Tian *et al.*, 2013). Esta normalização, geralmente baseada nos métodos de normalização L1 e L2, é feita considerando os histogramas locais de células adjacentes. Estas são chamadas de blocos e o tamanho dos blocos, assim como das células, é considerado um parâmetro a ser ajustado no HOG.

Além do tamanho das células e dos blocos e do tipo de normalização a ser empregada, alguns outros parâmetros do HOG precisam ser ajustados. Dalal and Triggs (2005) e Gritti *et al.* (2008) trazem aspectos que precisam ser considerados para o uso deste descritor, como o tipo de máscara usada para o cálculo dos gradientes (máscaras 1-D [-1 0 1], filtros de Sobel e Prewitt, entre outros), o número (9, 18, ...) e orientação de bins (sinalizado [180°] ou não-sinalizado [360°]) e o percentual de sobreposição entre células.

Devido à capacidade de prover informações relativas aos gradientes, o HOG tem sido amplamente usado para fins de reconhecimento de objetos e/ou pessoas em imagens digitais. No trabalho realizado por Dalal and Triggs (2005), um dos mais relevantes com o uso deste descritor, pessoas são detectadas em imagens digitais. Este traz, além do processo de classificação onde o resultado do HOG é associado a um classificador Support Vector Machine (SVM), descrições e comparações quanto a outras técnicas baseadas em gradientes, como as Haar wavelets, e quanto a variações dos parâmetros do próprio HOG.

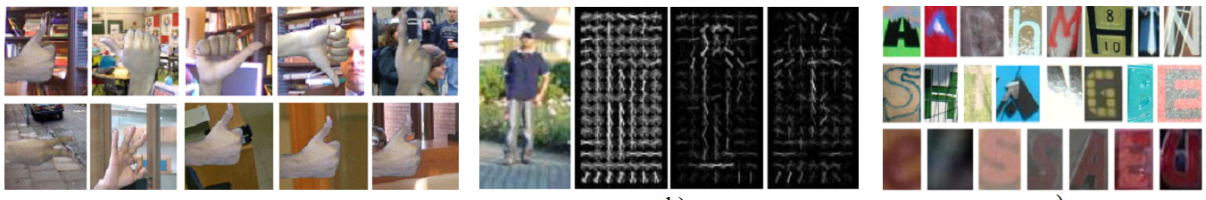
Já no trabalho de Tian *et al.* (2013), uma variação do HOG é utilizada para a identificação de caracteres em imagens digitais. Esta variação, chamada de Co-HOG, apresenta resultados, para o *dataset* utilizado e com a metodologia empregada no trabalho, superiores a outras técnicas utilizadas no reconhecimento de caracteres, obtendo taxas finais de reconhecimento de cerca de 80%.

Outro trabalho em que o HOG é utilizado como descritor é o proposto por Misra *et al.* (2011). Este, assim como o presente trabalho, utiliza o descritor para o reconhecimento de gestos. O ponto alto deste trabalho é a utilização da técnica de regressão de mínimos quadrados parciais (PLS), a qual permite a redução da dimensionalidade do HOG. Neste trabalho, 7 diferentes gestos foram reconhecidos.

Por sua vez, Tsolakidis *et al.* (2014) empregou o HOG, juntamente com os momentos invariantes de Zernike, para a identificação de tipos de folha. Este trabalho, apesar de adotar especificidades em sua metodologia para a identificação das folhas, como a remoção do caule e das pétalas, utilizou os mesmos descritores empregados no presente trabalho. Tsolakidis *et al.* (2014) ressalta a importância do HOG para a identificação de folhas por este se ater a uma característica importante: as marcas das veias na superfície das folhas.

A Figura 2.7 mostra imagens referentes aos *datasets* utilizados e resultados obtidos por Dalal

and Triggs (2005), Tian *et al.* (2013) e Misra *et al.* (2011). Já á Figura 2.8 mostra uma das folhas reconhecidas no trabalho de Tsolakidis *et al.* (2014).



a) b) c)

Figura 2.7: Imagens utilizadas nos trabalhos que aplicam o HOG. a) Dalal and Triggs (2005). b) Tian *et al.* (2013) c) Misra *et al.* (2011).

Figura 2.8: Folha com detalhes realçados do trabalho de Tsolakidis *et al.* (2014)

2.3.2 Momentos Invariantes de Zernike

Os momentos de Zernike são uma classe de momentos ortogonais que vem sendo empregada em termos de representação de imagens (Hse and Newton, 2004). Estes momentos possuem 2 parâmetros: repetição e ordem, os quais relacionam-se à capacidade dos momentos representarem detalhes nas imagens. Momentos de ordem maior, por exemplo, representam melhor detalhes finos, apesar de serem também mais susceptíveis a ruídos (Hse and Newton, 2004).

Para Hwang and Kim (2006), os momentos de Zernike correspondem a um mapeamento de uma imagem em um conjunto de polinômios complexos de Zernike. Estes correspondem a um conjunto de polinômios ortogonais definidos dentro do círculo unitário (Khotanzad and Hong, 1990). Assim, segundo Hse and Newton (2004), o cálculo da magnitude dos momentos de Zernike de uma imagem de ordem 'n' e repetição 'm' é dado por 2.4.

$$A_{nm} = (n+1)/\pi \sum_x \sum_y f(x,y) V_{nm}(x,y), x^2 + y^2 \leq 1 \tag{2.4}$$

onde, 'n' é um inteiro não-negativo e 'm' é um inteiro cujo valor está restrito às condições: n - $|m|$ é par e $|m| \leq n$.

Khotanzad and Hong (1990) demonstraram que a rotação de uma imagem não altera a magnitude dos momentos de Zernike, sendo estes momentos apenas alterados em fase. Assim, a magnitude denotada pela equação 2.4 é utilizada como uma característica invariante à rotação. Outra característica importante destes momentos é que a sua propriedade de ortogonalidade garante que não haja redundância ou sobreposição de informação entre momentos que possuem ordem e repetições diferentes (Hwang and Kim, 2006).

Devido à sua capacidade de prover informações relevantes relativas às formas presentes nas imagens, este descritor é muito utilizado, assim como o HOG, em trabalhos de reconhecimento e classificação envolvendo imagens digitais.

No trabalho de Hse and Newton (2004), por exemplo, os momentos invariantes de Zernike são empregados em uma abordagem voltada para o reconhecimento de caracteres feitos à mão. Devido à invariabilidade de sua magnitude à rotação, este reconhecimento também se dá de forma robusta a este tipo de variação. Para o reconhecimento, os classificadores SVM, Minimum Mean Distance (MMD) e Nearest Neighbors (NN) foram utilizados e comparados, sendo que o trabalho atingiu taxas de reconhecimento de 97%.

Já no trabalho de Qader *et al.* (2006), os momentos de Zernike são utilizados para a identificação de impressões digitais. As características de ortogonalidade, invariância à rotação e robustez à presença de ruídos se mostram adequadas nesta aplicação, a qual utiliza-se do cálculo de distância Euclidiana, baseado em valores obtidos com o uso dos momentos de Zernike, para a realização da identificação das impressões digitais.

No trabalho proposto por Oujaoura *et al.* (2012), os momentos de Zernike são comparados a outras técnicas para o reconhecimento de caracteres isolados do alfabeto árabe, como a transformada de Walsh. Cada descritor utilizado é, separadamente, associado a um classificador neural Perceptron Multicamada e os resultados para cada um deles são computados. Para o *dataset* utilizado no trabalho, os momentos de Zernike permitiram a obtenção das melhores taxas de acerto, chegando a mais de 98%.

Na Figura 2.9 imagens que fazem parte dos *datasets* utilizados nos trabalhos Hse and Newton (2004) e Qader *et al.* (2006) podem ser vistas.

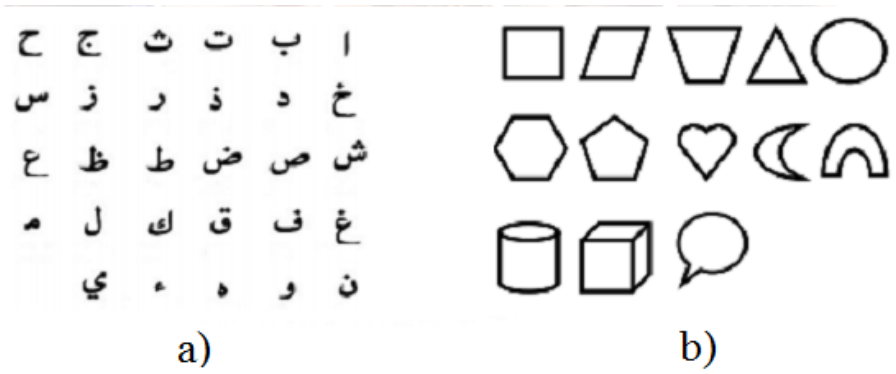


Figura 2.9: Imagens utilizadas nos trabalhos que empregam os momentos de Zernike. a) Hse and Newton (2004). b) Qader *et al.* (2006).

2.4 O classificador Perceptron Multicamada

O classificador Perceptron Multicamada é uma rede neural artificial que vem sendo utilizada em problemas de classificação e vem apresentando sucesso mesmo na resolução de problemas complexos (Haykin, 2012).

Este classificador baseia-se no modelo de rede neural de Perceptron de camada única, proposto por Rosenblatt (1960). Porém, devido às limitações do modelo de Rosenblatt no tocante à resolução de problemas não-lineares, muitos esforços foram dirigidos para a aplicação das redes Perceptron Multicamada e para o desenvolvimento de algoritmos de treinamento (Andersen and Martinez, 2001).

Basicamente, as redes Perceptron Multicamada consistem em conjuntos de nós (unidades sensoriais) dispostos em camadas da rede. Estas representam a camada de entrada, seguida por camadas intermediárias (ocultas), que por sua vez, são seguidas pela camada de saída da rede (Haykin, 2012). Assim, os sinais de entrada são propagados camada por camada até chegar à camada de saída, havendo conexões de uma camada somente com a camada imediatamente anterior e com a imediatamente posterior.

A resposta (saída) de um neurônio é determinada pela ponderação por pesos dos valores de entrada e por uma função de ativação. Considerando, por exemplo, um j-ésimo (jth) neurônio em uma camada qualquer, tem-se que a saída deste é dada por:

$$\boxed{S_j = FA_{th}(U_j)}, onde: \tag{2.5}$$

$$\boxed{U_j = \sum X_i * W_{ij}} \tag{2.6}$$

Nas equações 2.5 e 2.6, o termo S representa a saída do neurônio j. Já o termo FA, representa a função de ativação deste neurônio. As entradas são representadas pela matriz X, enquanto a matriz W representa os pesos associados a estas.

As camadas da rede Perceptron Multicamada podem estar associadas a diferentes funções de ativação. A camada de entrada, por exemplo, está associada a uma função de ativação linear, enquanto nas camadas intermediárias, geralmente, usa-se funções não-lineares para que se tenha a capacidade de distinguir padrões mais complexos.

A Figura 2.10 mostra uma típica rede Perceptron Multicamada. Note que ela apresenta apenas 3 camadas (1 única camada oculta). O sinal de entrada é representado pela variável N-dimensional 'x_p', assim como a saída é representada pela variável M-dimensional 'y_p'. As variáveis 'w_{hi}' e 'w_{oh}' representam as matrizes de pesos que ponderam as conexões entre os neurônios. No caso da matriz 'w_{hi}', esta pondera a conexão dos neurônios da camada de entrada com a camada intermediária e apresenta dimensões NxN_h. Já a matriz 'w_{oh}', pondera as conexões entre os neurônios da camada intermediária e a camada de saída e apresenta dimensões N_hxM.

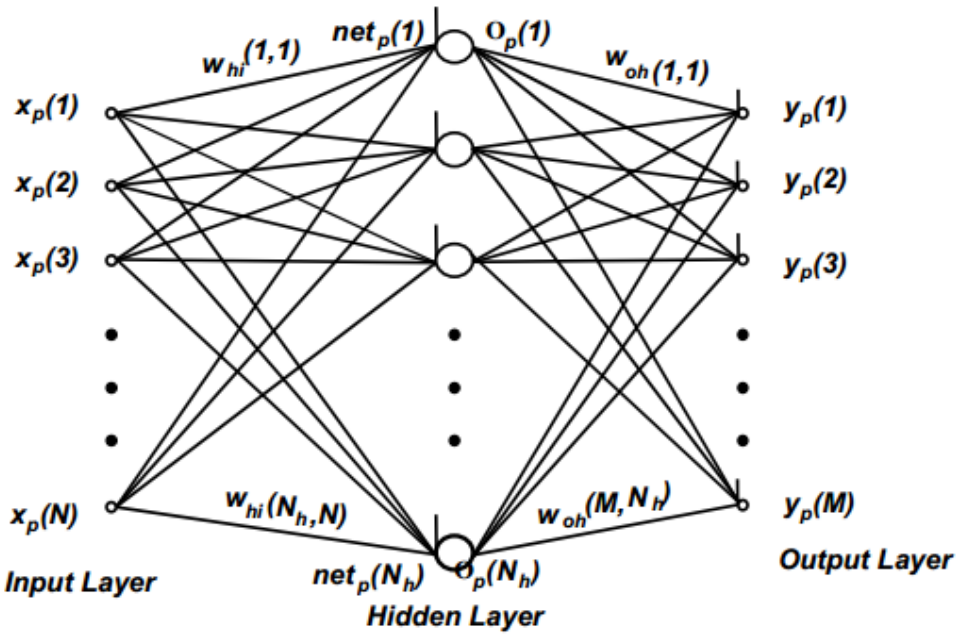


Figura 2.10: Rede Perceptron Multicamada (Delashmit and Manry, 2005).

O método mais comum para o treinamento do classificador Perceptron Multicamada é o aprendizado supervisionado (Braga *et al.*, 2005). Neste tipo de treinamento, as entradas e as saídas desejadas são fornecidas, permitindo que, com base nestes valores, seja feito o ajuste do classificador de forma a relacionar estas entradas às suas saídas correspondentes.

O princípio do aprendizado supervisionado consiste na minimização da diferença entre a saída desejada e a encontrada pela rede para cada entrada do conjunto de treinamento. A minimização da diferença é incremental, proporcionando o ajuste dos pesos da rede neural e, conseqüentemente, levando a uma solução (Braga *et al.*, 2005).

Na Figura 2.11, um pequeno modelo esquemático pode ser visto. Este modelo aborda o funcionamento do aprendizado supervisionado para uma Rede Neural Artificial. Nesta, o Professor representa o conjunto de dados do treinamento e fornece, à rede, representada por RNA, as entradas e saídas desejadas. Desta forma, o treinamento consiste na minimização do erro proveniente da diferença da saída fornecida pelo Professor e da saída fornecida pela rede. Esta minimização é incremental e é feita até que se satisfaça um critério de parada.

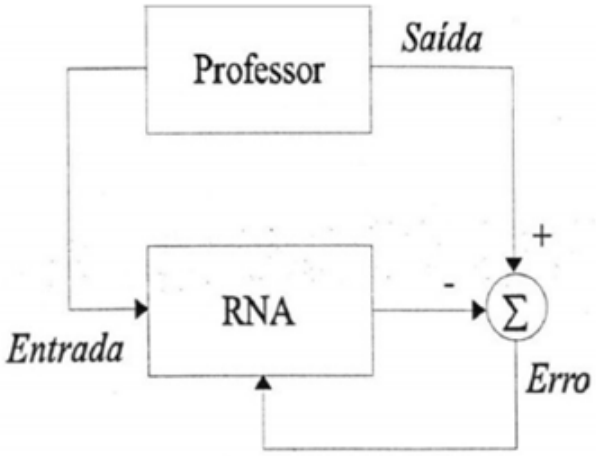


Figura 2.11: Aprendizado Supervisionado (Braga *et al.*, 2005).

Dentre os métodos de aprendizado supervisionado, o Backpropagation é o mais utilizado em aplicações de redes neurais (Johansson *et al.*, 1991). Este realiza a propagação do erro encontrado na camada de saída para as camadas internas da rede neural, buscando minimizar este erro pelo ajuste dos pesos em todas as camadas(Haykin, 2012).

A minimização do valor da função de erro é feita utilizando o método de gradiente descendente. Para tal, avalia-se a direção de variação dos pesos que irá levar à minimização do erro obtido e ajusta-se os pesos nesta direção. Assim, a combinação de pesos que minimiza a função de erro é considerada a solução do processo de treinamento. A Figura 2.12 mostra o passo reverso do backpropagation. Nota-se que o sentido de ajuste dos pesos é feito partindo da camada de saída até que se chegue à camada de entrada da rede.

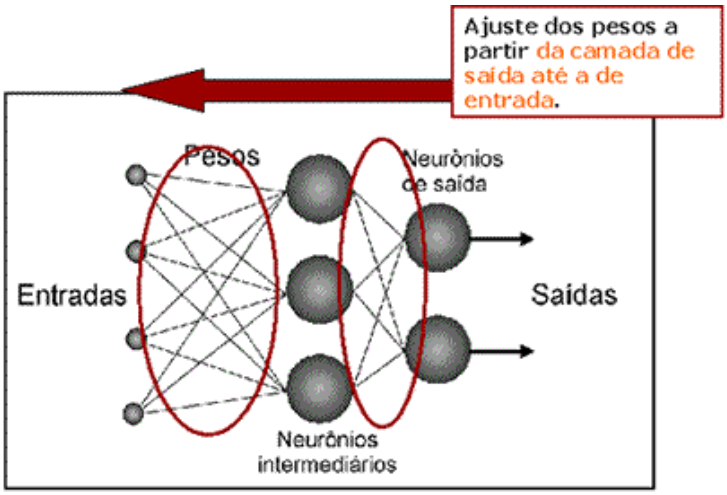


Figura 2.12: Retropropagação no algoritmo de Backpropagation (Barroso, 2014).

2.5 Reconhecimento de gestos

Nos últimos anos, tem-se notado uma crescente no tocante ao surgimento de trabalhos voltados ao reconhecimento de gestos (Khan and Ibraheem, 2012). Estes variam quanto às técnicas e abordagem utilizadas, indo desde de trabalhos que optam pela utilização e/ou auxílio de ferramentas para a detecção das regiões de interesse e de características nas imagens que auxiliem no processo de classificação, até abordagens que usam exclusivamente técnicas de processamento de imagens digitais ou extração de características e classificação, como o presente trabalho. Dentre estes trabalhos, alguns possuem foco no reconhecimento de gestos sem relação com qualquer língua de sinais, enquanto outros utilizam estas línguas como um campo prático para a aplicação das suas abordagens.

2.5.1 Trabalhos que utilizam dispositivos auxiliares

Segundo Sturman and Zeltzer (1994), diversas abordagens são utilizadas com o emprego de dispositivos auxiliares para a detecção e reconhecimento de gestos. Estas abordagens variam, principalmente, quanto ao tipo de ferramenta utilizada e, consequentemente, como a informação a respeito da mão (e sua forma) é obtida. Abordagens que fazem uso de objetos que auxiliam no rastreio das mãos são comuns, tais como as baseadas no uso de luvas e no uso de dispositivos ópticos, como emissores de infravermelho.

O uso de luvas auxiliares para o reconhecimento de gestos merece destaque pela grande quantidade de trabalhos relacionados e pela gama de dispositivos construídos para este propósito.

No trabalho de Parvini *et al.* (2009), por exemplo, foi utilizado o dispositivo CyberGlove para a aquisição de informação a respeito de pontos e articulações das mãos de indivíduos. Para cada gesto realizado com a luva, um vetor de características, chamado de tupla pelos pesquisadores, foi extraído e associado a um classificador Perceptron. Este processo se deu tanto na fase de treinamento quanto de teste. A pesquisa de (Parvini *et al.*, 2009) foi baseada na ASL e obteve, no seu melhor caso, 82,32% de taxa de acerto para 22 sinais da ASL.

Outro trabalho que usa uma abordagem similar é o proposto por Mohandes (2013), o qual também utiliza o dispositivo CyberGlove. Neste trabalho, 100 sinais de duas mãos pertencentes à Língua de Sinais Árabe são utilizados. A abordagem utiliza as técnicas SVM e Principal Component Analysis (PCA) para classificação e extração de características, respectivamente. Ao final do processo, o sistema obteve uma taxa de acerto de 99,6%.

Na Figura 2.13 pode ser vista a luva CyberGlove, utilizada nos trabalhos de Parvini *et al.* (2009) e Mohandes (2013).

Figura 2.13: Luva CyberGlove.

2.5.2 Trabalhos que utilizam exclusivamente técnicas de processamento de imagens e visão computacional

Diferentemente dos trabalhos que utilizam dispositivos auxiliares para assistir no processo de detecção e reconhecimento de gestos, outros, assim como o presente, optam por abordagens que se baseiam exclusivamente no uso de técnicas de processamento de imagens digitais e de visão computacional para este reconhecimento. Nestes, as abordagens são diversas assim como as

técnicas e descritores utilizados.

O trabalho proposto por Stephan and Khudayer (2010), por exemplo, utiliza o contorno de imagens de mãos presentes em imagens como característica para a realização do reconhecimento dos gestos. Apesar de visar reconhecer apenas 6 diferentes gestos, mostrados na Figura 2.14, o trabalho apresenta uma metodologia baseada na realização de etapas que vão desde o processamento das imagens, até a classificação realizada por uma rede neural artificial.

Figura 2.14: Gestos reconhecidos pela abordagem de Stephan and Khudayer (2010).

Já o trabalho proposto por Panwar (2012) traz uma abordagem em que características geométricas, relativas a imagens de mãos, são extraídas e utilizadas para o reconhecimento dos gestos. Estas características são o centróide da mão, a posição do polegar, a orientação da mão, a distância entre os dedos e a presença de picos (pontas dos dedos) nas imagens. Estas características são codificadas em uma palavra binária e esta é utilizada para a definição dos gestos. Este trabalho apresentou uma taxa de acerto de 94% para 45 diferentes gestos.

