CLASSIFICAÇÃO DE EMOÇÕES FACIAIS UTILIZANDO A REDE NEURAL SEM PESOS WISARD

Leopoldo André Dutra Lusquino Filho

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Rio de Janeiro
Março de 2018

CLASSIFICAÇÃO DE EMOÇÕES FACIAIS UTILIZANDO A REDE NEURAL SEM PESOS WISARD

Leopoldo André Dutra Lusquino Filho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Felipe Maia Galvão França, Ph.D.

Prof. Priscila Machado Vieira Lima, Ph.D.

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Alberto Ferreira De Souza, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2018

Lusquino Filho, Leopoldo André Dutra

Classificação de Emoções Faciais utilizando a Rede Neural sem Pesos WiSARD/Leopoldo André Dutra Lusquino Filho. – Rio de Janeiro: UFRJ/COPPE, 2018.

XIII, 52 p.: il.; 29, 7cm.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2018.

Referências Bibliográficas: p. 49 – 52.

1. Redes neurais sem peso. 2. Classificação de emoções. 3. Classificação de expressões faciais. I. França, Felipe Maia Galvão *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A Krishna-Balarama

Agradecimentos

nama om visnu-padaya krishna-presthaya bhu-tale
srimate chandramukha swamin iti namine
namas te guru presthaya hrdayananda padashraye
prabhupada-karuna-shakti gita-sara-pracharine

nama om visnu-padaya krishna-presthaya bhu-tale
srimate bhaktivedanta swamin iti namine
namas te sarasvate deve gaura-vani-pracharine
nirvisesha-sunyavadi-paschatya-desa-tarine

jaya sri krishna chaitanya prabhu nityananda
sri advaita gadadhara srivasadi-gaura-bhakta-vrinda

hare krishna hare krishna krishna krishna hare hare
hare rama hare rama rama rama hare hare

Sempre serei endividado com meus pais, Balabhadra e Carmen, por terem feito seu melhor para que eu desenvolvesse senso de dever e retidão, e com a minha companheira, Damodara, pelos seus esforços para me apoiar em qualquer atividade.

Agradeço aos meus professores Felipe e Priscila, por serem muito mais que orientadores acadêmicos, sendo naturalmente fontes de amizade e inspiração. Agradeço também a COPPE, especialmente ao PESC/LabIA, por criar um ambiente tão agradável e descontraído para o trabalho de pesquisa, a CAPES, pelo suporte financeiro, e a todos os amigos que fiz ao longo desta jornada.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CLASSIFICAÇÃO DE EMOÇÕES FACIAIS UTILIZANDO A REDE NEURAL SEM PESOS WISARD

Leopoldo André Dutra Lusquino Filho

Março/2018

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Programa: Engenharia de Sistemas e Computação

Classificação automática de emoções em expressões faciais é uma questão central em Computação Afetiva e uma das principais premissas na construção de modelos de interface homem-máquina cada vez mais responsivos, com uma vasta gama de aplicações. Muitos sistemas baseados em inteligência artificial são capazes de resolver este problema com acurácia elevada, mas em geral tais modelos possuem um processo de aprendizado lento e custoso.

O reconhecimento de expressões faciais através do uso de um classificador de n-uplas baseado em WiSARD é explorado neste trabalho. A eficácia desta rede neural sem peso é testada no desafio específico de identificar emoções em fotografias de faces, limitadas às seis emoções básicas descritas no trabalho seminal de Ekman e Friesen (1977) sobre a identificação de expressões faciais. Experimentos realizados com os dois principais *datasets* encontrados na literatura demonstraram sua competitividade com o atual estado-da-arte, assim como sua grande velocidade tanto na fase de aprendizado, quanto na de classificação. Diferentes abordagens de pré-processamento, assim como estudos sobre a melhor forma de representação de imagens em entradas binárias neste problema específico também estão são descritos neste texto.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

FACIAL EMOTION CLASSIFICATION USING A WISARD WEIGHTLESS NEURAL NETWORK

Leopoldo André Dutra Lusquino Filho

March/2018

Advisors: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Department: Systems Engineering and Computer Science

Automatic classification of emotions in facial expressions is a central issue in Affective Computing and one of the main premises in the construction of increasingly responsive man-machine interface models with a wide range of applications. Many systems based on artificial intelligence are able to solve this problem with high accuracy, but in general such models have a slow and expensive learning process.

The recognition of facial expressions through the use of a WiSARD-based n-tuple classifier is explored in this work. The competitiveness of this weightless neural network is tested in the specific challenge of identifying emotions from photos of faces, limited to the six basic emotions described in the seminal work of Ekman and Friesen (1977) on identification of facial expressions. Experiments carried out with the two main datasets found in the literature demonstrated their competitiveness with current state-of-the-art, as well as their great speed in both the learning and classification phases. Different preprocessing approaches as well as studies on how best to represent images in binary inputs in this specific problem are also described in this text.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Sendo um dos elementos básicos na relação humana, a classificação de emoções tem sido estudada por diversas áreas do conhecimento científico, como biologia, psicologia e antropologia, e com o avanço da computação, ela se tornou útil para variados tipos de aplicações, desde tutores inteligentes, até sistemas de segurança, investigação forense, redes sociais, computação gráfica e games.

Entre as diferentes manifestações de emoções humanas, uma das mais significativas é certamente através de expressões faciais, que são capazes de transmitir uma gama muito valiosa de informações sobre o estado mental do indivíduo. Com o crescente avanço da interação homem-máquina, torna-se necessário criar construtos computacionais que entendam de forma empática as emoções humanas e que sejam suficientemente responsivos a elas. Alguns exemplos de aplicações que podem ser beneficiadas com tal reconhecimento de expressões faciais incluem produtos capazes de obter um feedback da reação do usuário, sistemas pedagógicos digitais adaptáveis ao humor do aluno, serviços de aconselhamento psicológico, carros capazes de monitorar o nível de estresse do motorista, etc.

Uma vez que agentes humanos bem treinados podem extrair da face dados sofisticados, como a ocorrência de mentiras, ou descobrir se houve a simulação de um sentimento e mesmo qual emoção tentou se ocultar, sistemas suficientemente treinados deveriam ser capazes de minimamente se aproximar do desempenho de peritos humanos nestas operações. Segundo alguns pesquisadores (Wang e Kosinski, 2017), tais sistemas poderão perceber características e padrões faciais que são completamente imperceptíveis para humanos. Neste texto apresentaremos nossa contribuição para este importante problema.