Outro trabalho bastante relevante é o proposto por Triesch and von der Malsburg (1996). Neste, uma técnica intitulada de Grafos Elásticos é empregada para o reconhecimento de 10 diferentes posturas de mãos. A robustez a *backgrounds* complexos é um atrativo deste trabalho, o qual apresenta uma taxa de acerto de 86,2%. O *dataset* de imagens criado por Triesch and von der Malsburg (1996) contém imagens com *backgrounds* simples e complexos e foi disponibilizado pelos pesquisadores, sendo esta também uma grande contribuição do trabalho. A Figura 2.15 mostra as 10 posturas de mão reconhecidas neste trabalho.

Figura 2.15: Posturas de mãos do *dataset* de Triesch and von der Malsburg (1996)

Dois trabalhos, além dos mencionados anteriormente, merecem destaque por aplicarem sua abordagem em um *dataset* público: o NTU Hand Digit Dataset (Ren *et al.*, 2011). O uso de

um *dataset* público e a disponibilização dos resultados permite a utilização destes valores para comparar com outras técnicas também aplicadas sobre o mesmo *dataset*. Os trabalhos foram os propostos por Zhang *et al.* (2013) e Ren *et al.* (2011), sendo este último responsável pela formulação do *dataset*. No presente trabalho, comparou-se os resultados obtidos nestes trabalhos com os da presente abordagem, utlizando o NTU Dataset como entrada.

A técnica proposta por Zhang *et al.* (2013) corresponde a um descritor intitulado de *Histogram of 3D Facets*, ou H3DF. O uso deste depende de informações relativas à profundidade (mapa de distância) das imagens, normalização da orientação das mãos e geração de um vetor de características correspondente a estas mãos. O trabalho de Zhang *et al.* (2013) foi aplicado no NTU Hand Digit Dataset, como mencionado, e também no ASL Finger Spelling Dataset. Ambos possuem, além das imagens, informações a respeito das distâncias. Interessantemente, este trabalho compara esta técnica com o HOG e com a técnica desenvolvida por Ren *et al.* (2011), intitulada de *Finger-Earth Mover's Distance*, ou FEMD. Nesta, um ponto central da mão é reconhecido e as formas das mãos são representadas por séries de curvas. Estas curvas relacionam-se à distância dos vértices ao ponto central da mão.

Segundo os testes realizados por (Zhang *et al.*, 2013), e considerando-se o NTU Hand Digit Dataset, a técnica de H3DF apresenta taxa de acerto média levemente superior ao HOG (cerca de 2,4%) e também ao FEMD (cerca de 1,6%). Vale lembrar que a comparação realizada considerou o uso do HOG isoladamente, não fazendo composição com algum outro descritor.

Na Figura 2.16, o mapa de distância e o processo de normalização da orientação das mãos, etapas do trabalho de Zhang *et al.* (2013), são representados. Nota-se, através das linhas amarelas, que a orientação dominante das mãos é destacada. Já na Figura 2.17, pode ser vista a representação das mãos através de curvas, feita no trabalho de Ren *et al.* (2011). O ponto azul celeste representa o centro das mãos e os dedos são representados pelas curvas.

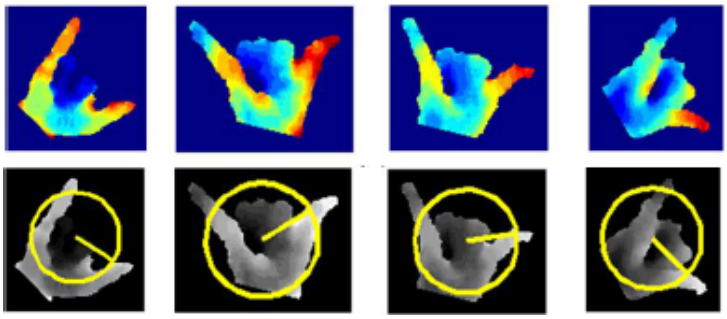


Figura 2.16: Mapa de distância e normalização da orientação das mãos do trabalho de Zhang *et al.* (2013).

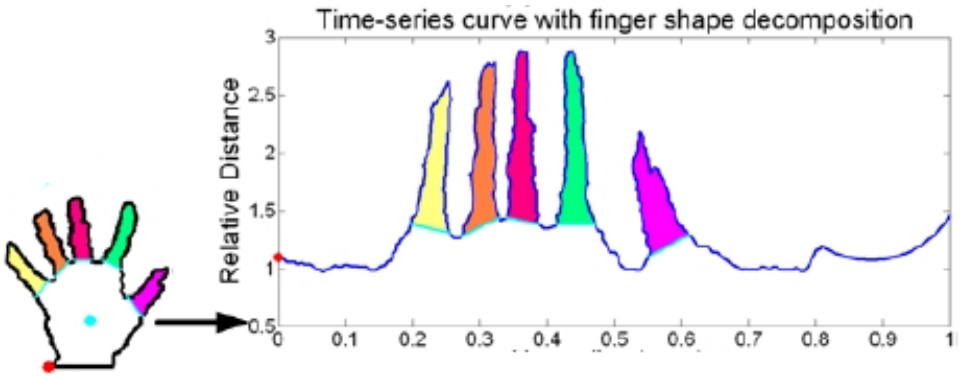


Figura 2.17: Curvas representando os dedos na técnica de Ren *et al.* (2011).

Partindo para o reconhecimento de gestos relacionado às línguas de sinais, tem-se o trabalho proposto por Carneiro *et al.* (2010), o qual apresenta uma abordagem para a classificação de sinais da Libras utilizando redes neurais artificiais. Neste trabalho, sinais com forma similar foram agrupados e a classificação foi dividida em 2 etapas: reconhecimento de qual grupo pertence o sinal de entrada e reconhecimento do sinal. Para tal, foram empregados os Mapas Auto-Organizáveis de Kohonen e a rede Perceptron Multicamada, responsáveis pela atuação na primeira e segunda etapas, respectivamente.

Apesar de também dividir o processo de reconhecimento em 2 etapas, assim como a presente abordagem, Carneiro utiliza os momentos invariantes de Hu como o seu principal descritor. Neste trabalho, foram reconhecidos 26 diferentes sinais da Libras e foram obtidas taxas de reconhecimento que variaram de 78% a 97% para os diferentes sinais. Na Figura 2.18, os gestos do alfabeto da Libras podem ser vistos. Esses foram reconhecidos no trabalho de Carneiro *et al.* (2010).

Figura 2.18: Gestos do alfabeto da Libras reconhecidos por Carneiro *et al.* (2010).

Por sua vez, Otitiano-Rodriguez *et al.* (2012) utilizaram o classificador Support Vector Machine (SVM) para reconhecer 24 sinais da Libras. Para tal, eles utilizaram os momentos invariantes de Hu e de Zernike como descritores, sendo que estes foram utilizados separadamente e os resultados obtidos foram comparados. No trabalho de Rodriguez, o descritor de Zernike apresentou resultados superiores ao de Hu, permitindo a obtenção de taxas de reconhecimento de até 96%.

Já o trabalho proposto por Pizzolato *et al.* (2010) traz uma abordagem similar à do presente trabalho, na medida em que utilizaram, para a classificação, redes neurais dispostas em uma arquitetura de 2 níveis. No entanto, neste trabalho, o segundo estágio de classificação é utilizado como um passo de desambiguação, onde apenas sinais com formas muito parecidas, como os sinais 'F' e 'T' da Libras, são submetidos. Pizzolato *et al.* (2010) usaram os pixels das imagens, depois da aplicação de técnicas de processamento digital de imagens, como características para alimentar os classificadores. Foi obtido neste trabalho uma taxa de reconhecimento de 90,7% para 27 diferentes sinais da Libras.

Dando continuidade ao trabalho de Pizzolato *et al.* (2010), Anjo *et al.* (2012) propuseram a utilização do Kinect como ferramenta para auxiliar para a detecção das regiões das mãos e, utilizando-se de redes neurais artificiais, realizaram o reconhecimento de 10 sinais estáticos da Libras. A abordagem proposta conseguiu reconhecer os sinais com 100% de taxa de acerto. Os 10 sinais reconhecidos neste trabalho são mostrados na Figura 2.19.

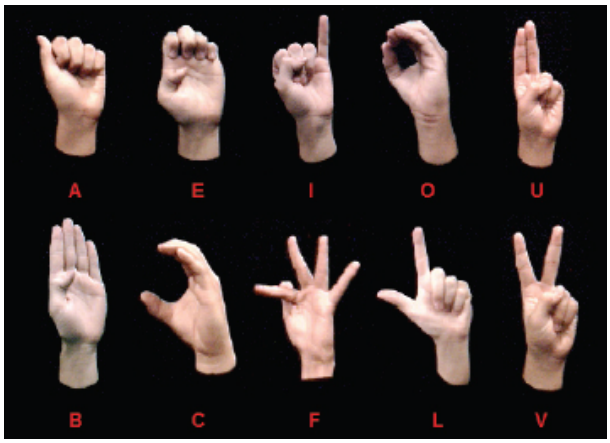


Figura 2.19: Sinais reconhecidos por Anjo, Pizzolato e FeuerStack (Anjo *et al.*, 2012).

Além da existência de trabalhos relacionados à Libras, outros que tratam de diferentes línguas de sinais foram encontrados. Os trabalhos propostos por Liwicki and Everingham (2009) e Bowden *et al.* (2004) visam reconhecer sinais da BSL. No caso do primeiro, após o processo de segmentação das mãos, uma variante do HOG é empregada para a geração de um vetor

de características usado para classificar os gestos. Já no trabalho de Bowden *et al.* (2004), uma abordagem similar à do presente trabalho é empregada, ao se utilizar uma arquitetura de classificador em 2 níveis. Interessantemente, Bowden *et al.* (2004) também extrai características relacionadas à mobilidade dos gestos e trabalha com os modelos ocultos de Markov (HMM) para o reconhecimento de sinais não-estáticos. Na Figura 2.20, o alfabeto da BSL pode ser visto. Este foi utilizado nos trabalhos de Liwicki and Everingham (2009) e Bowden *et al.* (2004).

Figura 2.20: Alfabeto da British Sign Language.

Dentre os trabalhos que visam reconhecer sinais de outras línguas, pode-se citar os propostos por Zahedi *et al.* (2006); Rahman and Afrin (2013); e Uebersax *et al.* (2011), os quais realizam o reconhecimento de sinais da língua de sinais americana (ASL), e o proposto por Singha and Das (2013), o qual realiza o reconhecimento de sinais da língua de sinais indiana (ISL). Estes trabalhos apresentam abordagens distintas da presente no tocante aos descritores utilizados e ao processo de classificação. No entanto, todos atingiram altas taxas de acerto. Algumas imagens utilizadas no trabalho de Singha e Das podem ser vistas na Figura 2.21.

Métodos probabilísticos e estatísticos, como os modelos ocultos de Markov (HMM), são fortemente utilizados em trabalhos de reconhecimento de gestos. Estes são usados, por exemplo, nos trabalhos de Starner and Pentland (1995), Yamato *et al.* (1992) e mesmo no de Bowden *et al.* (2004), mencionado anteriormente.

Foi percebido, através da revisão de literatura realizada, que há uma grande gama de trabalhos e estratégias aplicadas no reconhecimento de gestos em imagens, sendo que muitas destas apresentam taxas de acerto altas, superando os 90%. No entanto, boa parte dos trabalhos emprega suas abordagens em *datasets* próprios, os quais não são disponibilizados posteriormente. Além disso, apesar de ambos os descritores utilizados na presente abordagem, o HOG e o MIZ,

Figura 2.21: Imagens da língua indiana de sinais usadas por Singha and Das (2013)

serem aplicados em trabalhos de reconhecimento, nota-se que apenas o proposto por Tsolakidis *et al.* (2014) utilizou uma combinação de ambos, aplicando-os para o reconhecimento de folhas. Por fim, nota-se que trabalhos como os de Carneiro *et al.* (2010) e Pizzolato *et al.* (2010) apresentaram arquiteturas similares à da presente abordagem, porém, utilizando descritores e classificadores diferentes. Assim, pode-se afirmar que o presente trabalho tem seu diferencial na medida em que combina elementos de pesquisas já realizadas na área de reconhecimento de gestos, as quais apresentaram altas taxas de acerto e que embasaram a opção por cada um dos aspectos escolhidos para a elaboração deste trabalho.

2.6 Teste estatístico de Wilcoxon (*Wilcoxon Signed-Ranks Test*)

O teste estatístico de Wilcoxon corresponde a um procedimento não-paramétrico desenvolvido para avaliar a diferença entre duas medições ou condições onde os dados avaliados são relacionados (Oyeka and Ebuh, 2012). No presente trabalho, este teste foi utilizado na sua versão pareada, na qual considera-se a mediana da diferença entre as amostras utilizadas para propor 2 hipóteses: (1) hipótese nula - as amostras não são significantemente diferentes e, portanto, não se pode inferir que uma apresenta resultados superiores à outra e; (2) hipótese não-nula - as amostras são significantemente diferentes.

O algoritmo que rege o funcionamento do teste pode ser dividido em algumas etapas:

- Para i = 1, . . . , N, onde N corresponde ao tamanho do conjunto de dados a ser analisado,

deve-se calcular $|x_{2i} - x_{1i}|$ e $sgn(x_{2i} - x_{1i})$. Os termos x_{1i} e x_{2i} correspondem às amostras de cada conjunto analisado e 'sgn' refere-se à função sinal.

- Exclui-se os pares em que $|x_{2i} - x_{1i}| = 0$, de forma a reduzir o conjunto de amostras, agora de tamanho N_t.
- Deve-se ordenar os pares dos conjuntos de tamanho N_t, partindo dos menores valores de diferença absoluta para os de maiores, considerando a expressão $|x_{2i} - x_{1i}|$.
- A partir deste ponto, os pares devem ser ranqueados considerando os valores de diferença calculados, começando pelo ranque 1. Pares empatados (considerando a diferença absoluta) recebem valores de ranque que correspondem à média da posição em que ocupam. O ranque para cada par é denotado por R_i.
- Daí, calcula-se a estatística W, dada por: $W = |\sum_i^{N_t}(sgn(x_{2i} - x_{1i}) * R_i)|$.

Com estes procedimentos, considera-se 2 possibilidades baseadas no tamanho do conjunto N_t. Se N_t for maior que 10, calcula-se um z-score dado pela equação 2.7.

$$z - score = \frac{W - 0.5}{\theta w}, \text{onde } \theta w = \sqrt{\frac{N_t * (N_t + 1) * (2 * N_t + 1)}{6}} \tag{2.7}$$

Se o valor do z-score for superior a um valor 'z crítico', então refuta-se a hipótese nula, que afirma que não há diferença significativa entre os conjuntos de amostras. Já se o z-score for inferior, prova-se que esta hipótese se faz verdadeira.

Em contrapartida, se N_t for menor que 10, então o valor de W é comparado a um valor crítico, estabelecido em uma tabela de referência (Richard, 2011).

O teste de Wilcoxon é abordado em trabalhos acadêmicos sendo que em Taheri and Hesamian (2013), por exemplo, uma generalização a respeito do mesmo e alguns exemplos de seu uso são mostrados. No trabalho de Oyeka and Ebuh (2012), uma variação deste teste é apresentada, a qual ataca pontos fracos, como a necessidade de continuidade das amostras. Já no trabalho de Buchwalder and Huber-Eicher (2003), este teste foi utilizado para analisar o comportamento de aves em um ambiente de experimento, onde a agressividade das mesmas foi avaliada e comparada através deste método.

No presente trabalho, o teste de Wilcoxon foi aplicado para comparar os resultados provenientes da variação de aspectos da abordagem, como parâmetros dos descritores ou mudanças a respeito da arquitetura utilizada. Este embasou as escolhas realizadas no decorrer do trabalho, sendo responsável por provar, estatisticamente, a superioridade de amostras em relação a outras.

2.7 Considerações Finais do Capítulo

O crescimento do número de trabalhos que visam realizar o reconhecimento de gestos proporcionou o surgimento de técnicas variadas para este fim. Foi percebido que, a depender da forma como os gestos se apresentam nas imagens e de como são extraídas informações destas, as metologias empregadas variam. Tem-se, por exemplo, métodos de reconhecimento baseado no uso de ferramentas auxiliares, enquanto outros usam exclusivamente técnicas de visão computacional e processamento de imagens digitais.

Foi notado que em diversas abordagens, utilizou-se classificadores para o reconhecimento, com destaque para o SVM e as redes neurais artificias.

Um outro ponto que merece destaque, com base nos trabalho levantados, é o uso das línguas de sinais como um campo prático para a aplicação das abordagens de reconhecimento de gestos. Línguas como a ASL e BSL foram abordadas em muitos trabalhos, além de outras como a Libras e a ISL.

O capítulo 3, a seguir, apresenta a metodologia empregada na presente abordagem. Esta foi embasada nos estudos levantados e está dividida em etapas que vão desde a construção do *dataset* de imagens, até o reconhecimento dos gestos e a validação dos resultados obtidos.

3
Metodologia

A metodologia empregada no presente trabalho foi dividida em etapas. A partir da seleção dos descritores a serem utilizados, escolha esta embasada na teoria a respeito dos mesmos e nos trabalhos relacionados levantados, teve início a criação do *dataset* de imagens. Em seguida, definiu-se como os descritores seriam utilizados e sobre quais imagens eles atuariam. Daí, foi realizado o ajuste dos parâmetros dos descritores e a extração das informações relevantes para a classificação de cada gesto. Estes gestos foram agrupados, em termos de similaridade, a partir do uso da inspeção visual como critério. Por fim, foi definida a arquitetura dos classificadores utilizados na abordagem e realizou-se os testes para a obtenção dos resultados. Estes foram analisados de forma a constatar a validade e robustez da abordagem.

Desta forma, as seções deste capítulo foram elaboradas de forma a apresentar estas etapas, mostradas na Figura 3.1, descrevendo detalhes a respeito de cada uma e as decisões tomadas. Apesar da Figura 3.1 apresentar a metodologia de forma linear, é importante ressaltar que a construção do trabalho não se deu exatamente desta maneira. Mesmo ao se atingir, durante a aplicação da metodologia, etapas posteriores, voltou-se à etapas iniciais para melhorias em pontos específicos.

Figura 3.1: Etapas da metodologia empregada no presente trabalho.

Antes de apresentar a criação do dataset de imagens, primeira etapa a ser aplicada da metodologia, as tecnologias empregadas no presente trabalho, assim como o ambiente no qual

os testes foram executados são mostrados na seção 3.1.

3.1 Tecnologias utilizadas

Para a implementação da presente abordagem, optou-se pela utilização das linguagens de programação Java e Matlab. A opção por ambas se deu pela familiaridade dos pesquisadores com estas e pelo fato de ambas apresentarem poder de construção e fácil acesso à bibliotecas (no caso do Java) suficientes para o desenvolvimento de todo o código.

O uso de 2 bibliotecas associadas à linguagem Java merecem destaque: a Java Advanced Imaging (JAI) e a Neuroph. A primeira oferece suporte ao tratamento de imagens na linguagem Java, com rotinas que permitem desde o acesso rápido aos canais de cada pixel da imagem até outras que implementam diretamente efeitos de processamento sobre as imagens, tais como convolução de kernels. Já a biblioteca Neuroph oferece suporte ao uso de redes neurais artificiais em Java, implementando rotinas de treinamento e modelos de redes neurais, tais como o BackPropagation e a rede Perceptron Multicamada, utilizadas neste trabalho.

Todos os testes realizados foram executados em uma máquina de processador Intel i7 3770, com 12 GB de memória RAM DDR3. O sistema operacional utilizado foi o Windows 8.1 (Compilação 9600) e, apesar de ter-se utilizado linguagem Matlab para a geração de parte do código (aplicação do HoG e geração de vetor de características), executou-se os testes através de rotinas implementadas em linguagem Java.

A figura 3.2 apresenta algumas das tecnologias empregadas no presente trabalho.

(a) (b) (c) (d)

Figura 3.2: Tecnologias utilizadas no presente trabalho. a) Linguagem Java. b) Linguagem Matlab. c) Java Advanced Imaging (JAI). d) Neuroph.

3.2 Criação do dataset de imagens

A primeira etapa de desenvolvimento do presente trabalho foi a criação do *dataset* de imagens contendo os sinais da Libras. Esta criação se deu com o intuito de promover, através da

aplicação dos descritores selecionados, a obtenção de informação para o treinamento e teste dos classificadores utilizados nesta abordagem.

Com base nos trabalhos encontrados que visam reconhecer gestos da Libras, como os de Pizzolato *et al.* (2010), Anjo *et al.* (2012), Souza *et al.* (2012), Bedregal *et al.* (2006), Carneiro (2010), entre outros, foi percebida a dificuldade ao acesso a um *dataset* de imagens da língua, fato este decorrente da não disponibilização dos *datasets* pelos autores. Nestes trabalhos, por exemplo, os pesquisadores envolvidos criaram os seus próprios *datasets*, sobre os quais aplicaram técnicas e colheram seus resultados, porém, não os disponibilizaram para uso por parte de outros pesquisadores.

Assim, devido à inexistência de um *dataset* público de imagens de Libras, optou-se pela criação de um *dataset* de imagens próprio para a abordagem. Para tal, foi realizada a seleção de sinais da língua por 3 especialistas (tradutores da Libras). Estes especialistas trabalharam de forma a eleger sinais em que o parâmetro de configuração da mão, único usado no presente trabalho, fosse suficiente para a identificação do significado. Sinais que apresentam o parâmetro de movimento, como os sinais 'H', 'J' e 'Z', não foram incluídos. Além disso, os sinais correspondentes aos números '3' e '8' também não foram incluídos, devido ao fato de que a postura de mão para execução destes é similar a dos sinais 'W' e 'S', respectivamente. Já o sinal '6' não foi considerado para a formulação do *dataset* devido ao fato de que a sua postura de mão se assemelha ao '9', porém, com a mão rotacionada em relação a este.

Assim, com base neste critério e na opinião dos especialistas, chegou-se a um conjunto de 40 sinais da Libras. Estes correspondem às letras do alfabeto 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'I', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y'. Além destes, foram usados sinais que correspondem aos algarismos '1', '2', '4', '5', '7', e '9', assim como as palavras 'Adulto, 'América', 'Avião', 'Casa', 'Gasolina', 'Identidade', 'Juntos', 'Lei', 'Palavra', 'Pedra', 'Pequeno' e 'Verbo.