1.1 Objetivos e contribuições

Muitos sistemas robustos tem atingido acurácia semi-ótima na detecção de faces e classificação de sentimentos, mas praticamente todos os modelos neuronais que

se prestam a tal tarefa possuem treinamento consideravelmente lento. Uma vez que muitos desses sistemas devem operar online, garantir uma aprendizagem rápida ainda é um requisito altamente desejado e pode ser indispensável em muitos casos. Foi recentemente demonstrado que uma rede neural baseada em WiSARD razoavelmente complexa, aplicada ao problema da análise do crédito financeiro, superou a SVM em algumas ordens de grandeza no tempo de treinamento, mantendo-se muito competitiva com a precisão (Cardoso et al. 2016).

Nossos objetivos aqui se concentram em modelar um sistema baseado em redes neurais sem peso para contornar esta limitação dos paradigmas tradicionais. Outro tópico aqui explorado é a geração de protótipos de faces expressando emoções básicas através de "imagens mentais". Nos deteremos especificamente em como processar imagens de uma forma representativa e ao mesmo tempo simples, transformando-as em entradas binárias, em detectar faces e classificar as emoções por elas expressas.

Obtendo 97,3% e 99,3% de acurácia em dois dos principais *datasets* da literatura, Cohn-Kanade Extended e MMI Database, respectivamente, a solução baseada em WiSARD se mostrou competitiva com o estado-da-arte. Nossos resultados parciais foram publicados em (Lusquino Filho et al., 2018).

1.2 Estrutura da dissertação

No Capítulo 2 é apresentado o escopo conceitual envolvendo o estudo das emoções faciais, assim como o sistema mais popular para sua classificação: o FACS. Também são relatados os principais sistemas classificadores de emoções faciais encontrados na literatura, assim como suas principais características e dificuldades em comum. No Capítulo 3 é introduzido o conceito de rede neural sem peso, sua inspiração biológica e o modelo que será explorado neste trabalho: a WiSARD. Sua expansão DRASiW é também explicada. Estes dois capítulos formam a base teórica que serve de alicerce para o nosso estudo. No Capítulo 4 apresentamos nossa arquitetura, as diferentes técnicas de pré-processamento estudadas e as diferentes abordagens testadas para o problema da detecção de faces. O Capítulo 5 trata de experimentos feitos com a nossa solução nos mais importantes *datasets* da literatura, assim como sua efetiva comparação com o estado-da-arte, além de uma análise das imagens mentais dos discriminadores WiSARD. O Capítulo 6 conclui apresentando nossa análise sobre os resultados encontrados, assim como algumas direções a serem seguidas no futuro.

Capítulo 2

Classificação de Emoções

Neste capítulo será descrito o escopo teórico sobre as emoções universalmente reconhecidas, assim como será dada uma visão geral dos sistemas automáticos desenvolvidos para a execução desta tarefa, suas limitações e o atual estado-da-arte.

2.1 As seis emoções básicas e o FACS

O estudo das expressões faciais tem se desenvolvido desde períodos pré-aristotélicos, através da Fisiognomia, método que tentava inferir a personalidade e alguns aspectos psicológicos de uma pessoa através das suas feições. Nos século 17 e 18 o estudo das feições novamente voltou a agenda de muitos pesquisadores, cujo principal interesse era catalogar todos os diferentes movimentos possíveis da cabeça humana. As principais compilações realizadas a partir dos resultados obtidos por estes estudos foram o livro "Pathomyotomia", do estudioso de anatomia John Bulwer, e as transcriçòes das palestras do pintor Le Brun na Royal Academy of Painting (fis 1606-1656). Mas foi no século 19 que este tipo de investigação começou a tomar contornos mais definidos, quando Charles Darwin escreveu um tratado categorizando várias expressões que havia observado em homens e animais (Darwin 1904).

No século XX, a antropologia voltou sua atenção em descobrir quais emoções eram de fato universais. Um divisor de águas no estudo das emoções faciais foi o trabalho de (Ekman e Friesen 1977), onde eles discutem a possibilidade de se catalogar emoções básicas comunais a todas as culturas e que sejam frutos apenas de fatores biológicos internos. Questionando a validade científica dos trabalhos anteriores de catalogação e análise de emoções, eles denunciam os parâmetros previamente utilizados para caracterização das emoções como sendo subjetivos e, possivelmente, influenciados por interpretações culturais. Buscando por uma abordagem mais objetiva, eles usaram como principal diretriz para identificação de uma emoção, os músculos faciais usados para expressá-la, ao invés da voz, como era o padrão até então. Desta forma, eles estabeleceram o Facial Action Coding System (FACS),

que nada mais é que um sistema catalográfico com todos os músculos capazes de alterarem uma expressão facial. Cada componente do FACS é conhecido como uma "Action Unit" (AU). Essas AUs se dividem em superiores (expressas no entorno dos olhos) e inferiores (expressas na região delimitada pelo nariz e boca).

Exemplos de AUs são: AU1 - franzir a testa internamente, AU2 - franzir a testa externamente, AU45 - fechar os olhos. No entanto, três AUs não envolvem músculos: AU19 - por a língua para fora da boca, AU33 - estufar a bochecha e AU66 - envesgar os olhos. A Figura 2.1 mostra alguns exemplos de AUs da FACS.

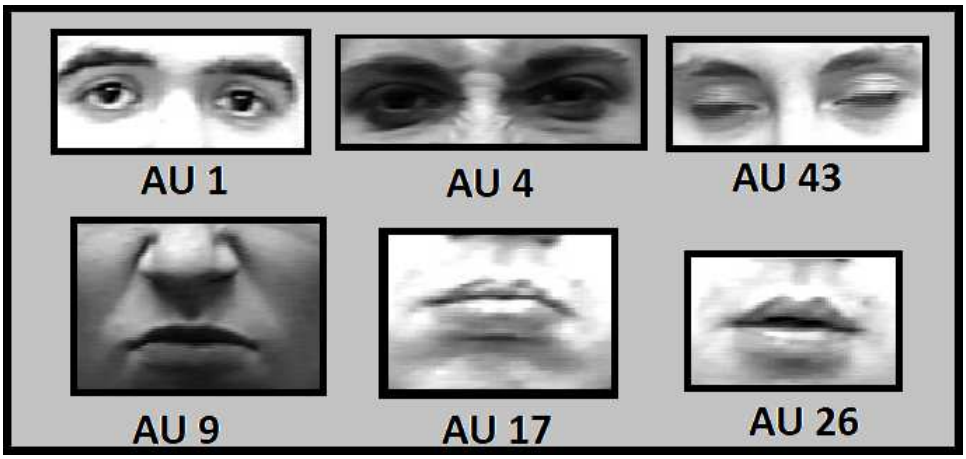

Figura 2.1: Algumas Action Units da FACS.

AUs se dividem em aditivas e não-aditivas. Quando uma AU é independente de qualquer outra, ela é dita ser aditiva. Uma AU não-aditiva é aquela que é alterada e altera outra AU não-aditiva, quando expressas simultaneamente (Cohn et al. 2005).

Outra contribuição significativa deste estudo foi definir quais emoções eram expressas pela combinação das AUs, chegando assim a uma nova proposta para um conjunto de emoções universais: felicidade, tristeza, raiva, repulsa, surpresa e medo, além da emoção neutra. Segundo as definições dadas por (Ekman e Friesen 2003):

- Felicidade é a única emoção necessariamente positiva (surpresa pode ser negativa ou positiva, enquanto as demais são necessariamente negativas) e é acionada por prazer ou excitação. (Figura 2.2)

  Características:

  - Os cantos dos lábios se voltam para cima;
  - A boca pode ou não ser separada, com os dentes expostos ou não;
  - Uma ruga (a dobra naso-labial) se manifesta do nariz até a borda externa além dos cantos dos lábios;
  - As bochechas são levantadas;

Figura 2.2: Felicidade

  - Rugas são manifestadas na parte inferior da pálpebra inferior, que pode estar erguida, mas nunca tensa;
  - Rugas ramificadas são manifestadas nos cantos externos dos olhos.

- Tristeza é uma emoção necessariamente passiva (ao contrário de todas as demais), que surge como consequência do sofrimento subsequente a expectativas não cumpridas, separação de objetos e situações agradáveis ou perda. (Figura 2.3)

Figura 2.3: Tristeza

Características:

  - Os cantos internos das sobrancelhas são erguidos;
  - A pele abaixo da sobrancelha é triangulada com a parte superior da pálpebra;
  - O canto interno da pálpebra superior é levantado;
  - Os cantos dos lábios estão abaixados ou os lábios estão tremendo.

- Raiva é o corolário natural da frustação, quando alguma espécie de ataque é preparado para a sua fonte. Quando alguém experimenta raiva há um aumento da pressão sanguínea, fazendo com que as veias se tornem mais aparentes. (Figura 2.4)

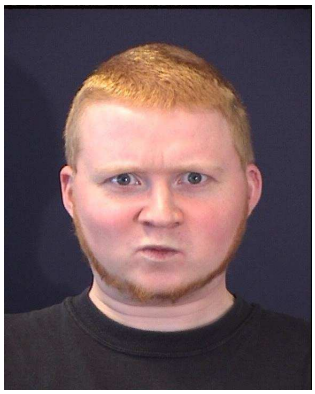

Figura 2.4: Raiva

Características:

  - As sobrancelhas são abaixadas e se aproximam uma da outra;
  - Linhas de expressão verticais aparecem entre as sobrancelhas;
  - A pálpebra inferior se torna tensa e pode ou não ser levantada;
  - A pálpebra superior se torna tensa e pode ou não ser abaixada pela ação da sobrancelha;
  - Os olhos têm um olhar rígido e podem ter uma aparência proeminente;
  - Os lábios estão em qualquer uma das duas posições básicas: pressionados firmemente juntos, com os cantos retos ou baixos, ou abertos, tensos, em uma forma quadrada como se estivessem gritando;
  - As narinas podem estar dilatadas, mas isso não é essencial para a expressão facial de raiva e também pode ocorrer em tristeza.

- Repulsa manifesta-se quando se quer expulsar um determinado elemento do ambiente onde se encontra. (Figura 2.5)

  Características:

  - O lábio superior é erguido;
  - O lábio inferior também é erguido e se aproxima do superior, ou é abaixado e ligeiramente contraído;
  - O nariz se enruga;
  - As bochechas são levantadas;

Figura 2.5: Repulsa

  - Formam-se linhas de expressão abaixo da pálpebra inferior e a própria pálpebra se ergue, mas sem ficar tensa;
  - A sobrancelha é abaixada, abaixando a pálpebra superior.

- Surpresa é a mais breve de todas as emoções. É acionada por eventos inesperados e normalmente é exibida com alguma segunda emoção menos intensa conjuntamente. (Figura 2.6)

Figura 2.6: Surpresa

  Características:

  - As sobrancelhas aparecem curvas e altas e longas rugas são produzidas na testa (em casos onde a pessoa já possui tais rugas naturalmente, elas se tornam mais profundas);
  - As sobrancelhas em um rosto surpreso geralmente são unidas por olhos bem abertos e mandíbulas caídas;
  - A intensidade da surpresa pode ser medida especialmente pela abertura do maxilar.

- Medo é a emoção que é desperta em indivíduos que estão experimentado a percepção de danos e perda e cuja integridade e segurança dependem de fuga.

Apesar de ser muito similar fisicamente a surpresa, estas emoções se distinguem principalmente pela duração, já que surpresa é necessariamente breve. (Figura 2.7)

Figura 2.7: Medo

Características:

- As sobrancelhas são levantadas e desenhadas juntas;
- As rugas ficam centralizadas na testa e não dispostas em toda sua extensão;
- A pálpebra superior é levantada, expondo a esclerótica e a pálpebra inferior fica tensa e esticada;
- A boca está aberta e os lábios estão ligeiramente ligados e esticados.

Muitas expansões e variações do FACS foram criados por Ekman e outros pesquisadores (Sayette et al. 2001). São elas: FAST (Facial Action Scoring Technique; percursor do FACS, lidava apenas com um hemisfério da face), EMFACS(Emotional Facial Action Coding System; correlaciona AUs com as emoções prototípicas), MAX(Maximally Discriminative Facial Movement Coding System; desenvolvido para ser utilizado na detecção de variações mínimas das AUs), EMG (Facial Electromyography; lida com as correntes elétricas em AUs ativas), AFFEX (Affect Expressions by Holistic Judgment; categoriza emoções a partir de todas as regiões da face simultaneamente), Mondic Phases (analisa a interação comportamental entre adultos e crianças), FACSAID(FACS Affect Interpretation Database; utiliza o código do FACS para análise de comportamentos) e FACS Infantil. Entre esses se destaca o EMFACS (Ekman e Friesen 1983), que cataloga combinações de AUs que podem ser usadas para definir emoções faciais. A Tabela 2.1 dá alguns exemplos de descrições de emoções do EMFACS.