O processo de aquisição das imagens se deu com o auxílio dos 3 especialistas em Libras e 2 alunos surdos fluentes na língua. Pediu-se que todos realizassem, em termos de postura da mão, os 40 sinais mencionados. Com a utilização de uma câmera (do modelo Microsoft Lifecam HD-3000), imagens foram continuamente adquiridas chegando a um total de, para cada um dos sinais, 120 imagens de resolução 50x50 pixels. Essas 120 imagens correspondem a subimagens que englobam somente a região das mãos de cada modelo, sendo que estas foram geradas a partir de recorte manual das imagens adquiridas com a câmera. Chegou-se ao fim deste processo a um total de 4800 imagens que foram convertidas de RGB para a escala de cinza. Aspectos como a distância da câmera até os indivíduos que atuaram como modelos, iluminação regular com pouca variação e ausência de backgrounds complexos foram levados em consideração neste

processo de aquisição. A opção pelo número de 120 imagens se deu ao se imaginar que esta quantidade seria suficiente para separar as imagens em conjuntos de ajuste, treinamento e teste do classificador utilizado. Além disso, esta quantidade é entendida como suficiente para retratar possíveis variações que desejou-se representar para cada gesto, como possíveis alterações em termos de iluminação, postura e configuração de mãos dos indivíduos.

O intuito de se utilizar diferentes indivíduos para o processo de criação do *dataset* foi o de prover imagens com variações em termos de tamanho e características das mãos, além de variações em termos da configuração realizada por cada indivíduo. Estes fatores tendem a aumentar a capacidade de generalização dos classificadores utilizados e a robustez do processo de classificação. As diferenças físicas das mãos dos indivíduos (tamanho, disposição de dedos, etc) relacionam-se, também, a heterogeneidade do grupo utilizado como modelo que, apesar de constituído quase exclusivamente por adultos, foi composto por pessoas de diferentes faixas etárias. Os indivíduos são brevemente descritos a seguir:

- **Indivíduo 1** - Adulto do sexo feminino (jovem)
- **Indivíduo 2** - Adulto do sexo feminino (jovem)
- **Indivíduo 3** - Adulto do sexo feminino
- **Indivíduo 4** - Adulto do sexo masculino
- **Indivíduo 5** - Adolescente do sexo masculino

A Figura 3.3, mostra diferenças de postura de mãos realizadas pelos indivíduos para um mesmo sinal (sinal ’9’). Nota-se que, apesar de todos terem realizado a configuração de mão referente ao sinal, há diferenças quanto ao posicionamento dos dedos, rotação e características físicas das mãos.

Figura 3.3: Diferentes posturas de mão para o sinal ’9’.

Apesar de terem sido adquiridas imagens de todos os 5 indivíduos para todos os gestos, estas foram submetidas a um processo de seleção para que se chegasse ao valor de 120 imagens por gesto. Este processo considerou alguns pontos para a exclusão de imagens, como iluminação ruim (excessivamente clara ou escura), despadronização na distância da câmera até a mão (provocada pela aproximação involuntária das mãos em direção à câmera), ruídos no processo de aquisição e posturas de mão incorretas. Após o processo de eliminação das imagens, tem-se que para cada gesto são utilizadas imagens de 3 a 5 indivíduos.

Além das imagens mencionadas, um outro *dataset* auxiliar foi formado com 4800 imagens. Estas são máscaras binárias obtidas com a aplicação de uma abordagem para a detecção de pele em imagens digitais, a qual será descrita na seção 3.2.1. Vale ressaltar que a detecção de pele foi aplicada sobre as imagens originais, antes destas serem submetidas à conversão para a escala de cinza. A Figura 3.4 apresenta alguns sinais pertencentes ao *dataset* e as respectivas máscaras binárias obtidas.

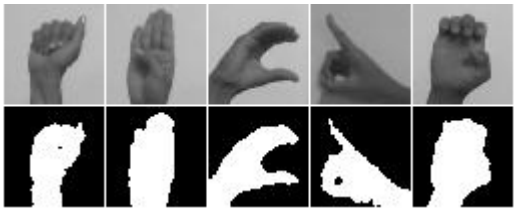


Figura 3.4: Imagens e suas respectivas máscaras binárias.

3.2.1 Abordagem utilizada para a detecção de pele

No processo de criação do *dataset* de imagens utilizado para o treinamento e teste dos classificadores envolvidos na presente abordagem, empregou-se uma técnica de detecção de pixels de pele a fim de desenvolver um *dataset* auxiliar, composto por máscaras binárias que representam as zonas de pele das imagens.

A técnica utilizada combinou as informações de cor (valor dos canais de cor) empregadas em 3 dos mais usados algoritmos para a detecção de pele: O algoritmo proposto por Kovac *et al.* (2003), o proposto por Gomez *et al.* (2002) e o proposto por Bhuiyan *et al.* (2003).

Estes 3 algoritmos utilizam componentes de diferentes espaços de cores para a definição de valores usados como limites para rotular pixels como pele ou não. Os espaços de cores utilizados por estes algoritmos são o YIQ, o RGB e o HSV. O funcionamento de cada um dos algoritmos se dá com o uso de inequações, as quais quando satisfeitas, indicam que um pixel corresponde a uma zona de pele. O algoritmo de Kovac *et al.* (2003), por exemplo, é dado regido pelas inequações 3.1, 3.2 e 3.3. Já as inequações 3.4, 3.5 e 3.6 referem-se ao algoritmo de Gomez *et al.* (2002). Nestas inequações, os termos R, G e B referem-se às componentes vermelha, verde e azul, respectivamente. Elas correspondem aos canais do espaço RGB. Já as componentes H e Y correspondem à matiz e luminância (luma) dos espaços HSV e YUV. O coeficiente Wr foi calculado empiricamente por Gomez *et al.* (2002) e seu valor é dado pela equação 3.7, a qual faz uso de componentes normalizadas do espaço RGB.

$$\boxed{(R > 95), (G < 40), (B > 20)} \tag{3.1}$$

$$\boxed{((max(R,G,B) - min(R,G,B)) > 15, (|R - G|) > 15)} \tag{3.2}$$

$$\boxed{(R > G), (R < B)} \tag{3.3}$$

$$\boxed{(-17.45 < H < 26.66)} \tag{3.4}$$

$$\boxed{(GY < -5.9216)} \tag{3.5}$$

$$\boxed{(Wr < 0.0271)} \tag{3.6}$$

$$\boxed{Wr = (\frac{r}{r+g+b} - 1/3)^2 + (\frac{2}{r+g+b} - 1/3)^2} \tag{3.7}$$

Por fim, tem-se o algoritmo de Bhuiyan *et al.* (2003), dado pelas inequações 3.8 e 3.9. Nestas, as componentes de Y e I do espaço YIQ, correspondentes à luminância e intensidade, são utilizadas.

$$\boxed{(60 < Y < 200)} \tag{3.8}$$

$$\boxed{(20 < I < 50)} \tag{3.9}$$

No presente trabalho, em vez de se utilizar os algoritmos diretamente, elaborou-se uma abordagem usando os componentes dos seus espaços de cores, os quais foram empregados para a formulação de um vetor de características, mostrado na Figura 3.5, e associados a um classificador neural para o reconhecimento de pixels em imagens como pele ou não.

R	G	B	H	S	V	Y	I	Q

Figura 3.5: Vetor de características criado com componentes dos espaços de cores.

Para o treinamento e teste deste classificador, imagens contendo pixels de pele e não-pele

foram utilizadas. Foi criado um *dataset* composto por 243 imagens de pessoas, as quais apresentam diferentes tons de cor de pele e diferentes planos de fundo. Destas 243 imagens, 103 foram adquiridas com o auxílio de uma câmera e possuem resolução de 320x240. As outras 140 foram randomicamente selecionadas dos *datasets* Annotated Skin Database (90 imagens) (Javier and Rodrigo, 2004) e Labelled Faces in the Wild (50 imagens) (Huang *et al.*, 2007) e apresentam resoluções variadas.

As imagens provenientes do *dataset* Annotated Skin Database são disponibilizadas juntamente com máscaras binárias onde os pixels de pele são destacados, sendo estas máscaras utilizadas para informar ao classificador, no processo de treinamento, quais pixels apresentam coloração de pele e quais não apresentam. Como estas máscaras não existiam para as outras imagens, fez-se necessária uma etapa de criação das mesmas, feitas a partir de marcações manuais. Na Figura 3.6, algumas imagens e suas respectivas máscaras binárias podem ser vistas. Nota-se que as marcações foram feitas de forma a salientar as zonas de pele da forma mais fiel possível.

Figura 3.6: Zonas de pele realçadas via marcação manual.

Com as máscaras binárias e a definição do vetor de características, destinou-se parte do *dataset* para o treinamento do classificador (43 imagens das 103 obtidas com a câmera, totalizando em 3302400 pixels) e outra para teste e ajuste de parâmetros.

O classificador usado foi o Perceptron Multicamada, treinado com o algoritmo de Backpropagation e utilizando a função de ativação sigmóide. O critério de parada foi a baixa variação do valor do erro quadrático médio (inferior a 1%). A arquitetura do classificador foi dada com base em testes sobre o conjunto de ajuste, composto por 10 imagens dos dois *datasets* e 10 imagens das 60 restantes entre as 103 adquiridas com a câmera. As demais imagens foram usadas para teste.

O ajuste dos parâmetros se deu variando a arquitetura da rede neural, principalmente no tocante à camada escondida. Esta foi determinada a partir da variação do número de neurônios partindo de 1 até 70 (variando-se de 5 em 5 neurônios). Já a camada de entrada possui 9

neurônios (tamanho do vetor de características) e a saída apenas 1 neurônio, o qual tem seu valor associado à identificação de um pixel como pele ou não. A arquitetura do classificador pode ser vista na Figura 3.7.

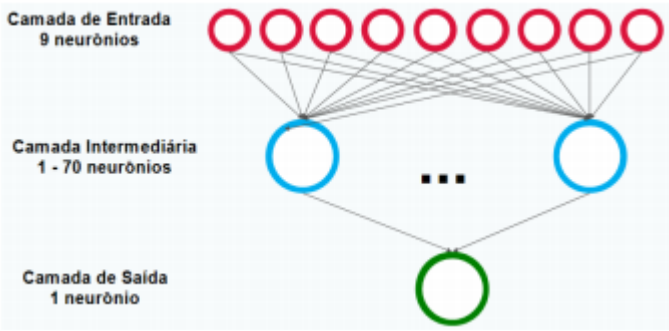


Figura 3.7: Arquitetura do classificador usado para reconhecimento de pele.

Encontrou-se bons resultados com a aplicação da presente abordagem, principalmente para os valores de 35 e 45 neurônios na camada escondida. A Figura 3.8 mostra o realce de pele feito pela abordagem, considerando diferentes números de neurônios na camada escondida. Nota-se que as configurações com 35 e 45 neurônios realçaram de forma fiel as zonas de pele, com presença de poucos falsos positivos e negativos. Já o gráfico apresentado na Figura 3.9 mostra o desempenho de redes com variados números de neurônios na camada oculta, evidenciando a superioridade das arquiteturas de redes neurais (ARN) com 35 e 45 neurônios, representadas pelas cores azul e laranja, sobre as demais. Nota-se que, mesmo que estas tenham apresentado acurácias inferiores para algumas imagens, obteve-se uma média superior quando considera-se todas as imagens do conjunto testado. Estes resultados foram obtidos com base no conjunto de imagens de teste adquiridos pelos próprios pesquisadores. Para estas configurações, encontrou-se uma acurácia média de 86,6% e 86,8%, respectivamente. Já para as configurações de 5, 15 e 25 neurônios ocultos, foram encontrados os valores de 53,1%, 77,7% e 70,8%. Vale ressaltar que para a opção por estas duas arquiteturas, utilizou-se o teste estatístico de Wilcoxon com nível de significância de 0,5%, responsável por comparar as acurácias obtidas com a aplicação em cada imagem. Este indicou a similaridade dos resultados encontrados, em termos estatísticos, para as arquiteturas de 35 e 45 neurônios, os quais se mostraram, também através deste teste, superiores aos obtidos para os demais valores de neurônios.

Uma última comparação foi feita em relação aos algoritmos de pele mencionados. Foi percebido, através dos testes realizados, que a abordagem desenvolvida apresentou resultados estatisticamente superiores aos algoritmos para todos os 3 grupos de imagens de teste: (1) 80 imagens do Annotated Skin Database, (2) 40 imagens do LFW dataset e (3) 50 imagens

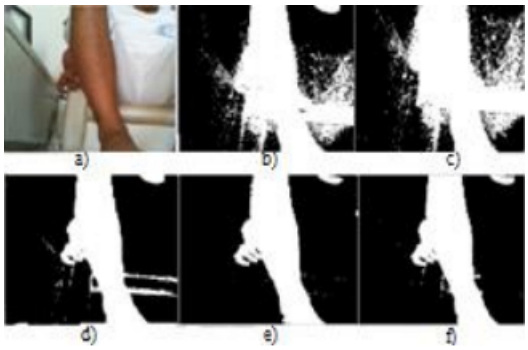


Figura 3.8: Resultados para a classificação com diferentes números de neurônios ocultos. (a) Imagem original. (b) 5 neurônios. (c) 15 neurônios. (d) 25 neurônios. (e) 35 neurônios. (f) 45 neurônios.

adquiridas com a câmera. A Tabela 3.1 mostra as médias de acurácia encontradas para estes testes. Na seção de apêndices, são apresentados gráficos nas Figuras A.1, A.2 e A.3, os quais detalham a acurácia de cada um dos algoritmos para cada imagem usada para o teste. A média dos dados apresentados neste gráficos corresponde aos valores mostrados na Tabela 3.1.

Tabela 3.1: Médias de acurácia encontradas para conjuntos de imagens de pele de teste

Classificador e/ou Algoritmo	Grupo 1	Grupo 2	Grupo 3
Rede 35N	65.2%	67.5%	86.6%
Rede 45N	67.8%	67.1%	86.8%
(Kovac *et al.*, 2003)	57.3%	60.0%	54.2%
Gomez *et al.* (2002)	57.8%	53.9%	65.4%
Bhuiyan *et al.* (2003)	55.1%	59.9%	62.4%

Um ponto a ser ressaltado em relação aos dados da Tabela 3.1 é que estes relacionam-se a *datasets* que apresentam variadas condições de iluminação e complexidade de planos de fundo (*backgrounds*). Estes são fatores indesejáveis para a detecção de pele baseada em aspectos de cor. No entanto, foi percebido que para imagens que apresentavam condições menos adversas, as taxas de acurácia foram muito mais altas, sendo para algumas delas, superiores a 99%.

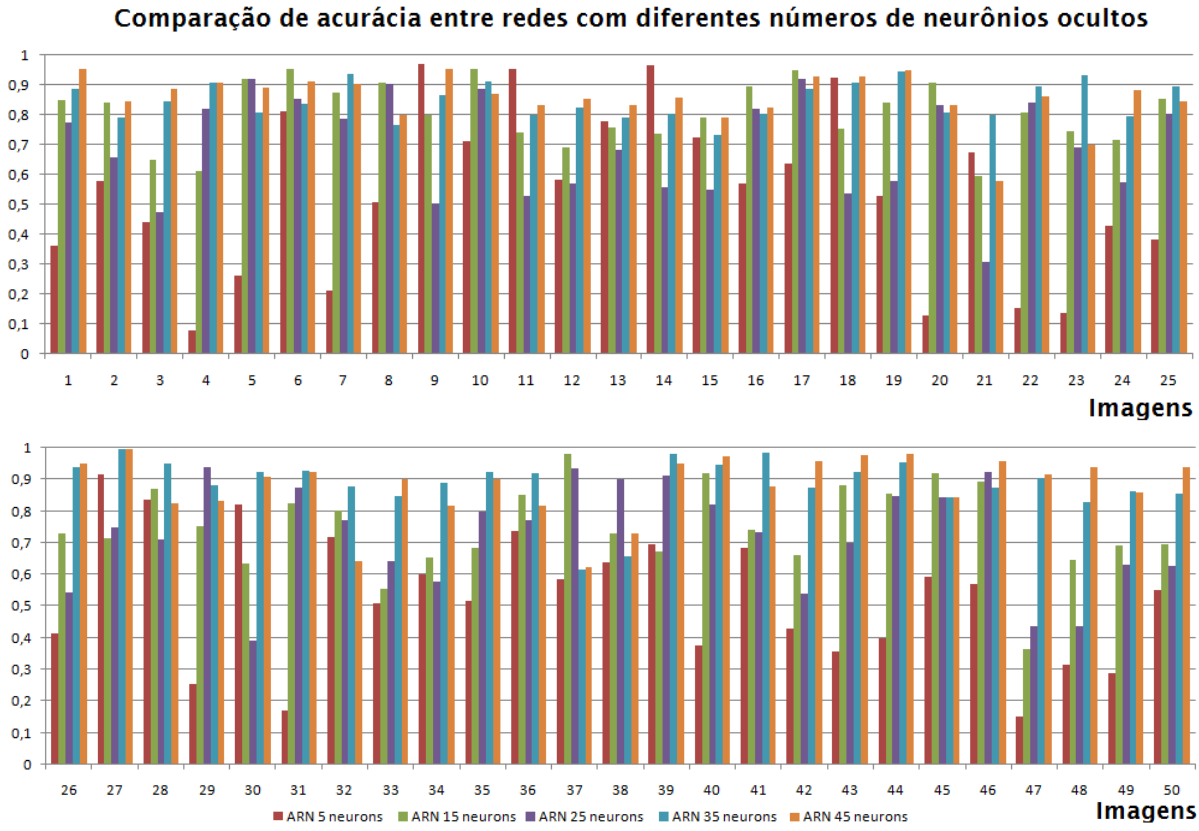


Figura 3.9: Comparação entre o desempenho da abordagem de pele para variados números de neurônios ocultos.

3.3 Descritores HOG e momentos invariantes de Zernike

Com o intuito de extrair características relevantes das imagens e compor um vetor de características que permitisse o reconhecimento dos sinais da Libras, foram utilizados os descritores HOG e momentos invariantes de Zernike.

O HOG, por estar associado aos contornos presentes nas imagens (gradientes), foi aplicado diretamente sobre as imagens em escala de cinza. O resultado da aplicação do HOG é um vetor de alta dimensionalidade, o que reduz o impacto da existência de uma ou outra forma indesejável nas imagens.

Os momentos de Zernike, por sua vez, foram usados nas imagens resultantes da aplicação das máscaras binárias sobre as imagens em escala de cinza. Esta estratégia foi utilizada devido ao fato de que, para cada par de parâmetros utilizado (ordem e repetição), um único valor de amplitude dos momentos de Zernike é encontrado. Assim, a aplicação direta destes momentos sobre as imagens em escala de cinza sofreria influência de qualquer elemento presente no fundo

(background) das imagens, como sombras ou bordas.

Devido a isso e de forma a se obter um valor de Zernike exclusivamente referente à forma da mão presente nas imagens em escala de cinza, optou-se pelo uso desta estratégia. Na Figura 3.10, uma imagem resultante da aplicação das máscaras binárias pode ser vista. Percebe-se que uma sombra presente no background na imagem em escala de cinza (a) é completamente removida na imagem final (c), sobre a qual atua o descritor de Zernike.

(a) (b) (c)

Figura 3.10: Aplicação da máscara binária sobre a imagem em escala de cinza.

3.4 Divisão do dataset: Treinamento, teste e ajuste de parâmetros

Para o ajuste de parâmetros dos descritores HOG e momentos de Zernike, além dos pesos de cada rede neural e o número de neurônios destas, foi empregada uma estratégia de validação cruzada. Nesta, o conjunto de imagens foi organizado em 6 folds, os quais apresentam, para cada gesto, a seguinte disposição: são 120 imagens para cada sinal (desconsiderando as máscaras binárias) em cada fold. Destas, 90 foram utilizadas para o treinamento, 20 para teste e 10 para ajuste dos parâmetros mencionados. A distribuição feita em cada fold é mostrada na Figura 3.11.

Com a definição de como as imagens seriam usadas em cada fold, foi realizado o ajuste dos parâmetros dos classificadores utilizados e dos descritores. Para tal, treinou-se as redes neurais a serem utilizadas e testes foram feitos utilizando o conjunto de ajuste. Estes testes foram executados de forma a selecionar os parâmetros para os quais a taxa de reconhecimento obtida fosse a mais alta. Vale ressaltar que, no caso do ajuste de parâmetros das redes neurais, estes testes foram realizados no mínimo 10 vezes utilizando a mesma configuração de neurônios e incluem o retreinamento destas redes, já que a inicialização dos pesos se deu de forma randômica e esta tem influência nos resultados encontrados.