Além destas variações do FACS, outro sistema de codificação facial baseado em AUs é o MPEG-4, um conjunto de métricas para modelagem facial, que se

Tabela 2.1: Alguns exemplos de classificação de emoções utilizando o EMFACS.

Emoção	Combinação de Action Units
Neutro	AU 41 + 42 ou AU 41 + 44
Felicidade	AU 12 ou AU 13
Tristeza	AU 1 + 4 + 15 ou AU 6 + 15
Raiva	AU 23 e 24
Medo	AU 1 + 2 + 3 + 4 ou AU 1 + 2 + 3 + 4 + 5 ou AU 20
Repulsa	AU9 ou 10
Surpresa	AU 1 + AU 2

tornou o padrão internacional para a comunidade de animadores. (Cowie et al. 2008) indica a relação entre o MPEG-4 FAPs e as FACS AUs: "MPEG-4 foca-se principalmente em sintetizar expressões faciais e animação, definindo os Facial Animation Parameters (parâmetros facias de animação), conhecidos como FAPs, que são fortemente relacionados com as Action Units, o núcleo do FACS."

Muitos pesquisadores criticam o FACS de Ekman, devido a fatores como o tamanho da amostra de cada grupo cultural, o fato de que os pesquisados não puderam exprimir o significado das suas emoções em suas próprias linguas, a ocorrência de múltiplas emoções simultaneamente e a descoberta posterior a publicação de seus resultados de pequenos grupos isolados que não expressavam suas emoções de acordo com a descrição do FACS (Russel e Fehr 1987). Apesar das críticas consistentes por meio da comunidade científica ao FACS e ao desenvolvimento de novos estudos mais completos sobre a codificação facial de emoções, utilizamos neste trabalho a classificação de Ekman já que os principais *datasets* da área ainda são baseados nela.

2.2 Sistemas de classificação de emoções

Computação Afetiva é um ramo da Inteligência Artificial ocupada em reconhecer, interpretar, processar e simular estados afetivos humanos. Sua origem formal se deu com a publicação do artigo de Rosalind Piccard (Piccard 1995), onde ela explica a importância de máquinas capazes de simular empatia e a possibilidade do seu desenvolvimento a partir da moderna teoria cognitiva e dos avanços da inteligência artificial na época, além de apontar as direções futuras a serem seguidas pelos pesquisadores da área.

Dada a relevância da detecção e análise de expressões faciais, a automação de tais atividades ocupou um lugar central na Computação Afetiva. Três principais tipos de sistemas foram são desenvolvidos neste sentido: detecção e rastreamento de faces, extração de *features* faciais e classificação de expressões faciais. Naturalmente, sistemas visando detectar as AUs do FACS, bem como as seis emoções de Ekman

ocuparam espaço significativo entre eles. (Bettadapura 2012) fez uma revisão ampla sobre tais sistemas.

A maior parte dos sistemas encontrados na literatura utiliza alguma técnica de visão computacional para extrair *features*, que depois são fornecidas a algum classificador. Entre as técnicas para extração de *features*, algumas soluções populares são: Optical Flow (vetor de movimentação de objetos por *frame*), Gabor Wavelets (transformada capaz de obter informações espaço-temporais de um sinal, sendo capaz de representar completamente imagens), Modelos Multi-estados (modelam um processo com várias transições, onde cada estado pode ser submetido a uma heurística distinta), KLT Tracker (método que utiliza informações a respeito da intensidade espacial da imagem para reduzir custo computacional da busca por *features* específicas) e PBVD Tracker (similar ao anterior, utiliza o modelo de deformação parcial de volume de Bézier).

Já os classificadores com melhor resultado são: redes bayesianas (modelos de representação do conhecimento que trabalham com o conhecimento incerto e incompleto por meio do Teorema de Bayes), Modelo Oculto de Markov (modelo estatístico em que o sistema modelado é assumido como um processo de Markov com parâmetros desconhecidos), SVM (modelo de aprendizado supervisionado que separa classes distintas através de um hiperplano) e redes neurais.

Notadamente, as redes convolucionais tem se destacado ultimamente como estado-da-arte para diferentes formas de classificação facial. Este tipo de rede se inspira no cortex visual dos animais para realizar o minímo de pré-processamento possível, tornando a rede invariante a deslocamento e escala (Lecun et al. 1998).

O atual estado-da-arte para classificação de emoções em faces, a rede convolucional DeXpression (Burkert et al. 2015), funciona de acordo com a arquitetura esboçada na Figura 2.8 (Representação do original delineada em [Burkert et al. 2015]):

1)Os dados são pré-processados através de uma primeira convolução, responsável por aplicar 64 filtros na imagem de entrada;

2)Os dados pré-processados são enviados para outra camada que normaliza as imagens da amostra;

3)Depois as imagens são enviadas a dois blocos FeatEx (Bloco de Extração de Features Paralelas);

4)As *features* extraídas pelas FeatEx são enviadas a uma outra camada, responsável pela classificação. DeXpression conseguiu balancear uma substancial extração de *features* com seus FeatEx, ao mesmo tempo que é barato computacionalmente se comparado com outros modelos convolucionais.

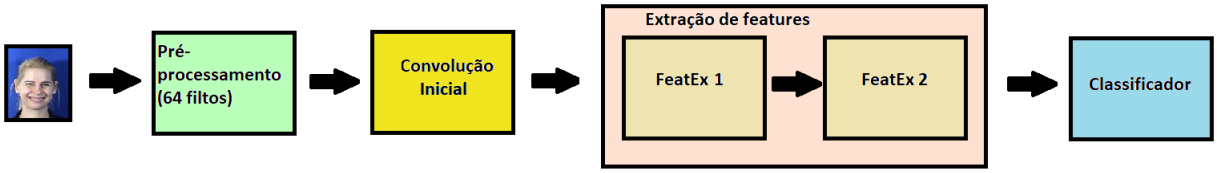

Figura 2.8: Arquitetura do DeXpression.

Comparando as matrizes de confusão destes diversos sistemas, alguns erros se mostraram comunais a todos eles: raiva e repulsa, medo e felicidade, medo e raiva e tristeza e raiva. Outro ponto em comum é o fato de surpresa e felicidade serem as emoções de mais fácil reconhecimento.

(Kotsia et al. 2008) estudou o efeito de oclusões em cada uma das seis expressões faciais prototípicas e mostrou como elas podem reduzir significativamente a acurácia da classificação de raiva, medo, felicidade e tristeza caso ocorram na região superior da face e de repulsa e surpresa caso ocorram na região inferior.

2.3 Métodos de detecção facial

Parte fundamental da classificação de emoções faciais está relacionada a própria detecção de faces em imagens. Sendo uma das primeiras áreas de pesquisa a ser desenvolvida em automação de análise facial, este tópico ainda é discutido atualmente e possui algumas questões em aberto. Seguem as descrições dos principais métodos de detecção facial encontrados na literatura:

2.3.1 Métodos baseados em conhecimento

Apoiam-se em uma base de regras sobre o que seria uma face, como suas *features* mínimas (tais como dois olhos, uma boca, um nariz), assim como seu posicionamento na face. Normalmente uma janela deslizante percorre a imagem, procurando as *features* obrigatórias, e uma vez que elas são encontradas, são validadas de acordo com a base de regras (Yang et al. 2002). A limitação tradicional desse tipo de abordagem é que se o conjunto de regras for muito generalista provavelmente haverá grande quantidade de falso-positivos e se for muito especificista haverá alta incidência de falso-negativos.

2.3.2 Métodos baseados em características invariantes

Baseiam-se em características comunais de uma face, que não irão ser afetadas por qualquer tipo de posicionamento ou oclusão que a face sofra, como por exemplo, a cor da pele e sua textura (Yang et al. 2002). A cor da pele tende a formar um

cluster cromático no espaço da imagem, sendo possível modelá-lo através de uma distribuição Gaussiana (Wang e Sung 1999), de forma que pode-se separar a imagem em regiões com face e sem face. A principal desvantagem de tais métodos é que eles normalmente são sensíveis a variações da luminância e outros ruídos semelhantes, envolvendo o ambiente.

2.3.3 Métodos baseados em templates

Procuram em uma imagem, geralmente através de alguma heurística ou algoritmo genético, candidatos a face devido a sua semelhança com elipses. Depois verificam através de alguma função de energia o grau de correspondência entre as imagens encontradas e o *template* de face proprietário do método específico (Alattar e Rajala, 1999).

2.3.4 Métodos baseados em amostragem

Utilizam um conjunto de imagens para treinarem um modelo. A abordagem tradicional entre estes métodos se baseiam-se em *eigenfaces* (Turk e Pentland 1991), vetores que melhor descrevem a distribuição da representação facial dentro do espaço da imagem, devido a sua semelhança com aquelas imagens que compõem a base de treinamento. Dentro desta categoria, muitas soluções tem utilizado abordagens baseadas em redes neurais e Modelo Oculto de Markov.

Capítulo 3

Redes neurais sem peso

Neste capítulo serão descritos os conceitos fundamentais sobre redes neurais, em especial o modelo WiSARD.

3.1 Redes neurais

Como se sabe, as típicas habilidades humanas de identificar similaridades em diferentes objetos e abstrair possíveis diferenças entre eles, identificando assim padrões e classificando-os, são possíveis devido à complexa estrutura do cérebro e sua unidade fundamental, o neurônio, exibido na Figura 3.1 (neu 2018).

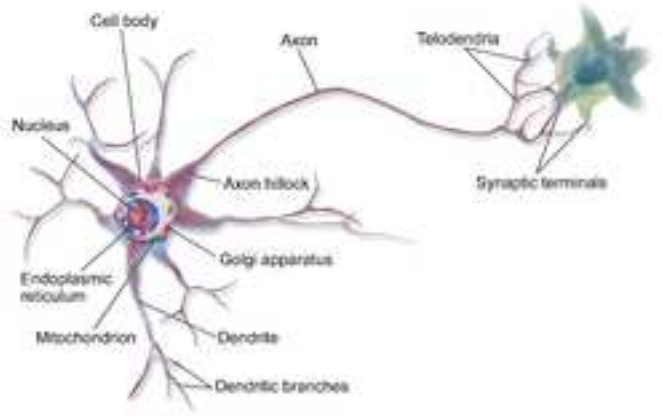

Figura 3.1: Neurônio biológico.

Na década de 40, o neurologista Warren McCulloch e o matemático Walter Pitts, tentaram simular a atividade neuronal artificialmente. Desta forma, eles desenvolveram um modelo matemático que captura o comportamento neurológico, caracterizado pela transmissão da informação por uma série de células nervosas (McCulloch e Pitts 1943). Baseado nesse modelo, desenvolveu-se o conceito de rede neural artificial (RNA).

O cérebro humano é formado por bilhões de neurônios que se comunicam através de fibras protoplasmáticas chamadas axônios. Analogamente, a topologia da rede neural é elaborada de forma que ela seja dividida em camadas contendo nós, ou seja, estruturas de dados ligadas entre si, onde uma dada entrada é processada e transmitida para outros nós através de sinapses.

Nas redes neuronais tradicionais (Figura 3.2), uma função de ativação é atribuída a cada nó, realizando uma espécie de soma ponderada dos estímulos recebidos por ele, e um peso é atribuído a cada ligação entre nós e estes são atualizados em cada iteração do processamento da informação na rede. Estas iterações seguem-se até que haja convergência entre os resultados obtidos e as saídas esperadas pela rede. Os pesos sinapticos podem ser positivos ou negativos, de forma análoga as extremidades excitatórias e inibitórias do sistema neuronal biológico.