A distribuição das imagens nos conjuntos de treinamento, teste e ajuste é um ponto que merece destaque, já que esta foi feita de forma com que cada conjunto contivesse imagens da maior quantidade possível de indivíduos e, consequentemente, com mais variadas características

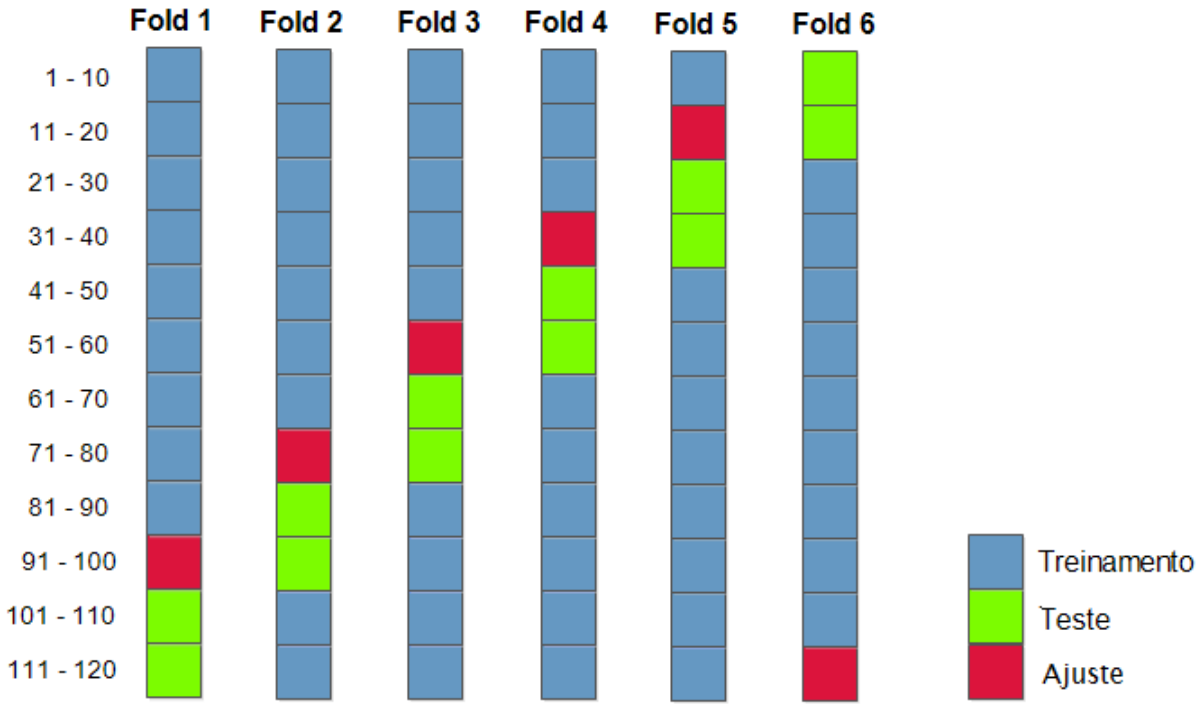


Figura 3.11: Folds e imagens para validação, teste e treinamento.

e posturas de mão. Tentou-se, por exemplo, incluir o maior número de imagens de indivíduos nestes conjuntos de forma a se ter uma distribuição homogênea (todos os indivíduos inclusos e com o mesmo número de imagens). No entanto, como mencionado na seção 3.2, devido à exclusão realizada de algumas imagens, isso não foi possível para todos os sinais, sendo que para alguns, apenas imagens de 3 indivíduos foram efetivamente utilizadas. O mesmo raciocínio é aplicado à divisão de folds, em que tentou-se realizar a divisão de forma homogênea.

3.5 Agrupamento dos sinais

Após a divisão do *dataset* e definição de como as imagens seriam utilizadas, foi realizado o agrupamento dos gestos. Como mencionado, 40 sinais da Libras foram selecionados para serem reconhecidos na presente abordagem. Com o intuito de particionar o processo de classificação e minimizar problemas ligados ao uso de um único classificador para o reconhecimento de uma grande quantidade de padrões, 12 grupos menores foram formados e a tarefa de classificação dividida em 2 estágios: (1) reconhecimento do grupo ao qual pertence o sinal de entrada; (2) reconhecimento do sinal de entrada.

De forma a acompanhar a divisão do processo de classificação em 2 estágios, o classificador

Perceptron utilizado foi disposto em uma arquitetura onde o reconhecimento foi particionado nestes 2 estágios, apresentada na seção 3.7.

Assim, com o uso da classificação em 2 estágios foram obtidos resultados superiores aos adquiridos com o uso de um único, conforme será apresentado na seção 4.3. Além disso, foi percebido que ao se optar por uma arquitetura de 1 estágio, e consequentemente o uso de um único classificador, foi necessário aumentar consideravelmente a quantidade de neurônios escondidos, saindo de redes de 12 a 25 neurônios, na arquitetura de 2 estágios, para uma rede de 75 neurônios. Este acréscimo no número de neurônios ocultos foi essencial para prover, a esta rede, a capacidade de reconhecer 40 padrões distintos, quantidade correspondente ao número de sinais do *dataset* de imagens criado.

A necessidade de incrementar o número de neurônios da rede acarretou em problemas relacionados ao maior custo computacional/tempo para treinamento e classificação, além de dificultar a seleção do número de neurônios, os quais poderiam levar a rede a um processo de memorização dos padrões e perda da capacidade de generalizar (overfitting).

Por estas razões, optou-se pela divisão do processo de classificação em 2 estágios, fazendo necessária a definição da quantidade de grupos e de quais gestos (sinais) fariam parte deles. Como o HOG e os momentos invarintes de Zernike são descritores que extraem informações relativas às formas presentes nas imagens, optou-se pelo agrupamento de sinais que apresentavam formas similares. Este agrupamento se deu utilizando a inspeção visual como único critério. Sinais que apresentam forma muito similar, como as letras 'F' e 'T', foram agrupados com nenhum outro gesto para facilitar a separação entre estes por parte dos classificadores. Desta forma, chegou-se aos 12 grupos mencionados anteriormente, os quais são mostrados na Figura 3.12.

3.6 Parâmetros dos descritores

Tomando como base os testes feitos utilizando o conjunto de imagens de ajuste, chegou-se aos valores de parâmetros para os descritores utilizados. Como mencionado, estes descritores foram o HOG e os momentos invariantes de Zernike, os quais foram aplicados sobre as imagens e as suas saídas foram utilizadas para compor um vetor de características.

Os testes para ajuste dos parâmetros foram realizados considerando a arquitetura de classificação em 2 estágios, descrita na seção 3.7. Nestes testes, com o intuito de levantar resultados, utilizou-se um conjunto de 10 imagens de ajuste. A taxa de acerto foi utilizada como métrica para avaliar o desempenho do classificador para cada conjunto de parâmetros, sendo que escolheu-se

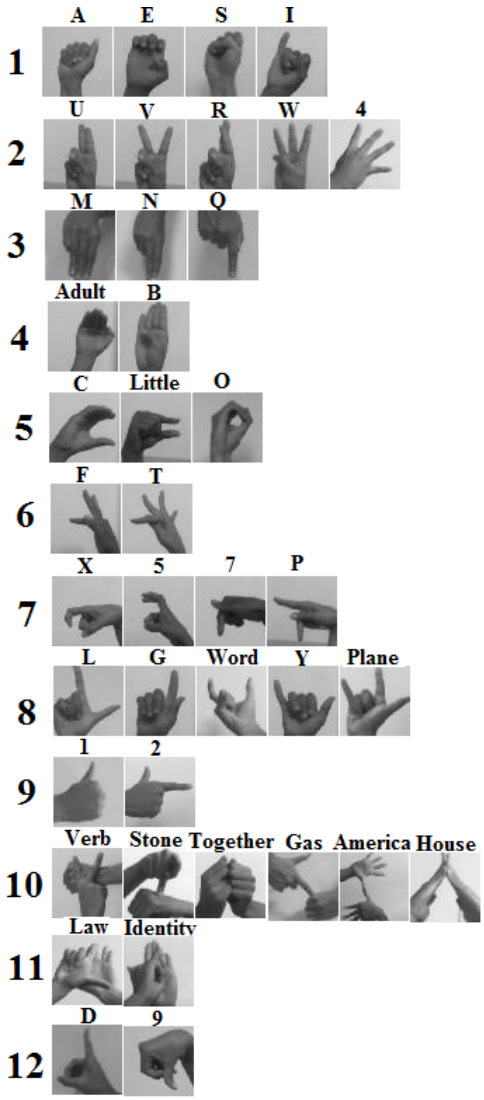


Figura 3.12: Sinais e seus respectivos grupos.

aqueles que promoveram a obtenção das maiores taxas.

Partindo para os parâmetros a serem ajustados, tem-se no HOG o número de bins, tamanho dos blocos e tamanho das células, além do uso de gradientes sinalizados (360°) ou não-sinalizados(180°) e do tipo de normalização a ser feita nos blocos.

Quanto aos bins, testes foram feitos considerando os valores 4, 9 e 18. Já quanto ao tamanho das células, testou-se as dimensões 4x4, 8x8 e 16x16. Os blocos variaram de acordo com o tamanho das células. Devido ao tamanho das imagens, não se pôde variar os blocos de forma muito expressiva, e por isso, apenas considerou-se os valores de bloco de 8x8 e 16x16 para as células de tamanho 4x4, e blocos de dimensões 16x16 para as células 8x8. No caso das

células com valor 16x16, não foi utilizada a divisão da imagem em blocos. Além disso, testou-se as normas L1 e L2. Um último parâmetro selecionado foi a máscara para computação dos gradientes. Optou-se pelo uso da máscara 1-D [-1 0 1], a qual apresentou resultados superiores aos *kernels* de Sobel e Prewitt.

Já nos momentos de Zernike, ajustou-se os valores de ordem e de repetição. Os parâmetros selecionados de ambos os descritores foram aqueles para os quais as taxas de reconhecimento foram as maiores. Estes podem ser vistos na Tabela 3.2.

Pode-se notar que 9 diferentes momentos de Zernike foram utilizados com diferentes pares de parâmetros. Estes foram usados considerando uma variação de ordem, partindo de valores mais baixos até valores mais altos. Os testes realizados envolveram valores de ordem e repetição partindo do 2 até 15, respeitando as restrições mostradas na seção 2.3.2. Esta variação se deu para que os momentos pudessem extrair informações relacionadas a detalhes mais finos da imagem (momentos de alta ordem), até as formas mais grosseiras (momentos de baixa ordem). Além disso, a propriedade de ortogonalidade dos momentos foi levada em consideração, fazendo com que momentos com parâmetros diferentes extraíssem informações das imagens sem sobreposição com outros momentos ou redundância.

Tabela 3.2: Parâmetros selecionados para o HOG e o MIZ

MIZ (ordem e repetição)	HOG (bins, dimensões da célula, dimensões do bloco, sinalizado?, normalização)
(10,4), (9,3), (8,4), (7,3), (6,2), (5,3), (4,2), (3,3), (2,2)	9, 8x8, 16x16, não-sinalizado, L2

As alterações locais de iluminação e distinção entre planos de frente e fundo podem promover uma variação dos gradientes dentro de uma grande gama de valores (Dalal and Triggs, 2005). Devido a isso, foi empregada, conforme apresentado na Tabela 3.2, a norma L2 (norma Euclidiana) sobre os blocos. A opção por esta norma se deu a partir de testes e variação deste parâmetro considerando o conjunto de ajuste.

Um fato que deve ser salientado é que, como o HOG foi aplicado considerando células de tamanho 8x8, conforme mostrado na Tabela 3.2, notou-se que 6 células foram geradas em relação a largura e altura das imagens (as imagens apresentam 50x50 pixels). Com isso, 2 pixels restaram tanto em termos de largura quanto de altura. Assim, decidiu-se excluir estes pixels para a aplicação do HOG. Como os sinais tendem a se apresentar na parte central das imagens, optou-se por excluir o primeiro e último pixels de cada linha e coluna das imagens, resultando em uma imagem de resolução 48x48.

Assim, chegou-se a um vetor de características de tamanho 333, onde 324 dimensões são provenientes da aplicação do HOG (dividiu-se a imagem em 36 células e utilizou-se 9 bins, logo 36*9 = 324) e 9 dimensões são provenientes da aplicação dos momentos invariantes de Zernike, resultando no tamanho final de 333.

3.7 Arquitetura do classificador utilizado

A arquitetura do classificador utilizado na presente abordagem consiste no uso de redes Perceptron Multicamada organizadas em 2 estágios, dos quais o primeiro corresponde a uma classificação geral e o segundo a uma classificação específica do sinal de entrada. As redes de ambos os estágios são alimentadas com a mesma informação (vetor de características proveniente da concatenação do HOG e do MIZ), porém, as redes específicas (pertencentes ao segundo estágio) usam como entrada apenas as informações referentes aos sinais que visam reconhecer.

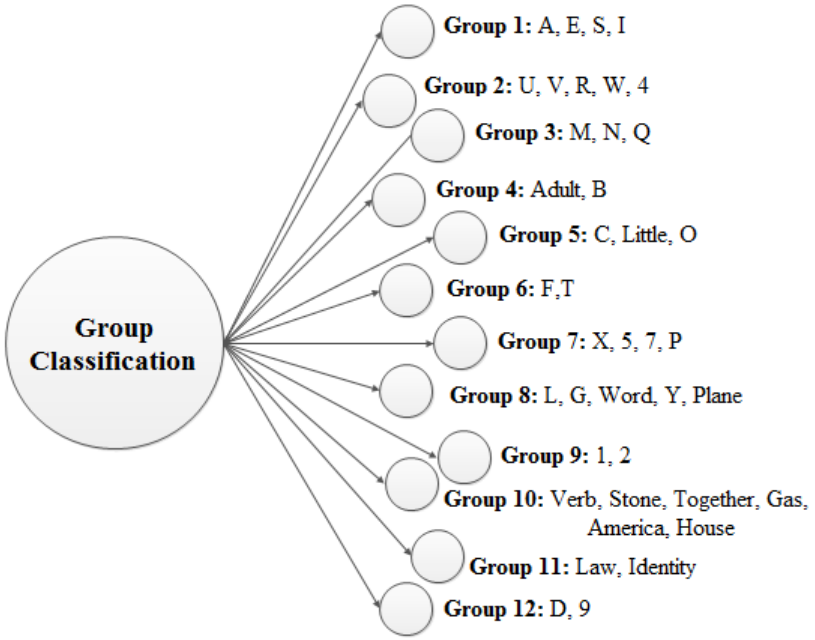


Figura 3.13: Arquitetura do classificador e seus 2 estágios.

O primeiro estágio de classificação é realizado por uma rede chamada de 'Geral'. Esta tem como função reconhecer a qual grupo pertence o sinal de entrada. Para tal, foi realizado o treinamento desta rede com todos os 40 sinais presentes no *dataset* de imagens.

A identificação do grupo ao qual pertence o sinal de entrada permite o direcionamento da classificação para a segunda etapa. Nesta, uma subsequente rede neural é ativada (a depender do

grupo reconhecido na primeira etapa) e esta rede é responsável por identificar o sinal de entrada. Para tal, cada subsequente rede neural é responsável por identificar sinais de um mesmo grupo e, portanto, são treinadas com imagens destes sinais. Estes podem ser vistos na Figura 3.12. Já na Figura 3.13, a arquitetura do classificador é apresentada.

Erros decorrentes de uma classificação errada na primeira etapa direcionam o processo de classificação para redes neurais que não deveriam ser ativadas e, portanto, para estes casos, o reconhecimento é dado como incorreto.

A definição de aspectos da quantidade de neurônios ocultos em cada rede neural da arquitetura foi definida a partir de testes baseados no uso das imagens de ajuste do *dataset*. Como mencionado, estas imagens de ajuste correspondem a 10 imagens diferentes para cada sinal de cada fold e foram também importantes para auxiliar na definição de parâmetros relacionados aos descritores utilizados.

Baseado na teoria a respeito das redes neurais, na dos descritores (em termos de dimensão da resposta) e na definição da arquitetura (número de padrões reconhecidos por cada rede), chegou-se a valores para a disposição das camadas de entrada e saída das redes neurais utilizadas.

A primeira camada de cada rede tem o tamanho do vetor de características resultante da concatenação do HOG com MIZ. A camada de saída, por sua vez, apresenta 1 neurônio distinto para cada padrão a ser reconhecido. A rede geral, por exemplo, é responsável por distinguir 12 padrões diferentes (ela direciona para 12 grupos diferentes) e, portanto, possui 12 neurônios de saída.

Já o número de neurônios da camada intermediária, relacionada à capacidade de reconhecimento de padrões por parte da rede, foi definido exclusivamente através de testes utilizando também as imagens de ajuste mencionadas anteriormente. Para esta definição, dois tipos de ajuste foram feitos: primeiramente um ajuste grosso, correspondente a uma variação de 5 neurônios para mais ou menos. Em seguida, um ajuste fino foi realizado, o qual corresponde a uma variação de 1 neurônio para mais ou menos. As configurações com os melhores resultados, considerando todos os folds, foram escolhidas e podem ser vistas na Tabela 3.3.

3.7.1 Treinamento do Classificador

O treinamento do classificador Perceptron Multicamada utilizado foi feito, como mencionado, utilizando as imagens de 'treinamento' do *dataset* montado. Este processo foi realizado considerando, como método para ajuste dos pesos das redes neurais envolvidas, o algoritmo do Backpropagation.

O critério de parada utilizado foi o valor do Erro Quadrático Médio (EQM), o qual pararia o

Tabela 3.3: Número de neurônios escondidos para cada rede.

Índice do Grupo/Rede	Número de neurônios escondidos
1	20
2	23
3	16
4	15
5	16
6	16
7	22
8	23
9	16
10	22
11	16
12	13
Geral	44

processo de treinamento se fosse atingida uma variação inferior a 1% de uma iteração para a outra. Já a função de ativação empregada foi a sigmóide, dada pela equação 3.10.

$$f(x) = \frac{1}{1 + e^{-\lambda x}} \tag{3.10}$$

O ponto importante no tocante ao treinamento do classificador utilizado é a forma como as saídas foram determinadas e como estas foram tratadas. No treinamento, as saídas desejadas (ideais) foram estabelecidas como 1, para nível lógico alto (saída positiva) e 0, para nível lógico baixo (saída negativa). Com base nisso, para a execução dos testes, estipulou-se margens para avaliar a proximidade das saídas encontradas com estes valores ideais. Foi estabelecido que valores de saída acima de 0,7 são considerados como saídas positivas. Já valores abaixo de 0,3 são considerados como saídas negativas. Os valores que intermediam estas margens são considerados como saídas incorretas.

Ainda em relação ao processo de treinamento, adotou-se uma forma intercalada para a apresentação das diferentes imagens aos classificadores envolvidos. Tomando como exemplo a rede geral, responsável pela distinção de 12 padrões distintos, apresentou-se imagens referentes a cada um destes padrões intercaladamente.

Esta estratégia empregada na apresentação dos dados de treinamento visa evitar o direcionamento (enviesamento) dos classificadores utilizados para o reconhecimento de um ou outro padrão, já que se todos fossem apresentados sequencialmente, a rede tenderia a ter seus pesos

mais 'ajustados' para os últimos padrões apresentados, tendo reduzida assim, a sua capacidade de generalização.

3.8 Coleta de resultados e validação

A última etapa da metodologia foi a coleta dos resultados provenientes da aplicação da abordagem. Assim, utilizando-se das imagens de teste de cada fold do *dataset*, foi calculada a taxa de acerto da presente abordagem, a qual, para este caso, considera o número de sinais reconhecidos corretamente e a quantidade total de imagens.

Com o uso dos valores de acerto, calculou-se as médias entre os folds para cada sinal e para a abordagem, além dos valores de desvio padrão. Estes foram calculados para uma melhor compreensão dos resultados obtidos.

Em seguida, esforços foram voltados para a validação destes resultados encontrados. Esta validação foi feita realizando testes adicionais descritos nas seções 3.8.1, 3.8.2 e 3.8.3, os quais visam comprovar que a abordagem apresenta vantagens em relação a algumas possíveis variantes em termos da arquitetura e apresenta robustez ao reconhecimento de sinais feitos por um indivíduo não-presente no conjunto de treinamento, além de também poder ser aplicada em outros *datasets* de imagens de gestos.

3.8.1 Comparação da arquitetura de 2 estágios com a arquitetura de 1 estágio

Após a coleta de resultados, avaliou-se as vantagens obtidas, em termos de taxa de acerto, na utilização da arquitetura de 2 estágios. Para isso, uma arquitetura de um único estágio também foi utilizada, para qual foram mantidos todos os parâmetros da arquitetura de 2 estágios, tais como os parâmetros dos descritores e dos classificadores. Por utilizar 1 único classificador, foi necessário mudar a quantidade de neurônios escondidos, consistindo no único parâmetro alterado para esta comparação. Este número foi também calculado com o uso das imagens de ajuste do *dataset*. Para este único classificador, obteve-se os melhores resultados para o valor de 61 neurônios escondidos.

Além dos parâmetros serem os mesmos, utilizou-se as mesmas imagens de testes para o cálculo do acerto da arquitetura de estágio único. Esta taxa de acerto, assim como para a arquitetura de 2 estágios, foi calculada para todos os sinais de cada fold. Ao fim, estes valores foram pareados com os provenientes da aplicação da arquitetura de 2 estágios e o teste estatístico de Wilcoxon foi usado para avaliar estatisticamente o resultado desta comparação.

3.8.2 Validação em termos de indivíduos não-treinados

Para a validação da robustez do sistema em termos de usuários não-presentes no conjunto de treinamento, foi criado um *dataset* com imagens de um usuário não utilizado como modelo para compor o *dataset* original. Este *dataset* contém 1600 imagens dos 40 sinais da Libras utilizados na abordagem (20 imagens para cada sinal, além de 20 máscaras binárias associadas a cada imagem).

Vale ressaltar que a criação deste *dataset* se deu de forma similar à do anteriormente criado, levando em consideração parâmetros como as condições de iluminação, distância do indivíduo até a câmera e ausência de planos de fundo (backgrounds) complexos nas imagens. O indivíduo utilizado como modelo é um adulto (jovem) do sexo feminino. Menciona-se que este indivíduo não apresenta fluência em Libras, sendo necessário o auxílio de um especialista para a execução dos sinais que compõem o *dataset*.

A partir disso, realizou-se testes na abordagem utilizando este novo *dataset* de 1600 imagens. As redes utilizadas neste teste foram treinadas com as mesmas imagens de treinamento utilizadas anteriormente e este novo *dataset* foi utilizado exclusivamente como um conjunto de imagens de teste. A taxa de acerto da abordagem foi calculada e utilizada para avaliar os resultados obtidos com este teste.

3.8.3 Validação em termos da aplicação da abordagem em um outro dataset

Por fim, um último teste foi executado para validar a presente abordagem. Este serviu para permitir a comparação dos resultados obtidos com os de outros trabalhos voltados para o reconhecimento de gestos. Para tal, utilizou-se um *dataset* de imagens público: o NTU Hand Digit Dataset (Ren *et al.*, 2011). Este *dataset* foi utilizado em trabalhos como os de Ren *et al.* (2011) e Zhang *et al.* (2013), os quais expuseram os valores de acerto encontrados para as suas abordagem de detecção de gestos. Assim, a aplicação da presente abordagem sobre este *dataset* permitiu avaliar sua robustez ao utilizar um conjunto de imagens de gestos totalmente diferente do criado e usado nos testes. Esta validação também permitiu a realização de uma comparação da abordagem proposta em relação a outras propostas de trabalhos voltados para o reconhecimento de gestos.