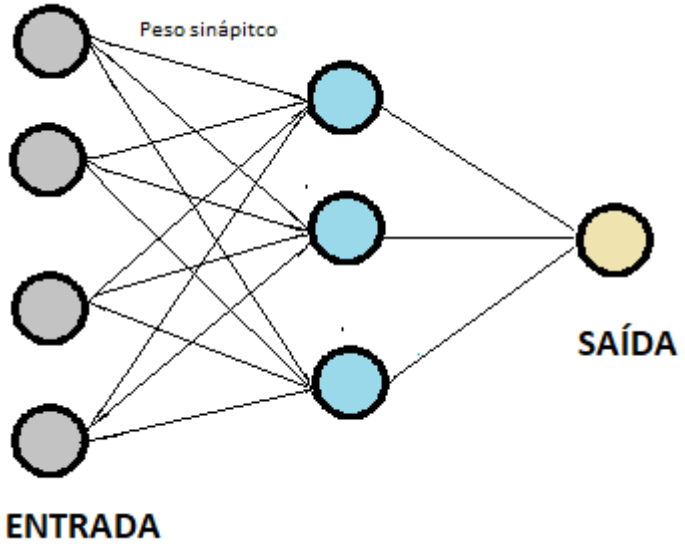

Figura 3.2: Rede neural tradicional

Uma descrição precisa é dada em (Haykin 1998), onde é dito que rede neural é: "Um processador distribuído paralelamente de forma maçiça, constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso. Ela assemelha-se ao cérebro em dois aspectos: (1) o conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem; (2) forças de conexão entre neurônios (os pesos sinápticos) são utilizadas para armazenar o conhecimento adquirido."

3.2 Os modelos sem peso

Uma alternativa ao modelo tradicional é a rede neural sem peso, que não se preocupa em usar modelos matemáticos para reproduzir em um sentido mais estrito o

comportamento do neurônio. Esse tipo de rede foi originalmente apresentada em (Aleksander et al. 1984), para se tratar problemas relacionados a processamento de imagem, tendo sua primeira implementação comercial no modelo WiSARD. Muitos outras redes sem peso foram desenvolvidas (Aleksander et al. 2009), como PLNs, GSNs e GRAMs.

Uma rede sem peso utiliza pequenas partes de memória de acesso aleatório, chamadas aqui apenas de RAMs, para armazenar informações obtidas com o processamento dos padrões usados em seu treinamento (Figura 3.3). Essas endereçam tantos *bits* quantos forem aqueles que formam a entrada. Uma rede de nós-RAM que classifique n-uplas é completamente análoga a uma rede neural com peso com n inputs, com a diferença que não existe necessidade de calibragem sináptica. Tais nós-RAMs foram usados pela primeira vez na década de 50, por (Bledsoe e Browning 1959) eem problemas de identificação de padrões. Posteriormente, Aleksander introduziu nós RAMs juntamente com Stored Logic Adaptative Microcircuits (SLAM) como componentes básicos de uma rede de aprendizado adaptativo (Aleksander 1966).

Aqui pode-se dizer que esse tipo de rede explora um fato relacionado ao funcionamento neurológico (Koch e Poggio, 1987): redes neurais com peso se preocupam com a modelagem matemática dos neurônios biológicos, mas apesar disso as suas conexões sinápticas sempre terminam no neurônio de soma. Enquanto biologicamente isso seja possível não é obrigatório, uma vez que muitas sinapses terminam na árvore dendrítica, que é a estrutura mais notável do neurônio. Pode-se fazer uma analogia entre os dendritos e o conteúdo dos endereços de memória da rede neural sem peso: quando o dendrito está longe do neurônio de soma e sua influência na formação do output do neurônio de soma é fraca, então o conteúdo de memória é um booleano falso, e vice-versa.

Para que haja consistência na informação, os *bits* de endereçamento devem se referir sempre à mesma informação. No caso do processamento de imagens, por exemplo, isso significa que cada RAM sempre endereçará cada *pixel* na mesma posição de memória.

Cada neurônio RAM tem suas posições inicializadas com zero e, durante o processo de treinamento elas são populadas de acordo com o valor dos *bits* de entrada. Desta forma, durante o processo de reconhecimento é verificado se existe algum conhecimento armazenado nos *bits* endereçados pela entrada fornecida.

Cada RAM equivale a uma característica da imagem. Ademais, quanto maior for o número de RAMs maior será a quantidade de características similares a serem procuradas entre os exemplos usados no treinamento e as imagens a serem submetidas a reconhecimento. Por exemplo, caso haja apenas uma RAM, então para que uma imagem seja reconhecida pela rede ela deve ser idêntica àquela que foi aprendida. Logo, é possível verificar que quanto maior for a quantidade de RAMs mais

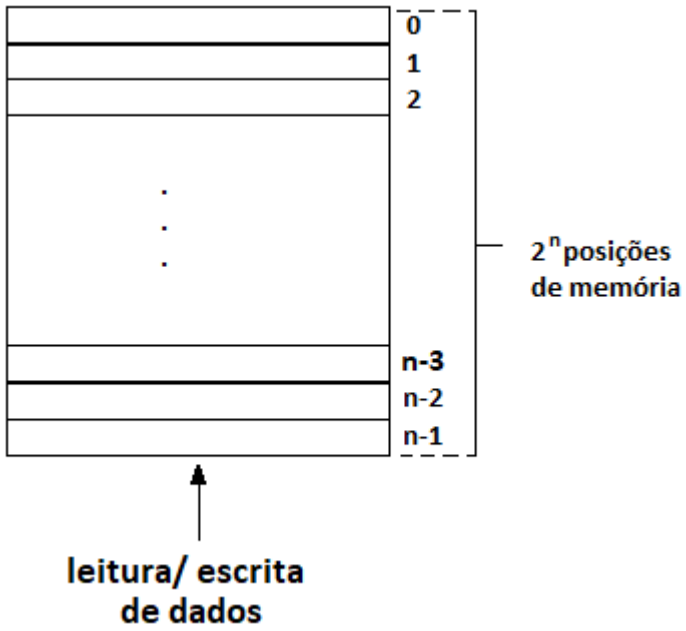

Figura 3.3: Neurônio RAM

generalista será a rede e quanto menor for a quantidade de RAMs mais especificista essa será. Logo, o tamanho da entrada da RAM de uma rede é um parâmetro importante. Uma vantagem das redes sem peso é que, por poupar processamento com a ausência dos cálculos de atualização dos pesos e consistindo a maior parte do seu aprendizado em acessos a memória, seu desempenho neste quesito é superior ao modelo com peso.

3.3 WiSARD

A implementação pioneira das redes neurais sem peso foi desenvolvida na década de 80 e é conhecida como WiSARD (Wilkie, Stonham and Aleksander's Recognition Device). Essa rede continua a se destacar pela sua fácil implementação e compreensão, assim como pelo seu baixo custo computacional. Muitas expansões da WiSARD foram desenvolvidas, como a NC-WiSARD (Bandeira 2010), que busca inspiração no Neocognitron para generalizar seu aprendizado quando os exemplos possuem características díspares (tais como tamanho e posicionamento de imagens), a StreamWiSARD (Cardoso 2012), que lida com fluxo contínuo de dados e a ClusWiSARD, cujo treinamento é não-supervisionado (Cardoso et al. 2016).

3.3.1 Estrutura

Tradicionalmente, devido à limitação de hardware, os endereços de memória da WiSARD eram booleanos, servindo apenas para indicar se houve algum acesso a

eles em algum momento do período de treino. Devido a não possuir o mecanismo de aprendizado que se origina da calibragem dos pesos sinápticos, havendo apenas o armazenamento das informações contidas no treinamento, os neurônios da WiSARD por si só não possuiriam capacidade de generalização. Por isso foi idealizado um mecanismo conhecido como discriminador, que é constituído por um conjunto de RAMs e uma função de soma, de forma análoga à função de ativação das redes com peso, e é o responsável pela capacidade de classificação da WiSARD.

Uma WiSARD terá tantos discriminadores quantas forem as classes possíveis de serem atribuídas aos objetos que devem ser reconhecidos pela rede. Por exemplo, caso a rede deva identificar qual letra maiúscula do alfabeto latino a imagem representa, então serão necessários 26 discriminadores, de forma que cada um seja treinado com a imagem da letra correspondente (Figura 3.4).

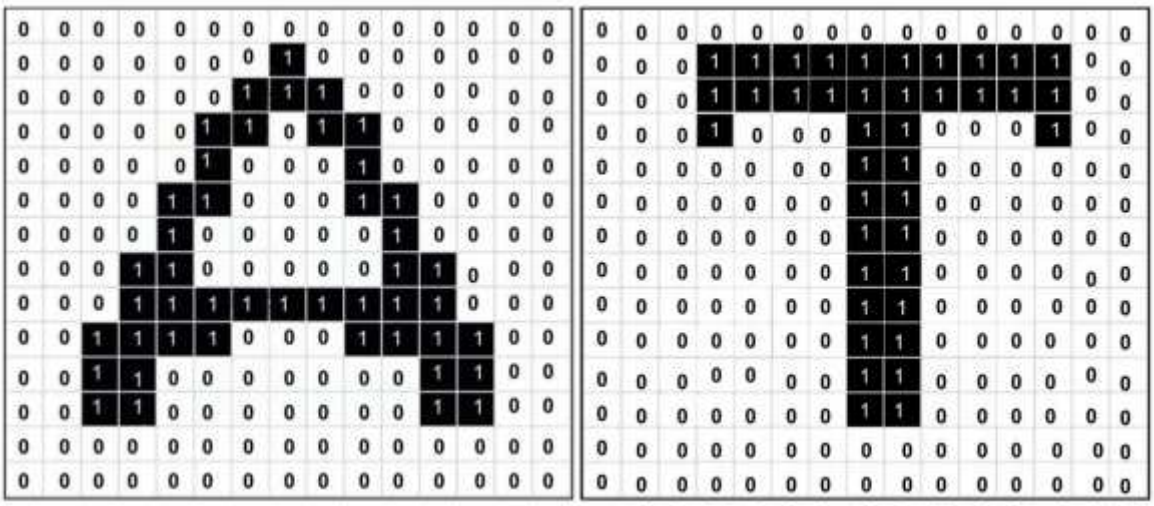

Figura 3.4: Discriminadores para as classes da letra A e T.

Pode-se definir a WiSARD como um classificador de n-uplas, composto por discriminadores de classe (Figura 3.5). Cada discriminador é um conjunto de N RAMs, com n linhas de endereço cada e uma função de soma Σ (Aleksander et al. 1984).

Antes das fases de treinamento e classificação, todos os conteúdos dos nós-RAM são definidos como zero. O treinamento de um padrão binário pertencente a uma determinada classe é feito da seguinte maneira: para qualquer padrão apresentado à retina de entrada, todos os endereços de memória do discriminador correspondentes ao booleano verdadeiro da entrada são também alterados para um.

Na fase de classificação, cada discriminador terá um placar r, formado pela soma das suas posições de memória que foram acessadas em algum momento da fase de treinamento e que são correspondentes a entrada agora fornecida. Caso o discriminador com maior placar tenha uma confiança em sua resposta superior aquela estipulada para a rede, a classe correspondente a ele é eleita como sendo

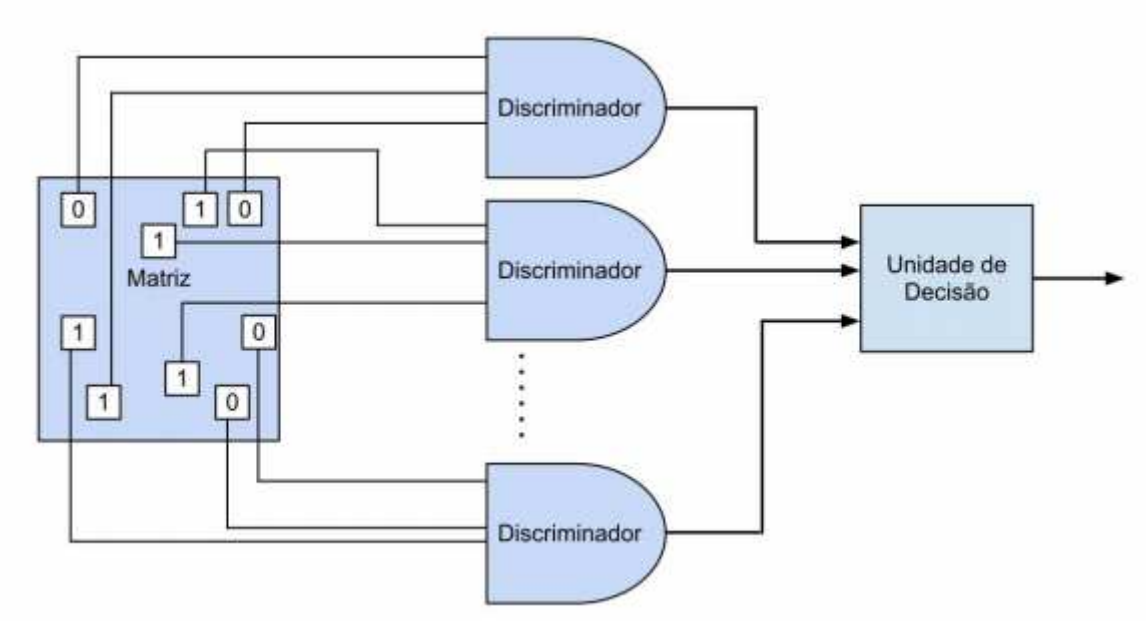

Figura 3.5: Estrutura de uma rede WiSARD com seus discriminadores.