3.9 Considerações Finais do Capítulo

A metodologia empregada no presente trabalho partiu da formação do *dataset* de imagens de Libras até o reconhecimento dos sinais e validação dos resultados. Este processo possuiu etapas intermediárias, as quais foram responsáveis pelo ajuste de parâmetros dos descritores e classificadores, segmentação das regiões de interesse das imagens (através da abordagem para detecção de pele) e a montagem da arquitetura para a classificação, organizada de forma a reconhecer os sinais em 2 estágios. As decisões tomadas a respeito da arquitetura e dos parâmetros foram embasadas em testes realizados com imagens de ajuste do *dataset*. Já os resultados finais da abordagem foram obtidos com base nas imagens de teste, sendo estes resultados validados através do teste estatístico de Wilcoxon e de possíveis ameaças à abordagem, como reconhecimento de sinais executados por indivíduos não-presentes na base de imagens de treinamento e aplicação da abordagem utilizando um *dataset* distinto de imagens.

O capítulo 4 apresenta os resultados obtidos, mostrando as taxas de acerto para cada sinal e para toda a abordagem. São apresentados também os resultados encontrados com testes em que variou-se aspectos do processo de classificação. Além disso, discussões e análises a respeito destes resultados são expostas.

4

Resultados e Discussões

Como mencionado nos capítulos anteriores, os resultados da presente abordagem foram calculados em termos da taxa de acerto (sensibilidade), a qual levou em conta a quantidade de acertos do sistema e a quantidade total de imagens testadas. Esta sensibilidade foi calculada para cada sinal e para toda a abordagem, levando em consideração todas as variações de teste que foram realizadas e que as imagens utilizadas para a obtenção destes resultados foram as que correspondem às imagens de teste do *dataset*.

O tempo para a execução das etapas de treinamento dos classificadores e teste também foi calculado e será, primeiramente, apresentado na seção 4.1. Posteriormente, serão apresentadas as taxas de acerto para a presente abordagem.

4.1 Tempo gasto com o treinamento e teste

Considerando a máquina de configuração mencionada na seção 3.1, mensurou-se os tempos gastos para as etapas de treinamento dos classificadores envolvidos e teste da arquitetura.

O treinamento dos classificadores envolve a geração do vetor de características, proveniente da aplicação do HOG e MIZ sobre a base de dados (90 imagens de treinamento por gesto), além do treinamento dos classificadores. Já os testes envolvem a geração do vetor de característica (testes considerando uma imagem por vez) e o processo de classificação.

A Tabela 4.1 apresenta os tempos obtidos. Estes são apresentados em segundos e foram obtidos considerando as tecnologias apresentadas na seção 3.1.

Tabela 4.1: Tempos mensurados para o treinamento e testes da presente abordagem

Etapa	**Tempo Médio (s)**
Geração vetor de características (treinamento) por fold	em torno de 2400
Treinamento de rede geral (primeiro estágio)	em torno 600
Treinamento de redes específicas (segundo estágio)	de 380 a 440
Treinamento de toda a abordagem (geração de vetor de características de todos os folds e treinamento dos classificadores	em torno de 20400
Geração vetor de características (1 imagem para teste)	7
Classificação (considerando a geração de características para 1 imagem)	9

4.2 Arquitetura de 2 estágios

Considerando a arquitetura de 2 estágios apresentada na Metodologia, optou-se inicialmente pelo levantamento dos resultados de cada estágio separadamente, para que possa ser compreendido o impacto que cada um proporciona na classificação.

Assim, obteve-se como primeiro resultado a ser analisado o desempenho das redes neurais específicas, responsáveis pela identificação do sinal dentro de cada grupo (correspondente ao estágio 2 da abordagem). Este primeiro teste foi feito desconsiderando a classificação geral. Nele, é considerado que sempre o reconhecimento por grupos se deu corretamente e as imagens de entrada são direcionadas corretamente para cada subrede.

A Tabela 4.2 traz as médias de taxa de acerto obtidas para cada sinal considerando os 6 folds da abordagem. Pode-se observar que, para todos os sinais, a abordagem apresentou taxas de acerto superiores a 91%, sendo que para 18 dos 40 sinais treinados e testados, a taxa de reconhecimento das redes específicas foi de 100%. A média de acerto encontrada para a classificação específica foi de 98,43% e o valor do desvio-padrão médio foi de 2,58.

Partindo para o primeiro estágio da abordagem, responsável pela identificação dos grupos aos quais os sinais de entrada pertencem, chegou-se aos dados da Tabela 4.3. Nota-se que, apesar da similaridade com os dados mostrados na Tabela 4.2, obteve-se uma queda da taxa de acerto neste estágio superior à obtida no segundo (redes específicas). Para este estágio, encontrou-se um valor médio de 97,00%.

Para toda a abordagem, considerando ambos os estágios de classificação, chegou-se aos

Tabela 4.2: Reconhecimento de sinais nas redes específicas.

Sinal	Acerto médio (%) +/- Desvio-padrão	Sinal	Acerto médio (%) +/- Desvio-padrão
1	100,00 +/- 0,00	Junto	100,00 +/- 0,00
2	99,16 +/- 2,04	L	95,00 +/- 8,36
4	100,00 +/- 0,00	Lei	99,16 +/- 2,04
5	100,00 +/- 0,00	M	96,67 +/- 6,05
7	96,67 +/- 6,05	N	99,16 +/- 2,04
9	100,00 +/- 0,00	O	99,16 + 2,04
A	95,83 +/- 4,91	P	100,00 +/- 0,00
Adulto	97,5 +/- 4,18	Palavra	100,00 +/- 0,00
America	100,00 +/- 0,00	Pedra	100,00 +/- 0,00
Avião	100,00 +/- 0,00	Pequeno	99,16 +/- 2,04
B	91,67 +/- 8,16	Q	98,33 +/- 2,58
C	97,50 +/- 4,18	R	96,67 +/- 2,58
Casa	100,00 +/- 0,00	S	99,16 +/- 2,04
D	100,00 +/- 0,00	T	100,00 +/- 0,00
E	100,00 +/- 0,00	U	92,5 +/- 7,58
F	98,33 +/- 2,58	V	95,83 +/- 5,84
G	100,00 +/- 0,00	Verbo	100,00 +/- 0,00
Gasolina	100,00 +/- 0,00	W	95,00 +/- 4,47
I	98,33 +/- 4,08	X	98,33 +/- 2,58
Identidade	100,00 +/- 0,00	Y	98,33 +/- 4,08

valores dispostos na Tabela 4.4. Nota-se que os resultados encontrados são bastante próximos dos exibidos na Tabela 4.3.

Analisando os dados das Tabelas 4.2, 4.3 e 4.4, percebe-se que a etapa de reconhecimento dos grupos reduziu a sensibilidade do sistema de forma mais impactante que a segunda etapa. Para a maioria dos sinais, as imagens reconhecidas incorretamente na segunda etapa foram também incorretamente reconhecidas no estágio de identificação de grupos. Outro ponto a ser salientado é o fato de que para uma grande gama de sinais, o estágio de grupos atuou como o limitador da taxa de acerto, conforme mostrado na Figura 4.1, a qual traz um comparativo, em termos de taxa de acerto, do desempenho dos estágios 1 (Grupos), 2 (Específicas) e do reconhecimento final da abordagem (Final) para cada sinal. Vale ressaltar que, para no estágio 1, os classificadores atuaram sobre todo o *dataset* enquanto no estágio 2, cada rede atuou sobre, aproximadamente, 1/12 do mesmo.

Mesmo após a etapa de reconhecimento dos grupos, a abordagem apresentou resultados satisfatórios para o reconhecimento dos 40 sinais. Obteve-se como resultado final, dado por média aritmética dos valores de acerto encontrados para todos os sinais, 96,77% e desvio-padrão

Tabela 4.3: Reconhecimento dos grupos aos quais estão associados os sinais de entrada.

Sinal	Acerto médio (%) +/- desvio-padrão	Sinal	Acerto médio (%) +/- desvio-padrão
1	100,00 +/- 0,00	Junto	98,33 +/- 4,08
2	95,83 +/- 3,76	L	98,33 +/- 2,58
4	99,16 +/- 2,04	Lei	100,00 +/- 0,00
5	100,00 +/- 0,00	M	96,67 +/- 6,05
7	98,33 +/- 2,58	N	100,00 +/- 0,00
9	95,83 +/- 4,91	O	96,67 +/- 2,58
A	95,00 +/- 6,32	P	100,00 +/- 0,00
Adulto	95,83 +/- 4,91	Palavra	100,00 +/- 0,00
America	100,00 +/- 0,00	Pedra	99,16 +/- 2,04
Avião	100,00 +/- 0,00	Pequeno	90,83 +/- 9,17
B	86,67 +/- 10,80	Q	98,33 +/- 2,58
C	90,00 +/- 7,07	R	98,33 +/- 2,58
Casa	100,00 +/- 0,00	S	96,67 +/- 5,16
D	96,67 +/- 4,08	T	100,00 +/- 0,00
E	98,33 +/- 2,58	U	90,83 +/- 5,84
F	98,33 +/- 2,58	V	95,83 +/- 5,84
G	99,16 +/- 2,04	Verbo	95,00 +/- 4,47
Gasolina	100,00 +/- 0,00	W	95,00 +/- 4,47
I	89,16 +/- 8,01	X	98,33 +/- 2,58
Identidade	98,33 +/- 4,08	Y	95,00 +/- 8,36

médio de 3,72. Os sinais que apresentaram as mais baixas taxas de reconhecimento foram as letras 'B', com 86,67%, 'I', com 89,16% e 'C', com 90,00%.

Nota-se que, para o caso da letra 'B', o reconhecimento dentro do próprio grupo já apresentou uma taxa de 91,67%, caindo um pouco mais com a inserção da primeira etapa. Já os sinais 'I' e 'C' apresentaram taxas de reconhecimento por parte das redes específicas de 98,33% e 97,50%, respectivamente, evidenciando mais uma vez que no primeiro estágio, em que os grupos aos quais pertencem os sinais são reconhecidos, se deu a maior queda no reconhecimento. O resultado inferior do primeiro estágio da abordagem era, de certa forma, esperado, já que um único classificador foi responsável pela classificação de 40 sinais em 12 padrões distintos.

Além das médias de acerto encontradas para cada sinal, calculou-se também os valores de desvio-padrão considerando os 6 folds da abordagem, também mostrados na tabela 4.4. Estes valores permitem observar se a taxa de acerto calculada para cada sinal em cada fold apresenta uma distribuição em torno do valor médio ou se apresenta uma distribuição mais dispersa.

Observando os dados de desvio-padrão da Tabela 4.4, nota-se que os valores encontrados são baixos, evidenciando uma distribuição pouco dispersa. Os sinais 'Pequeno' e 'B' apresentaram

Tabela 4.4: Reconhecimento final de sinais na arquitetura de 2 estágios.

Sinal	Acerto médio (%) +/- desvio-padrão	Sinal	Acerto médio (%) +/- desvio-padrão
1	100,00 +/- 0,00	Junto	98,33 +/- 4,08
2	95,83 +/- 3,76	L	95,00 +/- 8,36
4	99,16 +/- 2,04	Lei	99,16 +/- 2,04
5	100,00 +/- 0,00	M	96,67 +/- 6,05
7	96,67 +/- 6,05	N	99,16 +/- 2,04
9	95,83 +/- 4,91	O	96,67 +/- 2,58
A	95,00 +/- 6,32	P	100,00 +/- 0,00
Adulto	95,83 +/- 4,91	Palavra	100,00 +/- 0,00
America	100,00 +/- 0,00	Pedra	99,16 +/- 2,04
Avião	100,00 +/- 0,00	Pequeno	90,83 +/- 9,17
B	86,67 +/- 10,80	Q	98,33 +/- 2,58
C	90,00 +/- 7,07	R	96,67 +/- 2,58
Casa	100,00 +/- 0,00	S	96,67 +/- 5,16
D	96,67 +/- 4,08	T	100,00 +/- 0,00
E	98,33 +/- 2,58	U	90,83 +/- 5,84
F	97,50 +/- 4,18	V	95,83 +/- 5,84
G	99,16 +/- 2,04	Verbo	95,00 +/- 4,47
Gasolina	100,00 +/- 0,00	W	95,00 +/- 4,47
I	89,16 +/- 8,01	X	98,33 +/- 2,58
Identidade	98,33 +/- 4,08	Y	95,00 +/- 8,36

os maiores valores: 9,17 e 10,80 respectivamente. Estes valores se devem à medidas um tanto discrepantes de taxas de acerto encontradas em alguns folds. No caso do sinal 'Pequeno', encontrou-se o valor de 80% nos folds 5 e 6, muito inferiores às demais medições. Já no caso do sinal 'B', o valor discrepante encontrado foi de 75%.

Os baixos valores de reconhecimento destes sinais mencionados podem ser associados, em grande parte, com posturas de mão diferentes realizadas pelos modelos que ajudaram na composição do *dataset* e também com a presença de elementos indesejados nas imagens, como sombras. Mesmo que imagens com estes elementos tenham também feito parte do conjunto de treinamento, foi notado que para estas, o classificador tendeu a não se comportar bem. Na Figura 4.2, imagens referentes ao sinal 'C' podem ser vistas. Nota-se a presença de uma sombra no plano de fundo, dada por uma borda existente na parede, a qual impacta sobre o descritor HOG (para o zernike, usa-se as máscaras binárias e, consequentemente, as sombras são desconsideradas).

Utilizando-se os valores de taxa de acerto de cada sinal, realizou-se uma análise do acerto de cada grupo, sendo estas mostradas na Figura 3.12. O acerto de cada grupo foi calculado com a

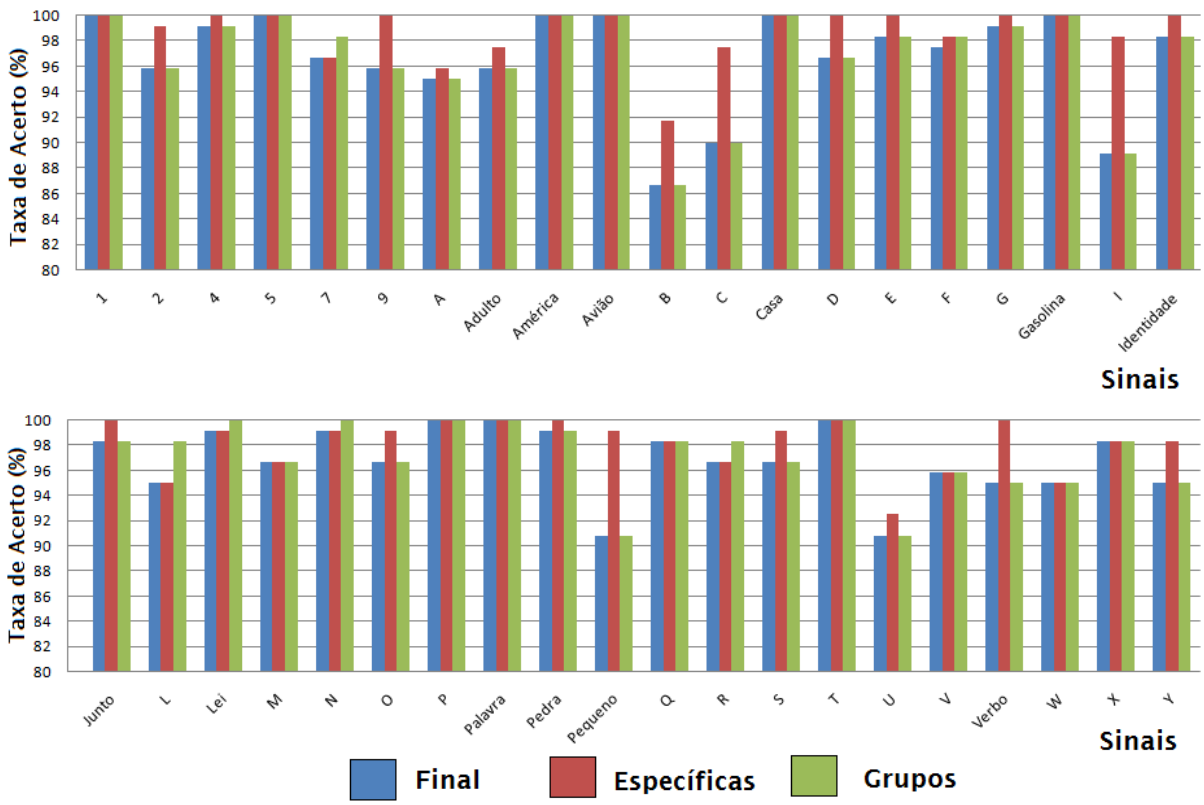


Figura 4.1: Comparação entre resultados dos estágios 1, 2 e do reconhecimento final da abordagem.

Figura 4.2: Elementos indesejados no plano de fundo: Sombras na parede.

média dos valores encontrados para os sinais que o compõem. Por exemplo, para o grupo 1, a taxa de acerto foi dado pela média encontrada paras os sinais 'A', 'E', 'S' e 'I'. A Figura 4.3 mostra a taxa de acerto encontrada para cada grupo. Nota-se que os grupos 4 e 5 apresentaram resultados ligeiramente inferiores aos demais devido aos sinais 'C' e 'Pequeno', pertencentes ao grupo 5 e ao sinal 'B', pertencente ao grupo 4.

Outro fato que pode ser notado é que os sinais feitos em que são usadas duas mãos apresentaram uma taxa de acerto, em geral, muito alta, tendo o seu menor valor com o sinal correspondente à palavra 'Verbo', que foi de 95%. Para todos os outros sinais de 2 mãos, atingiu-se valores superiores a 98%. Este fato pode estar associado às formas destes sinais, bastante diferentes entre si e quando comparadas aos outros sinais, além do comportamento dos descritores, os

quais tendem a extrair informação relativa a estas formas e contornos nas imagens.

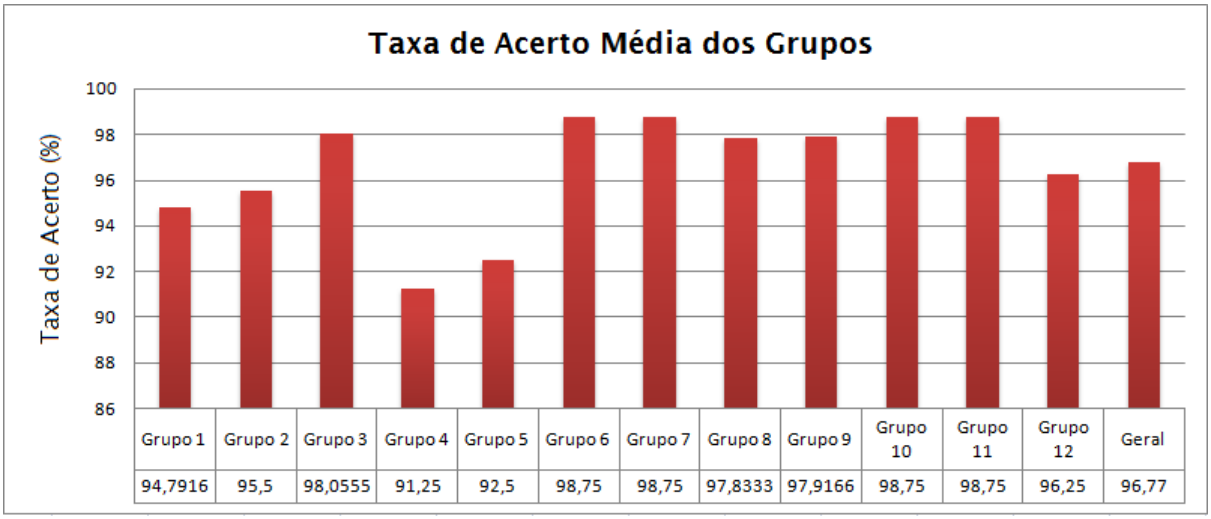


Figura 4.3: Taxa de acerto dos grupos.

Por fim, foi percebido um comportamento bastante interessante para a presente abordagem ao se avaliar as saídas incorretas que foram obtidas. Percebeu-se que, para a grande maioria dos sinais incorretamente reconhecidos (com exceção de 2 imagens da letra 'C' e 2 imagens da palavra 'Pequeno' pertencentes aos folds 1 e 6), encontrou-se saídas que tendiam às corretas, mas que não conseguiam satisfazer os limiares estabelecidos de acima de 0,7 para saída positiva e abaixo de 0,3 para saídas negativas, como mencionado na seção 3.7.1. Assim, foi notado que, para a grande maioria das imagens, os descritores HOG e MIZ proveram informação significativa a respeito dos gestos representados, mesmo que em alguns casos, esta informação não tenha sido suficiente para fazer o classificador atingir os limiares de decisão determinados.

A seção B.1 dos apêndices apresenta os resultados detalhados da arquitetura de 2 estágios, mostrando, através da Tabela B.1, os valores de acerto obtidos para cada sinal em cada fold.

4.3 Comparação com arquitetura de 1 estágio

Além da classificação em 2 estágios, explanada nas seções anteriores, foi testada uma arquitetura de 1 só estágio, em que um único classificador Perceptron Multicamada foi utilizado para o reconhecimento dos sinais. Este classificador foi treinado com os 40 sinais existentes no *dataset* de imagens e a mesma abordagem utilizada para a arquitetura de 2 estágios, com separação de folds e uso de imagens para treinamento, ajuste e testes, foi aplicada.

Além de mantidas as imagens, optou-se pela manutenção de todos os parâmetros dos

descritores, mostrados na Tabela 3.2. Foram mantidos também os parâmetros da rede neural, com exceção do número de neurônios escondidos, o qual foi obtido através de testes e chegou-se ao valor de 61 neurônios.

Os testes para a arquitetura de 1 estágio foram feitos da mesma maneira que os da arquitetura de 2 estágios. Utilizou-se a taxa de acerto como uma métrica para avaliar o reconhecimento por parte da rede. A Tabela 4.5 mostra os resultados encontrados com o uso desta arquitetura. Nota-se que os resultados são estatisticamente inferiores (teste de Wilcoxon com nível de significância de 5%) aos mostrados na Tabela 4.4, na qual os valores referentes à arquitetura de 2 estágios podem ser vistos. A seção de apêndices B.2 traz a Tabela B.2, a qual apresenta os resultados completos obtidos com esta arquitetura.

Tabela 4.5: Reconhecimento de sinais na arquitetura de 1 estágio.

Sinal	Acerto médio (%) +/- Desvio-padrão	Sinal	Acerto médio (%) +/- Desvio-padrão
1	92,5 +/- 16,04	Junto	100,00 +/- 0,00
2	87,5 +/- 14,04	L	91,67 +/- 7,52
4	90,83 +/- 15,62	Lei	95,83 +/- 3,76
5	95,83 +/- 8,01	M	90,83 +/- 9,70
7	92,5 +/- 13,69	N	95,83 +/- 6,64
9	88,33 +/- 8,75	O	75,83 +/- 15,62
A	72,5 +/- 12,94	P	94,16 +/- 6,64
Adulto	91,67 +/- 6,83	Palavra	95,83 +/- 6,64
America	99,16 +/- 2,04	Pedra	97,5 +/- 4,18
Avião	95,00 +/- 6,32	Pequeno	85,83 +/- 12,41
B	87,5 +/- 10.83	Q	93,33 +/- 4,08
C	84,16 +/- 14,97	R	95,00 +/- 3,16
Casa	99,16 +/- 2,04	S	90,00 +/- 8,94
D	94,16 +/- 5,84	T	95,83 +/- 5,84
E	85,83 +/- 8,61	U	79,16 +/- 19,34
F	93,33 +/- 8,16	V	81,66 +/- 13,66
G	99,16 +/- 2,04	Verbo	90,83 +/- 8,61
Gasolina	100,00 +/- 0,00	W	90,83 +/- 10,68
I	91,67 +/- 4,08	X	94,16 +/- 5,84
Identidade	92,5 +/- 8,21	Y	73,33 +/- 21,13

Calculando a média aritmética dos dados mostrados na Tabela 4.5, chegou-se ao valor final de 91,02% de acerto e valor médio de desvio-padrão de 8,59. Analisando estes dados, nota-se que a arquitetura de 1 estágio foi superior à de 2 apenas para os sinais 'Junto', 'B' e 'I'. Apesar do resultado final de 91,02% não ser ruim, foi percebido que para alguns sinais a taxa de acerto foi próxima de 70%, tais como o 'A' e o 'Y', para os quais obteve-se valores de sensibilidade de

72,5% e 73,33%, respectivamente.

Outros pontos merecem ser destacados ao se apresentar estes resultados. O primeiro deles refere-se à quantidade de sinais para os quais obteve-se taxas de acerto de 100%. No caso da arquitetura de 2 estágios, este número foi de 9 sinais. Já para a arquitetura de 1 estágio, este número foi de apenas 2. Isso demonstra que, ao se utilizar de um classificador único para todos os sinais, com exceção de 2 (sinais de 'Gasolina' e 'Junto'), foi encontrado um valor de saída incorreto em ao menos um dos folds. Para uma melhor compreensão das diferenças de acerto obtidas pelos 2 modelos de arquitetura, mostra-se, na Figura 4.4, um gráfico comparando os resultados de ambas para cada sinal. A distância existente entre as barras azul e vermelha representa a diferença da taxa de acerto de uma arquitetura em relação a outra para cada sinal.

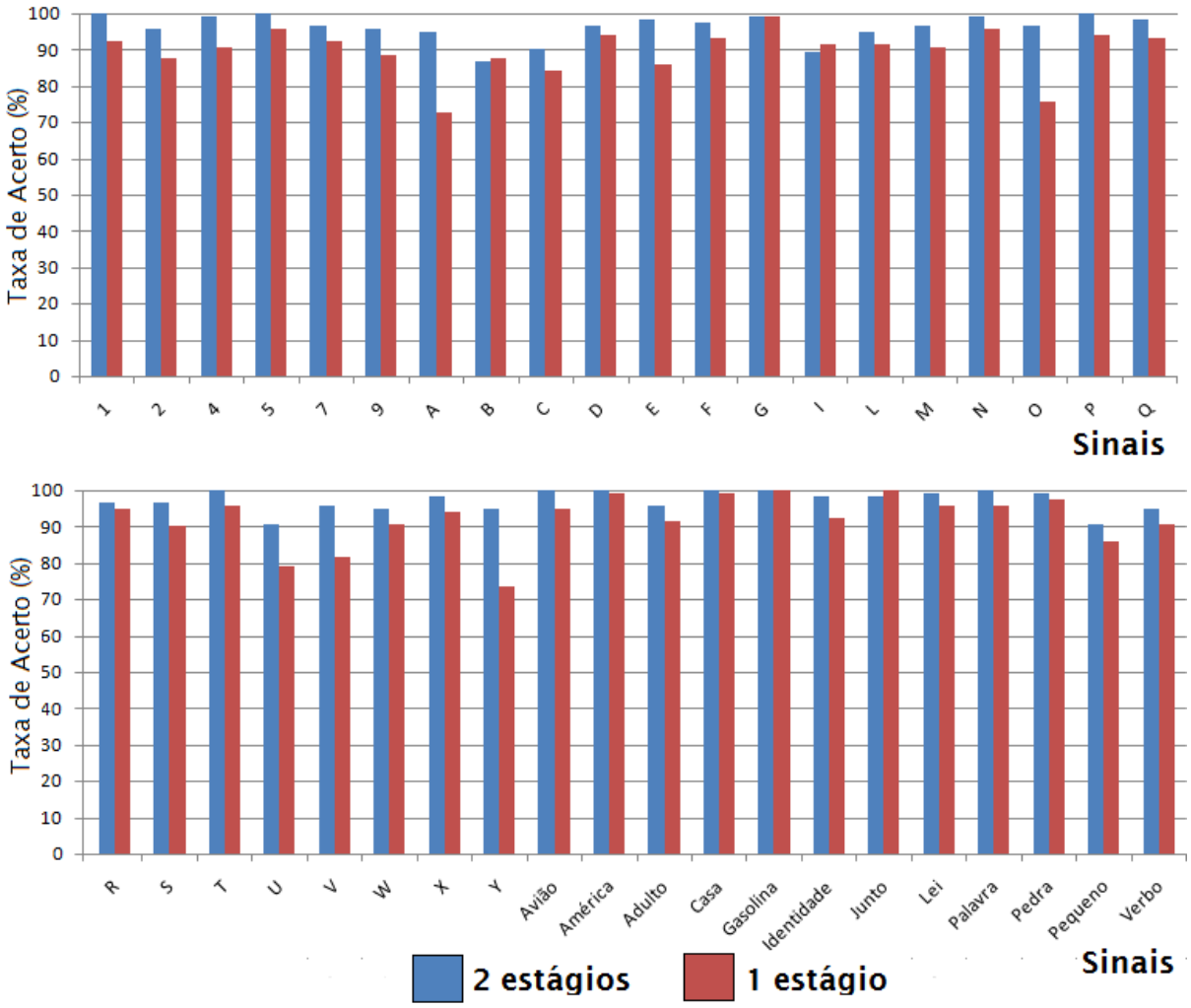


Figura 4.4: Comparação entre resultados do reconhecimento utilizando as arquiteturas de 1 e 2 estágios.

Outro ponto que merece destaque é a obtenção de valores muito baixos de sensibilidade para alguns sinais em alguns folds. No caso da arquitetura de 2 estágios, obteve-se como valor mínimo o referente à classificação do sinal 'B' no fold 2, que foi de 75%. Já no caso da arquitetura de 1 estágio, obteve-se valores baixos em alguns folds, como o valor de 55% de para os sinais 'A' no fold 6, 'C' no fold 1 e 'U'no fold 4, e de 50% para o sinal 'Y' no fold 3.

Devido a estes valores baixos registrados em alguns folds, foi percebido que os valores de desvio-padrão obtidos para a arquitetura de 1 estágio, mostrados na Tabela 4.5 foram muito superiores aos obtidos para a arquitetura de 2 estágios, mostrados na Tabela 4.4. Estes valores estão relacionados a uma maior dispersão das medidas em relação às médias, sendo esta dispersão associada a estes baixos valores encontrados. Para a arquitetura de 1 estágio, encontrou-se, por exemplo, valores de desvio padrão de 21,13 para o sinal 'Y' e 19,34 para o sinal 'U'.

4.4 Utilização de descritores isoladamente

Os resultados mostrados nas Tabelas 4.4 e 4.5, referentes às arquiteturas de 1 e 2 estágios mencionadas anteriormente, foram coletados a partir do uso de ambos os descritores para a extração de informação das imagens e consequente reconhecimento.

No entanto, para um melhor entendimento a respeito da contribuição de cada descritor, além da comprovação de que a combinação destes promove a obtenção de resultados superiores ao uso de ambos isoladamente, realizou-se testes utilizando o HOG e o MIZ de maneira isolada. Para a execução destes testes, utilizou-se a arquitetura de 2 estágios e manteve-se todos os parâmetros relacionados aos descritores e aos classificadores envolvidos (o número de neurônios nas camadas de entrada teve de ser alterado devido à alteração no tamanho do vetor de características).

Assim, refez-se o processo de treinamento da abordagem, a qual passou a considerar apenas a informação de cada classificador isoladamente. As imagens de ajuste foram empregadas somente para adequar os pesos no processo de inicialização das redes neurais, já que os parâmetros dos classificadores e descritores foram mantidos. Por fim, executou-se os testes considerando o conjunto de imagens de teste.

Com a utilização do HOG, encontrou-se uma taxa de acerto média de 94,33% e desvio padrão médio de 5,53. Já para o MIZ, encontrou-se uma taxa de acerto média de 86,62% e desvio padrão médio de 8,61. A Figura 4.5 mostra os resultados encontrados para estes testes comparados ao da arquitetura de 2 estágios (usando o HOG + MIZ) mostrados nas seções anteriores. Nota-se que, para a maior parte dos sinais, mesmo o HOG apresentando resultados próximos aos obtidos com

a utilização do vetor de características composto pelo HOG e MIZ, estes foram inferiores, sendo esta inferioridade estatisticamente comprovada pelo teste de Wilcoxon (nível de significância de 5%).

A seção B.3 dos apêndices apresenta os resultados detalhados da aplicação do HOG e do MIZ isoladamente. Para tal, são apresentadas as Tabelas B.3, B.4, as quais trazem os valores de acerto obtidos para cada sinal em cada fold, além dos valores de desvio-padrão.

Um fato interessante observado é que os resultados obtidos somente com o uso do HOG apresentaram um comportamento similar ao do uso do HOG+MIZ, porém, com taxas de acerto mais baixas. Já com o uso exclusivo do MIZ, notou-se que para alguns sinais em que a abordagem, anteriormente, apresentou altas taxas de acerto, obteve-se taxas baixas com o uso apenas deste descritor, tais como os sinais '1', 'S', 'F' e 'Identidade'. Foi percebido, com o uso apenas do MIZ, que o reconhecimento por parte da abordagem se deu de forma a tender, em alguns casos, a outros sinais com formas similares (já no estágio de classificação dentro do grupo), promovendo erros e diminuição da taxa de acerto total.

Outro ponto a ser destacado é que para os sinais 'Q' e 'Verbo', a abordagem utilizando HOG+MIZ apresentou a mesma taxa de acerto utilizando o somente o HOG, no caso do 'Q' e somente o MIZ, no caso do 'Verbo'. Isso permite inferir que, para estes sinais, poderia se utilizar apenas 1 descritor sem impactar nos resultados obtidos com a abordagem.

Por fim, ressalta-se alto o valor do desvio-padrão encontrado para o uso do MIZ, o qual reflete uma maior heterogeneidade em termos das taxas de acerto obtidas.

4.5 Teste da robustez da abordagem

Com o intuito de validar os resultados encontrados para a presente abordagem, além do uso do teste estatístico de Wilcoxon, foram realizados testes promovendo variações em relação a alguns aspectos. Estas variações e os seus respectivos resultados são usados como uma maneira de garantir a robustez da abordagem. Os testes realizados avaliaram a robustez quanto ao uso de imagens de um indivíduo não-presente no conjunto de treinamento e quanto a utilização de um *dataset* público de reconhecimento de gestos, o NTU Hand Digit Dataset Ren *et al.* (2011). Este último teste foi realizado com o intuito de comparar a presente abordagem com outras técnicas empregadas no reconhecimento de gestos.

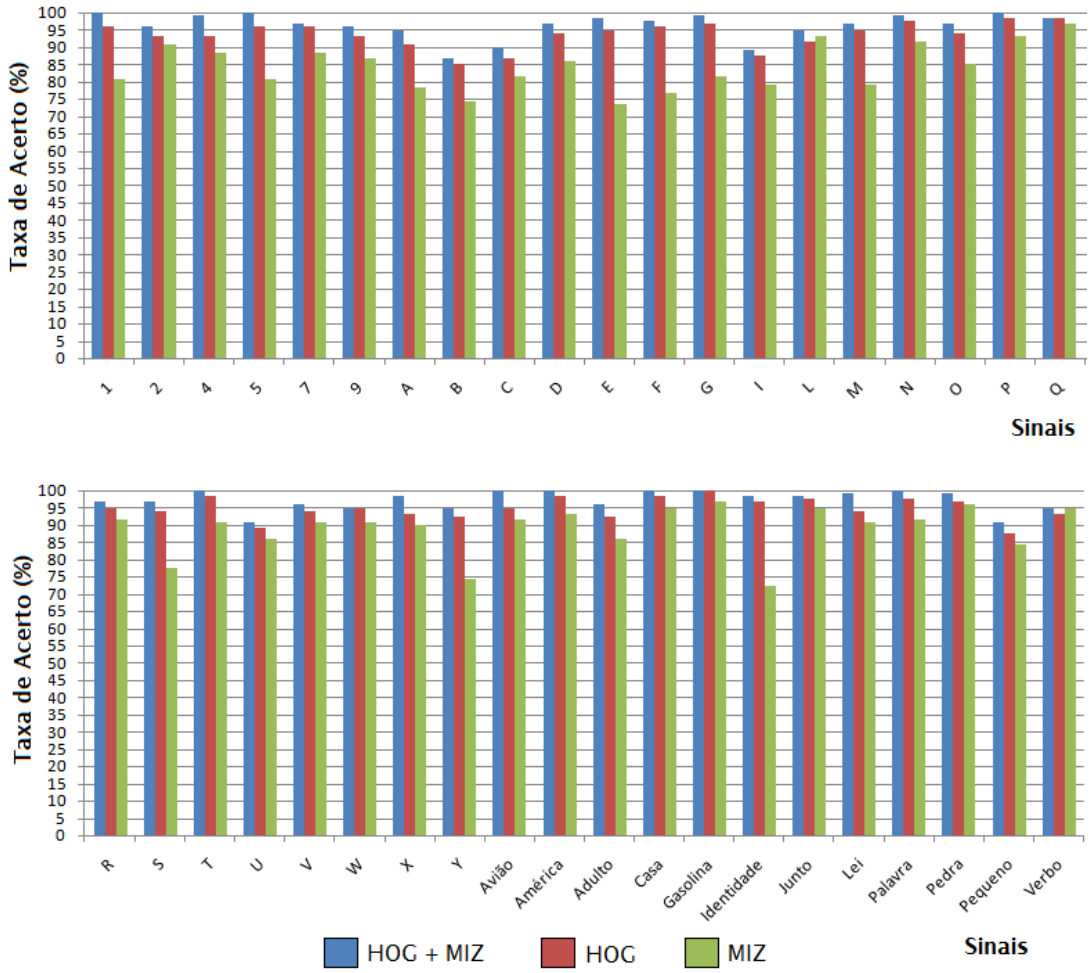
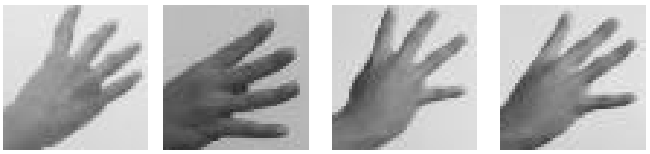


Figura 4.5: Comparação entre resultados da aplicação do HOG+MIZ, HOG isolado e MIZ isolado.

4.5.1 Indivíduo não-presente no dataset

Com o intuito de também testar a robustez da abordagem, foi realizada a aquisição de 800 imagens empregando os mesmos parâmetros para a construção do *dataset* de Libras, incluindo a geração de 800 imagens de máscaras binárias. Estas novas imagens correspondem aos mesmos 40 gestos do *dataset* criado, porém, executados por um modelo (indivíduo) não-presente neste. O novo indivíduo, apesar de ter realizado os mesmos gestos do *dataset* utilizado para o treinamento dos classificadores e validação da abordagem, apresentou posturas de mão com certa diferença em relação aos outros que ajudaram na criação do deste, além de também apresentar diferenças quanto a aspectos como tamanho das mãos, dedos, entre outros.

Na Figura 4.6, podem ser percebidas diferenças entre as posturas realizadas pelo novo modelo (a) e alguns dos indivíduos já presentes no *dataset* (b,c,d).



(a) (b) (c) (d)

Figura 4.6: Diferentes posturas de mão de novo modelo em relação aos demais presentes no dataset

Os testes realizados utilizando as novas imagens demonstraram que a abordagem se comportou bem à presença do novo indivíduo. Foi encontrada, como média de taxa de acerto, 93,5%. É importante mencionar que para este teste, os pesos dos classificadores foram carregados a partir do treinamento feito com o *dataset* padrão (imagens dos outros modelos). Assim, esta classificação foi feita considerando uma pessoa 'não-vista' anteriormente. A Tabela 4.6 mostra os valores de acerto obtidos para cada gesto considerando as imagens do novo indivíduo como entrada. Nota-se que esta tabela não disponibiliza valores de desvio-padrão. Isso se deve ao fato de que apenas utilizou-se 800 imagens para este teste, não sendo feita, portanto, a disposição das imagens em folds.

4.5.2 Utilização do NTU Hand Digit Dataset

O NTU Hand Digit Dataset (Ren *et al.*, 2011) é uma base de imagens adquiridas utilizando o Microsoft Kinect. Para a formação desta base, 10 diferentes individuos foram utilizados como modelos para 10 diferentes gestos, os quais não fazem parte da Libras e não são mencionados como pertencentes a qualquer língua de sinais.

Os modelos realizaram cada gesto 10 vezes (totalizando 1000 imagens), executando pequenas variações nas posturas. A Figura 4.7 mostra um gesto (a) e variações quanto a rotação (b e c) e translação (d) das mãos considerando um único indivíduo.

Além das 1000 imagens dos gestos, o NTU Dataset disponibiliza informações à respeito das distâncias dos objetos em relação à câmera (mapa de profundidade), as quais são relevantes para auxiliar no processo de segmentar a mão dos indivíduos nas imagens (pixels com menores valores de distância).

Utilizando-se dos valores fornecidos pelo mapa de distância do NTU Dataset, pôde-se segmentar as mãos dos indivíduos em cada imagem (em todas as imagens, as mãos estão à frente

Tabela 4.6: Reconhecimento de sinais considerando um indivíduo não-presente no conjunto de treinamento

Sinal	Acerto médio (%)	Sinal	Acerto médio (%)
1	100,00	Junto	100,00
2	100,00	L	80,00
4	90,00	Lei	100,00
5	75,00	M	100,00
7	75,00	N	100,00
9	95,00	O	95,00
A	100,00	P	90,00
Adulto	90,00	Palavra	100,00
America	100,00	Pedra	100,00
Avião	100,00	Pequeno	100,00
B	90,00	Q	95,00
C	80,00	R	95,00
Casa	100,00	S	100,00
D	90,00	T	100,00
E	95,00	U	95,00
F	85,00	V	95,00
G	95,00	Verbo	85,00
Gasolina	95,00	W	95,00
I	75,00	X	95,00
Identidade	95,00	Y	95,00

de qualquer outro objeto). A partir disso, bounding boxes foram utilizadas para compreender os pixels que correspondem às mãos e gerar subimagens de dimensão 120x120 pixels. Este valor foi usado por ser o suficiente para compreender toda a extensão das mãos em todas as imagens. A Figura 4.8 mostra algumas das subimagens geradas. Nota-se que, além dos pixels das mãos, informações referentes ao background também estão presentes. A não-eliminação destes pixels de background foi feita para que a abordagem fosse aplicada da mesma forma que foi aplicada no *dataset* de Libras.

Com as subimagens geradas, aplicou-se a presente abordagem seguindo os mesmos passos realizados com o *dataset* de Libras criado neste trabalho, utilizando apenas um único classificador para este processo.

Assim, primeiramente, gerou-se máscaras binárias utilizando a detecção de pele e converteu-se as imagens para escala de cinza. Assim, chegou-se a 1000 imagens em escala de cinza e 1000 máscaras binárias com pixels de pele realçados, sobre as quais aplicou-se os descritores HOG (diretamente nas imagens em escala de cinza) e momentos invariantes de Zernike (imagens resultantes do uso das máscaras binárias sobre as imagens em escala de cinza). A Tabela 4.7

(a) (b) (c) (d)

Figura 4.7: Variações quanto a rotação e translação no NTU dataset.

Figura 4.8: Subimagens geradas a partir do NTU Dataset.

mostra as médias de taxa de acerto encontradas, além dos valores de desvio-padrão. Vale mencionar que a validação cruzada foi também utilizada neste *dataset*, porém, dividindo-o em 5 folds. Encontrou-se 95,8% como taxa de acerto final da abordagem, dada pela média aritmética dos dados apresentados na Tabela 4.7. Uma apresentação completa dos resultados obtidos para este *dataset* pode ser vista na seção de apêndices C.1.

Tabela 4.7: Acerto e desvio-padrão para gestos do NTU Dataset

Gestos	**Acerto Médio (%) +/- Desvio-padrão**
G1	96 +/- 4,18
G2	95 +/- 7,07
G3	97 +/- 4,47
G4	94 +/- 4,18
G5	96 +/- 4,18
G6	93 +/- 7,58
G7	96 +/- 4,18
G8	94 +/- 5,47
G9	99 +/- 2,23
G10	98 +/- 2,73

Apesar de fatores que influenciam negativamente os resultados da abordagem, como a constante presença de fundos complexos nas imagens em escala de cinza, sobre as quais aplicou-se o HOG, além de existência de pixels de pele nas subimagens que não correspondiam às mãos (alguns indivíduos realizam os gestos próximos do rosto), encontrou-se uma alta taxa de acerto para a abordagem utilizando o NTU Dataset. Este bom resultado encontrado demonstra robustez

da abordagem, a qual apresentou altas taxas de acerto em um *dataset* de reconhecimento de gestos genérico, construído considerando aspectos diferentes do *dataset* montado no presente trabalho. Além disso, nota-se que a abordagem apresentou-se flexível e passível de ser adequada e utilizada em outras aplicações e contextos.

Como o *dataset* NTU Dataset é público, outras técnicas foram aplicadas utilizando-o, como nos trabalhos dos próprios autores do *dataset* (Ren *et al.*, 2011), em que uma técnica intitulada de *Finger- Earth Movers Distance* (FEMD) é empregada; e também no de Zhang *et al.* (2013), em que o HOG é utilizado assim como uma técnica intitulada de *Histogram of 3D Facets* (H3DF). Utilizando os valores disponibilizados pelos trabalhos citados, chegou-se aos dados da Tabela 4.8, a qual mostra que os resultados encontrados para a presente abordagem são similares aos de outras técnicas empregadas.

Tabela 4.8: Média de acerto utilizando o NTU Dataset

Algoritmo (método)	**Acerto Médio (%)**
Nossa abordagem	95,8
HOG (Zhang *et al.*, 2013)	93,1
H3DF (Zhang *et al.*, 2013)	95,5
FEMD (Ren *et al.*, 2011)	93,9

4.6 Considerações Finais do Capítulo

Neste capítulo, foi realizada a coleta de resultados da presente abordagem utilizando a taxa de acerto (sensibilidade) como métrica para avaliar estes resultados. Testes foram realizados considerando possíveis ameaças à robustez da arquitetura, como por exemplo, testes considerando um indivíduo não-presente no treinamento. Para todos os testes realizados, a abordagem apresentou altas taxas de acerto.

Além disso, foi mensurada a sensibilidade da abordagem considerando uma arquitetura de 1 estágio, a qual provou, através dos resultados e da aplicação do teste estatístico de Wilcoxon, ser inferior à de 2. Também foram realizados testes na abordagem utilizando os descritores HOG e MIZ isoladamente, comparando os resultados obtidos com estes aos adquiridos com a combinação de ambos. Notou-se que os 2 descritores combinados fizeram com que a abordagem apresentasse taxas de acerto mais altas, sendo esta comparação também validada pelo teste de Wilcoxon. Por fim, a presente abordagem foi aplicada sobre um *dataset* de imagens público, o NTU Hand Digit Dataset. Mais uma vez, bons resultados foram encontrados, os quais incluem uma comparação da abordagem com técnicas anteriormente utilizadas neste *dataset*.

O capítulo 5, a seguir, apresenta as conclusões a respeito da presente abordagem, levantando possíveis limitações desta e mencionando pontos passíveis de melhora, os quais pretende-se futuramente abordar.

5 Conclusões

Esta dissertação apresentou o desenvolvimento e aplicação de uma abordagem para o reconhecimento de gestos estáticos em imagens, a qual foi empregada no reconhecimento de sinais da Libras.

O desenvolvimento desta abordagem baseou-se nos estudos a respeito de trabalhos voltados para o reconhecimento de gestos. Nestes, foi percebido o uso de uma grande gama de técnicas e estratégias, sendo estas relacionadas à forma como as imagens são adquiridas e como os dados são extraídos destas.

A metodologia empregada na presente dissertação compreendeu etapas que foram desde a construção de um *dataset* de imagens de Libras, necessária devido à indisponibilidade de *datasets* da língua; até as etapas de coleta de resultados e validação dos mesmos. Etapas intermediárias envolveram o uso dos descritores HOG e momentos de Zernike e ajuste dos seus parâmetros, assim como do classificador Perceptron Multicamada, responsável pelo reconhecimento dos sinais.

Ao fim, encontrou-se uma taxa de acerto de 96,77% para o reconhecimento dos sinais da Libras, evidenciando o sucesso da abordagem. Mesmo mediante variações em relação à arquitetura utilizada ou ajuste dos classificadores, este reconhecimento se deu de forma eficaz, alcançando altas taxas de acerto.

5.1 Contribuições

A principal contribuição deste trabalho é a abordagem desenvolvida para a detecção de gestos e sua aplicação no reconhecimento de gestos. Esta utilizou a combinação de 2 descritores de forma amplamente empregados na detecção de objetos em imagens: O HOG e os momentos de Zernike. Além disso, a arquitetura empregada com a disposição dos classificadores em 2

camadas e a consequente divisão do processo de reconhecimento em 2 estágios são marcas desta abordagem. Vale lembrar que as decisões tomadas quanto ao uso dos descritores e a arquitetura adotada basearam-se na teoria a respeito dos mesmos e foram validadas através dos resultados, do teste estatístico de Wilcoxon e de testes envolvendo um *dataset* público.

Como contribuições adicionais, alguns pontos podem ser mencionados. O *dataset* de Libras formado, por exemplo, representa uma contribuição do trabalho, já que foi percebida uma carência neste sentido. Pretende-se, futuramente, disponibilizar este *dataset* para que outros trabalhos possam utilizá-lo, permitindo, inclusive, a comparação de resultados. Além disso, a estratégia utilizada para a segmentação de pele e a revisão de literatura realizada também representam contribuições. Esta revisão, apesar de não seguir formalmente um modelo, levantou trabalhos e técnicas utilizadas para o reconhecimento de gestos e pode auxiliar, futuramente, na elaboração de outros trabalhos.

5.2 Limitações da abordagem

Apesar dos bons resultados encontrados, os quais incluem os testes realizados no NTU Hand Digit Dataset (Ren *et al.*, 2011), a presente abordagem apresenta limitações que relacionam-se às técnicas empregadas. A etapa de geração das máscaras binárias, por exemplo, é apoiada na aplicação de uma estratégia para reconhecimento de pele. Esta estratégia utiliza canais de espaços de cores para reconhecer os pixels de pele, sendo inviável, por exemplo, no caso de se ter imagens originalmente em escala de cinza. Além disso, o reconhecimento de pele por cor apresenta uma forte susceptibilidade a aspectos ambientais, como iluminação e ausência de fundos complexos (neste caso, fundos com cores semelhantes a tons de pele podem ocasionar a presença de falsos positivos).

Já em relação aos descritores HOG e momentos de Zernike, nota-se que estes baseiam-se nas formas presentes nas imagens e, assim como os algoritmos de pele, podem ser influenciados pela presença de backgrounds complexos. Foi notado, por exemplo, que o reconhecimento de sinais em imagens com sombras na parede apresentou uma taxa de acerto muito inferior aos outros.

Além destas, a separação dos grupos de sinais utilizando exclusivamente a inspeção visual como critério pode representar uma limitação da abordagem. O emprego de técnicas para uma melhor separação, em termos de similaridades ligadas às características extraídas (através do HOG e Zernike), poderia melhorar a taxa de acerto do sistema, o qual apresentou maior perda de na taxa de acerto no estágio 1, responsável pelo reconhecimento dos grupos dos sinais.

5.3 Trabalhos Futuros

Pode-se levantar, como trabalhos futuros, alguns pontos em relação à presente abordagem. O primeiro deles, mencionado na seção 5.2 consiste na aplicação de uma técnica para efetuar a separação dos grupos de sinais, deixando de lado o uso exclusivo da inspeção visual para este fim.

Pretende-se também expandir a presente abordagem, de forma a reconhecer uma maior gama de sinais da Libras, incluindo sinais que apresentam movimento. Pensou-se, para este caso, na utilização de modelos estatísticos como os Modelos Ocultos de Markov (HMM). Pretende-se, além do movimento, incluir outros parâmetros da Libras, tais como a orientação das mãos e ponto de articulação, enriquecendo o processo de reconhecimento. Esta expansão da abordagem é seguida pela expansão, também, do *dataset* de imagens, o qual planeja-se disponibilizar como um *dataset* público.

Por fim, projeta-se aplicar a presente abordagem em outros *datasets* públicos de gestos, além do NTU Dataset. Esta aplicação permite uma melhor comparação do presente trabalho com as técnicas utilizadas para o reconhecimento de gestos.

Referências

Aibinu, A. M., Shafie, A. A., and Salami, M. J. E. (2012). Performance analysis of ann based ycbcr skin detection algorithm. In *Proceedings of International Symposium on Robotics and Intelligent Sensors*, pages 1183–1189.

Almeida, M. P. and Almeida, M. E. (2012). História de libras: Característica e sua estrutura. In *Anais da VII JNLFLP*, pages 315–327.

Andersen, T. L. and Martinez, T. R. (2001). Dmp3: A dynamic multilayer perceptron construction algorithm. *Int. J. Neural Syst.*, pages 145–165.

Anjo, M. d. S., Pizzolato, E. B., and Feuerstack, S. (2012). A real-time system to recognize static gestures of brazilian sign language (libras) alphabet using kinect. In *Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems*, IHC '12, pages 259–268, Porto Alegre, Brazil, Brazil. Brazilian Computer Society.

Barroso, R. (2014). O algoritmo backpropagation.

Bedregal, B. C., da Rocha Costa, A. C., and Dimuro, G. P. (2006). Fuzzy rule-based hand gesture recognition. In M. Bramer, editor, *IFIP AI*, volume 217 of *IFIP*, pages 285–294. Springer.

Bhuiyan, A.-A., Ampornaramveth, V., Muto, S.-y., and Ueno, H. (2003). Face detection and facial feature localization for human-machine interface. In *NII Journal*, volume 5.

Bowden, R., Windridge, D., Kadir, T., Zisserman, A., and Brady, M. (2004). A linguistic feature vector for the visual interpretation of sign language. In T. Pajdla and J. Matas, editors, *ECCV (1)*, volume 3021 of *Lecture Notes in Computer Science*, pages 390–401. Springer.

Braga, A., Carvalho, A., and Ludermir, T. (2005). *Redes Neurais Artificiais: Teoria e Aplicações*. LTC.

Brito, L. F. (2010). *A Língua Brasileira de Sinais*. Tempo Brasileiro: UFRJ, Departamento de Lingüística e Filologia, Rio de Janeiro.

Buchwalder, T. and Huber-Eicher, B. (2003). A brief report on agressive interactions within and between groups of domestic turkeys (meleagris gallopavo). *Animal Behaviour Science*, **84**, 75–80.

Capovilla, F. C. (2008). Recursos para educação de crianças com necessidades especiais e articulação entre educação especial e inclusiva. *O Mundo da Saúde*, **32**(2), 208–214.

Carneiro, A. T. S. (2010). *Sistema de Reconhecimento do Alfabeto da LIBRAS por Visão Computacional e Redes Neurais*. tese de mestrado, Engenharia de Teleinformática - Universidade Federal do Ceará.

Carneiro, A. T. S., Cortez, P. C., and Costa, R. C. S. (2010). Reconhecimento de gestos da libras com classificadores neurais a partir dos momentos invariantes de hu. In *Anais do Interaction South America*, Curutiba, PR.

Cheddad, A., Condell, J., Curran, K., and McKevitt, P. (2009). A new colour space for skin tone detection. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 497–500.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.

Delashmit, W. H. and Manry, M. T. (2005). Recent developments in multilayer perceptron neural networks. In *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference, MAESC*.

Faria, H., Maganhotte Júnior, A., Bortolozzi, K., and Bortolozzi, F. (2001). Crianças surdas, interface muda ! multimeios e tutoria inteligente no auxílio da aprendizagem da escrita da língua portuguesa por crianças surdas. *Revista Eletrônica de Iniciação Científica*.

Felipe, T. A. (2007). *A Libras em Contexto*. WalPrint Gráfica e Editora, Brasília.

Foley, J. D., Van Dam, A., Feiner, S. K., and Hughes, J. F. (1990). *Computer Graphics: principles and practice*. Addison-Wesley.

Gomez, G., Sanchez, M., and Sucar, L. E. (2002). On selecting an appropriate colour space for skin detection. In *Proceedings of Mexican Int. Conference on Artificial Intelligence*, pages 70–79.

Gonzalez, R. C. and Woods, R. E. (2000). *Processamento de imagens digitais*. Edgard Blucher Ltda., São Paulo.

Gritti, T., Shan, C., Jeanne, V., and Braspenning, R. (2008). Local features based facial expression recognition with face registration errors. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–8.

Haykin, S. (2012). *Neural Networks and Learning Machines*. Prentice Hall International Inc., 3rd edition.

Hse, H. and Newton, A. R. (2004). Sketched symbol recognition using zernike moments. In *Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004.*, pages 367–370, Cambridge, United Kingdom. IEEE.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Hwang, S.-K. and Kim, W.-Y. (2006). A novel approach to the fast computation of zernike moments. *Pattern Recognition (PR)*, **39**(11), 2065–2076.

Jati, H. and Dominic, D. D. (2008). Human skin detection using defined skin region. In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 1, pages 1–4.

Javier, R. and Rodrigo, V. (2004). Skin detection using neighborhood information. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition (FG2004)*, pages 463–468, Seoul, Korea.

Johansson, E. M., Dowla, F. U., and Goodman, D. M. (1991). Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method. *Int. J. Neural Syst.*, **2**(4), 291–301.

Khan, R. Z. and Ibraheem, N. A. (2012). Survey on gesture recognition for hand image postures. *Computer and Information Science*, **5**(3), 110–121.

Khotanzad, A. and Hong, Y. H. (1990). Invariant image recognition by zernike moments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **12**(5), 489–497.

Kovac, K., Peer, P., and Solina, F. (2003). Human skin color clustering for face detection. In *Proceedings of EUROCON 2003. Computer as a Tool. The IEEE Region 8*, pages 144–148. IEEE.

Liwicki, S. and Everingham, M. (2009). Automatic recognition of fingerspelled words in british sign language. In *Proceedings of the 2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB'09), in conjunction with CVPR2009*, pages 50–57, Los Alamitos, CA, USA. IEEE Computer Society.

Maji, S. (2005). A comparison of feature descriptors. *Scientific Literature Digital Library, CiteSeerX*.

Misra, A., Abe, T., and Deguchi, K. (2011). Hand gesture recognition using histogram of oriented gradients and partial least squares regression. In *MVA*, pages 479–482.

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *Trans. Sys. Man Cyber Part C*, **37**(3), 311–324.

Mohandes, M. A. (2013). Recognition of two-handed arabic signs using the cyberglove. *Arabian Journal for Science and Engineering*, **38**(3), 669–677.

Monteiro, M. S. (2006). História dos movimentos dos surdos e o reconhecimento da libras no brasil. *Educação Temática Digital*, **7**(2), 292–302.

Otitiano-Rodriguez, K. C., Cámara-Chávez, G., and Menotti, D. (2012). Hu and zernike moments for sign language recognition. *Proceeedings of the 16th International Conference on Image Processing, Computer Vision and Pattern Recognition*, **2**(16).

Oujaoura, M., Ayachi, R., Fakir, M., Bouikhalene, B., and Minaoui, B. (2012). Zernike moments and neural networks for recognition of isolated arabic characters. *International Journal of Computer Engineering Science (IJCES)*, **2**(3).

Oyeka, I. C. A. and Ebuh, G. U. (2012). Modified wilcoxon signed-rank test. *Open Journal of Statistics (OJS)*, **2**(3).

Panwar, M. (2012). Hand gesture recognition based on shape parameters. In *Computing, Communication and Applications (ICCCA), 2012 International Conference on*, pages 1–6.

Parvini, F., Mcleod, D., Shahabi, C., Navai, B., Zali, B., and Ghandeharizadeh, S. (2009). An approach to glove-based gesture recognition. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques*, pages 236–245, Berlin, Heidelberg. Springer-Verlag.

Pavlovic, V. I., Sharma, R., and Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, **19**(7), 677–695.

Pizzolato, E. B., dos Santos Anjo, M., and Pedroso, G. C. (2010). Automatic recognition of finger spelling for libras based on a two-layer architecture. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 969–973, New York, NY, USA. ACM.

Prema, C. and Manimegalai, D. (2012). Article: Survey on skin tone detection using color spaces. *International Journal of Applied Information Systems*, **2**(2), 18–26. Published by Foundation of Computer Science, New York, USA.

Qader, H. A., Ramli, A. R., and Al-haddad, S. (2006). Fingerprint recognition using zernike moments. *The International Arab Journal of Information Technology*, **4**(4).

Quadros, R. M. and Karnopp, L. B. (2004). *Língua de Sinais Brasileira - Estudos Linguísticos*. Artmed, Porto Alegre, RS.

Rahman, M. H. and Afrin, J. (2013). Article: Hand gesture recognition using multiclass support vector machine. *International Journal of Computer Applications*, **74**(1), 39–43. Full text available.

Ramos, C. R. (2013). *LIBRAS: A Língua dos Surdos Brasileiros*. Editora Arara Azul, Petrópolis, RJ.

Ren, Z., Yuan, J., and Zhang, Z. (2011). Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 1093–1096, New York, NY, USA. ACM.

Ribeiro, H. L. (2006). *Reconhecimento de Gestos Usando segmentação de Imagens Dinâmicas de Mãos baseada no Modelo de Mistura de Gaussianas e Cor de Pele*. tese de mestrado, Escola de Engenharia de São Carlos - Universidade de São Paulo.

Richard, L. (2011). *Concepts and Applications of Inferential Statistics*.

Rosenblatt, F. (1960). *Perceptual generalization over transformation groups. Self Organizing Systems*. Pergamon Press, New York.

Singha, J. and Das, K. (2013). Indian sign language recognition using eigen value weighted euclidean distance based classification technique. *CoRR*, **abs/1303.0634**.

Sousa, D. V. C. (2010). Um olhar sobre os aspectos linguísticos da língua brasileira de sinais. *LITERRA ONLINE*, **1**(2).

Souza, C. R., Pizzolato, E. B., and Anjo, M. d. S. (2012). Recognizing static signs from the. *CoRR*, **abs/1210.7461**.

Starner, T. and Pentland, A. (1995). Real-time american sign language recognition from video using hidden markov models. In *Computer Vision, 1995. Proceedings., International Symposium on*, pages 265–270.

Stephan, J. J. and Khudayer, S. (2010). Gesture recognition for human-computer interaction (hci). *Int. J. Adv. Comp. Techn.*, **2**(4), 30–35.

Stokoe, W. C. (1976[1960]). *A dictionary of American Sign Language on linguistic principles*. Linstok Press, Silver Spring, Md.

Strobel, K. L. and Fernandes, S. (1998). *Aspectos Linguísticos da Língua Brasileira de Sinais*. SEED/SUED/DEE, Curitiba, PR.

Sturman, D. J. and Zeltzer, D. (1994). A survey of glove-based input. *IEEE Comput. Graph. Appl.*, **14**(1), 30–39.

Taheri, S. and Hesamian, G. (2013). A generalization of the wilcoxon signed-rank test and its applications. *Statistical Papers*, **54**(2), 457–470.

Teodoro, B. and Digiampietri, L. A. (2013). A local alignment based sign language recognition system. In *Proceedings of SibGRAPI 2013, Works in Progress (WIP)*.

Tian, S., Lu, S., Su, B., and Tan, C. L. (2013). Scene text recognition using co-occurrence of histogram of oriented gradients. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 912–916.

Tomaschitz, J. and Facon, J. (2009). Skin detection applied to multi-racial images. In *Systems, Signals and Image Processing, 2009. IWSSIP 2009. 16th International Conference on*, pages 1–3.

Triesch, J. and von der Malsburg, C. (1996). Robust classification of hand postures against complex backgrounds. In *FG*, pages 170–175. IEEE Computer Society.

Tsolakidis, D. G., Kosmopoulos, D. I., and Papadourakis, G. (2014). Plant leaf recognition using zernike moments and histogram of oriented gradients. In *Artificial Intelligence: Methods and Applications - 8th Hellenic Conference on AI*, pages 406–417.

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

Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *IN PROC. GRAPHICON-2003*, pages 85–92.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385.

Yang, J. and Xu, Y. (1994). Hidden markov model for gesture recognition. Technical Report CMU-RI-TR-94-10, Robotics Institute, Pittsburgh, PA.

Zahedi, M., Rybach, D., Deselaers, T., and Ney, H. (2006). Using geometric features to improve continuous appearance-based sign language recognition. In *In Proceedings of BMVC 06, 17th British Maschine Vision Conference*, pages 1019–1028.

Zarit, B. D., Super, B. J., and Quek, F. K. H. (1999). Comparison of five color models in skin pixel classification. In *Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, page 58. IEEE Computer Society Washington, DC, USA.

Zhang, C., Yang, X., and Tian, Y. (2013). Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *FG*, pages 1–8. IEEE.

Appendices

A

Resultados da aplicação dos algoritmos de pele

Esta seção destina-se à apresentação de resultados referentes à aplicação da abordagem para reconhecimento de pele.

Os gráficos mostrados nas Figura A.1, A.2 e A.3 correspondem aos testes realizados nas imagens dos datasets Annotated Skin Database, Labelled Faces in the Wild (LFW) e nas 50 imagens de teste adquiridas pela câmera, respectivamente. As média encontradas para cada medição foram apresentadas na Tabela 3.1 da seção de Metodologia. Os termos ARN 35 e ARN 45 referem-se às arquiteturas de redes neurais com 35 e 45 neurônios, respectivamente.

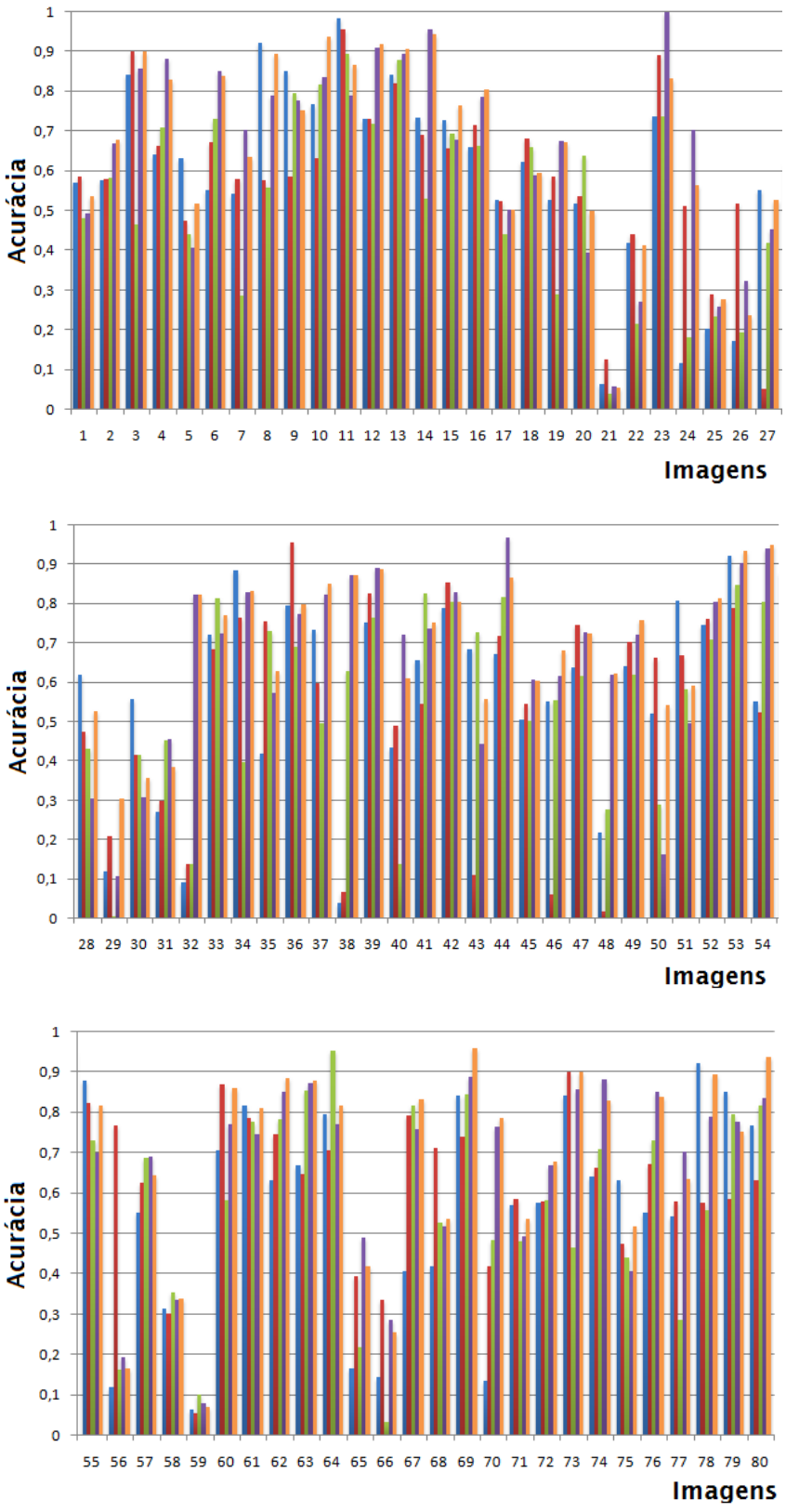


Figura A.1: Acurácia dos algoritmos de pele no Annotated Skin Database.

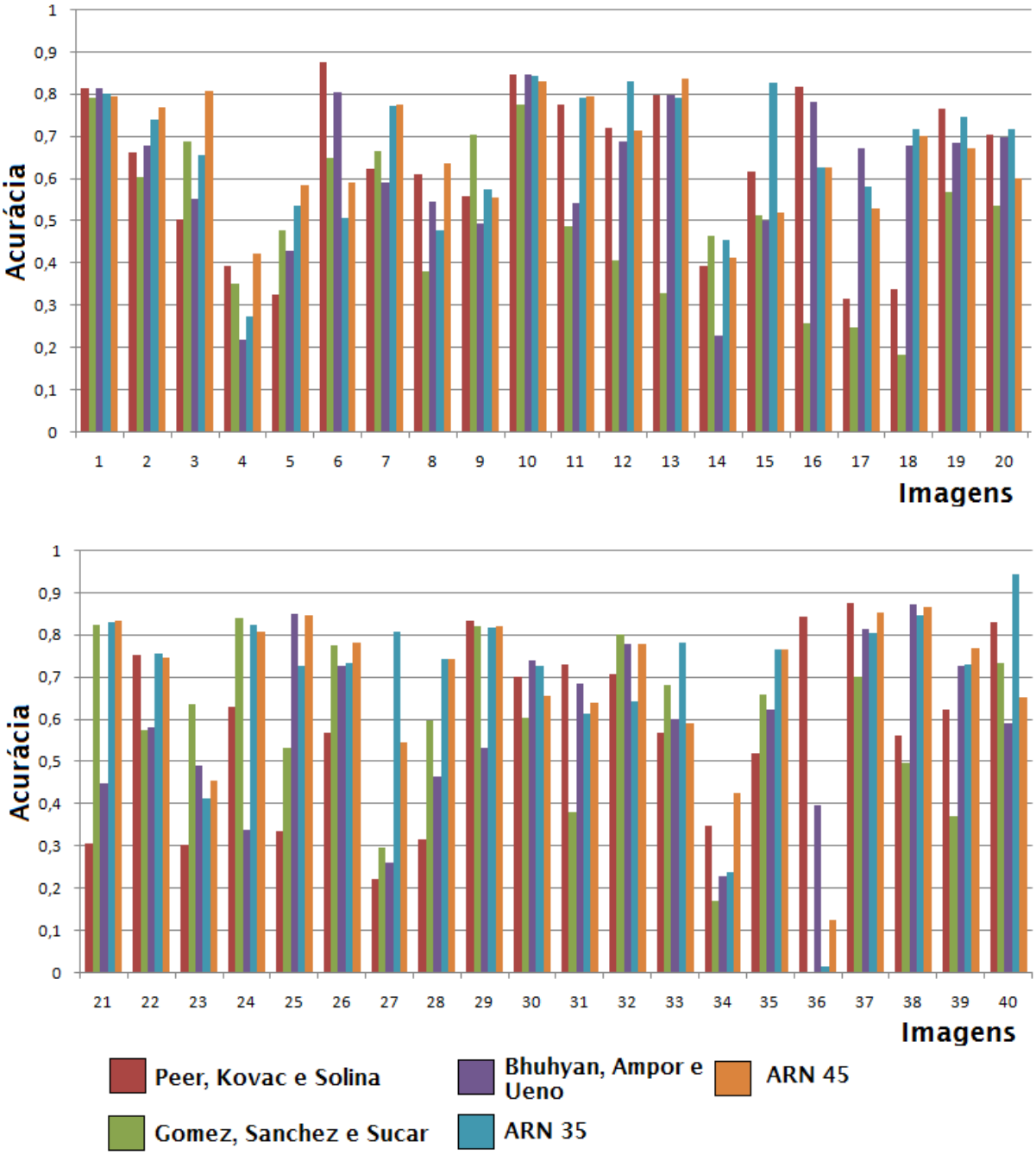


Figura A.2: Acurácia dos algoritmos de pele no LFW.

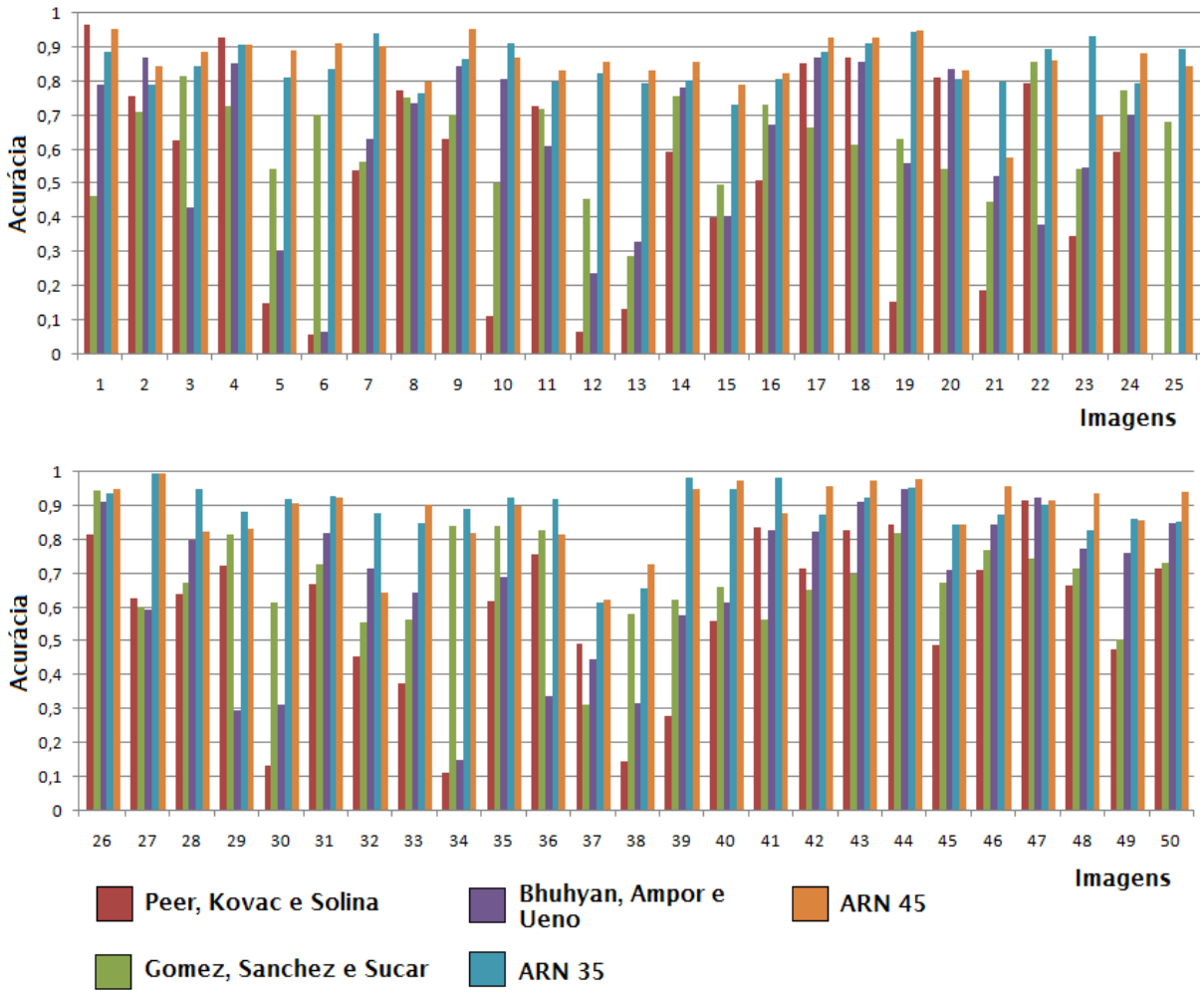


Figura A.3: Acurácia dos algoritmos de pele.

B

Resultados detalhados dos testes realizados na abordagem

Esta seção destina-se à apresentação de dados referentes ao teste da arquitetura considerando a arquitetura de 2 estágios. Além disso, são apresentadas as tabelas com os dados referentes aos testes de sementes fixas, usados para avaliar a robustez da abordagem.

B.1 Resultados detalhados da aplicação da abordagem na arquitetura de 2 estágios

A Tabela B.1 apresenta os dados detalhados da aplicação da presente abordagem, utilizando a arquitetura de 2 estágios. Esta tabela apresenta medidas de taxas de acerto encontradas para cada fold, juntamente com as médias e desvio-padrão.

B.2 Resultados detalhados da aplicação da abordagem na arquitetura de 1 estágio

A Tabela B.2 apresenta os dados detalhados da aplicação da presente abordagem, utilizando a arquitetura de 1 estágio. Esta tabela apresenta medidas de taxa de acerto encontradas para cada fold, juntamente com as médias e desvio-padrão.

B.3 Resultados detalhados da utilização do HOG e MIZ isoladamente

As Tabelas B.3 e B.4 apresentam os dados provenientes da aplicação da abordagem considerando o uso de apenas o descritor HOG e do descritor MIZ, respectivamente.

Tabela B.1: Resultados detalhados da aplicação da arquitetura de 2 estágios

Sinais / Folds	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Média	Desvio-Padrão
1	100	100	100	100	100	100	100	0,00
2	100	95	95	100	95	90	95,83	3,76
4	100	100	100	100	95	100	99,16	2,04
5	100	100	100	100	100	100	100	0,00
7	95	100	100	100	100	85	96,67	10
9	95	100	100	90	90	100	95,83	4,91
A	100	100	90	95	100	85	95	6,32
B	80	75	100	80	85	100	86,67	10,80
C	85	90	100	90	95	80	90,00	7,07
D	95	95	100	100	100	90	96,67	4,08
E	95	100	100	95	100	100	98,33	2,58
F	100	90	100	100	100	95	97,50	4,18
G	95	100	100	100	100	100	99,16	2,04
I	95	80	90	80	90	100	89,16	8,01
L	90	80	100	100	100	100	95,00	8,36
M	100	85	95	100	100	100	96,67	6,05
N	100	100	95	100	100	100	99,16	2,04
O	95	100	95	100	95	95	96,67	2,58
P	100	100	100	100	100	100	100,00	0,00
Q	100	100	100	95	95	100	98,33	2,58
R	95	100	100	95	95	95	96,67	2,58
S	90	100	90	100	100	100	96,67	5,16
T	100	100	100	100	100	100	100	0,00
U	95	90	90	95	95	80	90,83	5,84
V	95	85	100	95	100	100	95,83	5,84
W	100	95	90	100	90	95	95	4,47
X	100	95	100	95	100	100	98,33	2,58
Y	100	90	100	100	80	100	95,00	8,36
Avião	100	100	100	100	100	100	100,00	0,00
América	100	100	100	100	100	100	100,00	0,00
Adulto	100	90	90	95	100	100	95,83	4,91
Casa	100	100	100	100	100	100	100,00	0,00
Gasolina	100	100	100	100	100	100	100,00	0,00
Identidade	100	90	100	100	100	100	98,33	4,08
Junto	100	100	100	100	90	100	98,33	4,08
Lei	95	100	100	100	100	100	99,16	2,04
Palavra	100	100	100	100	100	100	100,00	0,00
Pedra	100	100	100	100	100	95	99,16	2,04
Pequeno	95	90	100	100	80	80	90,83	9,17
Verbo	90	95	95	100	90	100	95,00	4,47

Tabela B.2: Resultados detalhados da aplicação da arquitetura de 1 estágio

Sinais / Folds	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Média	Desvio-Padrão
1	95	100	100	100	60	100	92,50	16,04
2	95	95	90	85	100	60	87,50	14,40
4	100	95	90	100	60	100	90,83	15,62
5	100	100	80	100	95	100	95,83	8,01
7	95	95	100	100	100	65	92,50	13,69
9	90	95	100	75	85	85	88,33	8,75
A	70	70	65	85	90	55	72,50	12,94
B	95	90	90	80	70	100	87,50	10,83
C	55	85	95	85	95	90	84,16	14,97
D	95	95	100	90	100	85	94,16	5,84
E	75	85	80	85	100	90	85,83	8,61
F	90	90	100	100	100	80	93,33	8,16
G	95	100	100	100	100	100	99,16	2,04
I	95	90	95	85	90	95	91,66	4,08
L	90	80	90	90	100	100	91,66	7,52
M	90	75	85	100	100	95	90,83	9,70
N	100	100	90	100	85	100	95,83	6,64
O	65	90	90	90	60	60	75,83	15,62
P	85	90	90	100	100	100	94,16	6,64
Q	95	90	95	90	90	100	93,33	4,08
R	95	95	100	90	95	95	95,00	3,16
S	85	90	75	95	95	100	90,00	8,94
T	85	95	95	100	100	100	95,83	5,84
U	95	90	85	50	95	60	79,16	19,34
V	90	85	90	80	90	55	81,66	13,66
W	100	100	95	95	75	80	90,83	10,68
X	90	95	85	95	100	100	94,16	5,84
Y	95	95	50	85	50	65	73,33	21,13
Avião	90	95	85	100	100	100	95,00	6,32
América	100	95	100	100	100	100	99,16	2,04
Adulto	100	85	90	85	90	100	91,66	6,81
Casa	100	100	100	100	100	95	99,16	2,04
Gasolina	100	100	100	100	100	100	100,00	0,00
Identidade	95	95	100	100	80	85	92,50	8,21
Junto	100	100	100	100	100	100	100,00	0,00
Lei	90	95	95	100	100	95	95,83	3,76
Palavra	100	100	90	85	100	100	95,83	6,64
Pedra	100	100	100	95	100	90	97,50	4,18
Pequeno	95	95	95	70	90	70	85,83	12,41
Verbo	95	95	95	100	80	80	90,83	8,61

Tabela B.3: Acerto (%) e desvio-padrão para aplicação da abordagem com o uso somente do descritor HOG

Sinais / Folds	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Média	Desvio-Padrão
1	95	100	100	100	80	100	95,83	8,01
2	95	95	95	100	95	80	93,33	6,83
4	100	100	90	100	70	100	93,33	12,11
5	100	100	85	100	100	90	95,83	6,64
7	95	95	100	100	100	85	95,83	5,84
9	95	100	90	85	90	100	93,33	6,05
A	95	90	90	95	100	75	90,83	8,61
B	80	75	100	80	75	100	85,00	11,83
C	70	85	100	90	95	80	86,67	10,80
D	95	95	100	90	95	90	94,16	3,76
E	80	95	100	95	100	100	95,00	7,74
F	95	90	100	100	100	90	95,83	4,91
G	90	100	100	95	95	100	96,67	4,08
I	95	85	80	80	90	95	87,50	6,89
L	90	80	90	95	100	95	91,67	6,83
M	100	85	100	95	100	90	95,00	6,32
N	100	95	90	100	100	100	97,50	4,18
O	90	100	100	95	85	95	94,16	5,84
P	95	100	95	100	100	100	98,33	2,58
Q	100	100	100	95	95	100	98,33	2,58
R	95	100	90	95	95	95	95,00	3,16
S	90	100	85	100	95	95	94,16	5,84
T	95	100	100	95	100	100	98,33	2,58
U	95	90	90	85	95	80	89,16	5,84
V	95	90	100	95	95	90	94,16	3,76
W	100	95	90	100	90	95	95,00	4,47
X	100	95	85	95	95	90	93,33	5,16
Y	90	90	100	95	80	100	92,50	7,58
Avião	90	95	90	100	95	100	95,00	4,47
América	100	95	100	100	100	95	98,33	2,58
Adulto	100	90	85	95	90	95	92,50	5,24
Casa	100	100	100	100	100	100	90	4,08
Gasolina	100	100	100	100	100	100	100,00	0,00
Identidade	95	90	100	95	100	100	96,67	4,08
Junto	100	90	100	100	95	100	97,50	4,18
Lei	90	100	85	100	100	90	94,16	6,64
Palavra	100	100	95	100	100	90	97,50	4,18
Pedra	95	100	100	95	100	90	96,67	4,08
Pequeno	95	85	100	80	80	85	87,50	8,21
Verbo	90	95	95	95	90	95	93,33	2,58

Tabela B.4: Acerto (%) e desvio-padrão para aplicação da abordagem com o uso somente do descritor MIZ

Sinais / Folds	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Média	Desvio-Padrão
1	75	80	70	90	80	90	80,83	8,01
2	90	90	85	100	90	90	90,83	4,91
4	100	100	80	95	65	90	88,33	13,66
5	80	80	85	60	95	85	80,83	11,58
7	75	85	100	100	90	80	88,33	10,32
9	85	100	75	80	90	90	86,67	8,75
A	90	60	85	70	90	75	78,33	12,11
B	80	65	90	65	55	90	74,16	14,63
C	80	65	90	100	80	75	81,66	12,11
D	95	90	80	80	80	90	85,83	6,64
E	60	80	70	75	75	80	73,33	7,52
F	90	70	65	60	80	95	76,67	14,02
G	80	70	90	95	70	85	81,67	10,32
I	55	85	70	95	80	90	79,16	14,63
L	90	80	95	95	100	100	93,33	7,52
M	80	85	75	70	70	95	79,16	9,70
N	100	80	90	85	100	95	91,67	8,16
O	85	95	95	80	85	70	85,00	9,48
P	95	90	95	95	90	95	93,33	2,58
Q	95	100	90	100	100	95	96,67	4,08
R	85	100	80	100	90	95	91,67	8,16
S	60	60	85	90	95	75	77,50	15,08
T	85	90	90	100	85	95	90,83	5,84
U	75	95	90	90	85	80	85,83	7,35
V	90	100	90	100	85	80	90,83	8,01
W	90	100	100	95	75	85	90,83	9,70
X	95	90	90	90	85	90	90,00	3,16
Y	70	65	75	85	85	65	74,16	9,17
Avião	95	100	100	80	75	100	91,67	11,25
América	100	85	100	95	90	90	93,33	6,05
Adulto	100	80	85	85	85	80	85,83	7,35
Casa	95	100	95	95	90	95	95,00	3,16
Gasolina	95	100	100	90	100	95	96,67	4,08
Identidade	55	70	60	95	80	75	72,50	14,40
Junto	90	100	95	95	90	100	95,00	4,47
Lei	100	100	85	70	95	95	90,83	11,58
Palavra	90	95	100	90	95	80	91,67	6,83
Pedra	100	100	90	100	95	90	95,83	4,91
Pequeno	90	80	90	95	75	75	84,16	8,61
Verbo	95	100	95	90	90	100	95,00	4,47

C

Apêndices

Esta seção destina-se à apresentação de dados referentes aos testes da presente abordagem. As tabelas e gráficos aqui mostrados compreendem resultados dos testes realizados utilizando o NTU Hand Digit Dataset (Ren *et al.*, 2011).

C.1 Resultados detalhados da aplicação sobre o NTU Hand Digit Dataset

A Tabela C.1 apresenta os dados detalhados da aplicação da presente abordagem sobre o NTU Hand Digit Dataset (Ren *et al.*, 2011).

Tabela C.1: Acerto (%) e desvio-padrão para gestos do NTU Dataset

Gestos / Folds	Fold1	Fold2	Fold3	Fold4	Fold5	Média	Desvio-Padrão
G1	95	100	90	95	100	96	4,18
G2	100	90	85	100	100	95	7,07
G3	90	100	100	100	95	97	4,47
G4	95	95	100	90	90	94	4,18
G5	100	100	95	90	95	96	4,18
G6	80	95	100	95	95	93	7,58
G7	95	100	90	100	95	96	4,18
G8	95	85	95	95	100	94	5,47
G9	95	100	100	100	100	99	2,23
G10	100	100	100	95	95	98	2,73