aquela que define a entrada. Tal confiança é calculada como mostrado na equação 3.1.

$$\varphi = \frac{r_{max} - r_{max-1}}{r_{max}} \tag{3.1}$$

3.3.2 Retina

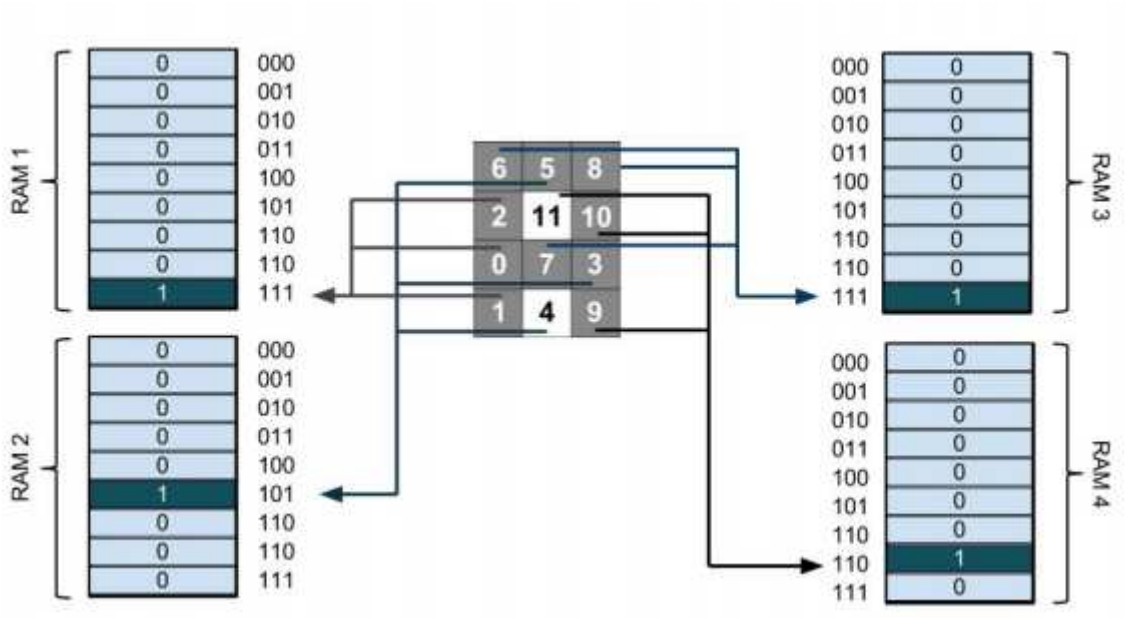

Figura 3.6: Input-retina.

Cada discriminador-RAM recebe um padrão de entrada binário com $N*n$ *bits*, chamado retina (Figura 3.6), que são relacionados a todas as linhas de endereçamento das N RAMs do discriminador por meio de uma ordenação pseudo-aleatória.

Segue um exemplo de funcionamento da retina:

1. Usaremos uma imagem com dimensões 3X3, logo precisaremos de nove endereços de memória;

2. Usaremos uma rede onde cada RAM tenha três entradas, logo serão usadas três RAM;

3. Atribui-se desta forma a ordenação dos *pixels* (Figura 3.7):

0	1	2
3	4	5
6	7	8

Figura 3.7: Ordenamento original dos *pixels*.

4. A retina cria uma relação simples entre os *pixels* da imagem e as entradas das RAMs: [(0,0); (1,1); (2,2); (3,3); (4,4); (5,5); (6,6); (7,7); (8,8)];

5. É feito um embaralhamento entre todos os pares do mapeamento, de modo que nenhum par fique inalterado e que o mapeamento permaneça 1:1 :

(a) Alterando a posição 0 - Escolhido aleatoriamente o número 3 - [**(0,3)**; (1,1); (2,2); **(3,0)**; (4,4); (5,5); (6,6); (7,7); (8,8)];

(b) Alterando a posição 1 - Escolhido aleatoriamente o número 8 - [(0,3); **(1,8)**; (2,2); (3,0); (4,4); (5,5); (6,6); (7,7); **(8,1)**];

(c) Alterando a posição 2 - Escolhido aleatoriamente o número 1 - [(0,3); **(1,2)**; **(2,8)**; (3,0); (4,4); (5,5); (6,6); (7,7); (8,1)];

(d) Alterando a posição 3 - Escolhido aleatoriamente o número 6 - [(0,3); (1,2);

(2,8); **(3,6)**; (4,4); (5,5); **(6,0)**; (7,7); (8,1)];

(e) Alterando a posição 4 - Escolhido aleatoriamente o número 2 - [(0,3); (1,2); **(2,4)**; (3,6); **(4,8)**; (5,5); (6,0); (7,7); (8,1)];

(f) Alterando a posição 5 - Escolhido aleatoriamente o número 1 - [(0,3); **(1,5)**; (2,4); (3,6); (4,8); **(5,2)**; (6,0); (7,7); (8,1)];

(g) Alterando a posição 6 - Escolhido aleatoriamente o número 5 - [(0,3); (1,5); (2,4); (3,6); (4,8); **(5,0)**; **(6,2)**; (7,7); (8,1)];

(h) Alterando a posição 7 - Escolhido aleatoriamente o número 3 - [(0,3); (1,5); (2,4); **(3,7)**; (4,8); (5,0); (6,2); **(7,6)**; (8,1)];

(i) Alterando a posição 8 - Escolhido aleatoriamente o número 0 - [**(0,1)**; (1,5); (2,4); (3,7); (4,8); (5,0); (6,2); (7,6); **(8,3)**];

6. A configuração final da entrada é [(0,1); (1,5); (2,4); (3,7); (4,8); (5,0); (6,2); (7,6); (8,3)].

3.3.3 Bleaching

Um problema com o fato do conteúdo das posições de memória serem apenas um *bit* que identificava se houve acesso a elas é que se houver excesso de treinamento possivelmente haverá muitos *pixels* preenchidos na imagem treinada no discriminador, impossibilitando o aprendizado de qualquer padrão. Isso faz com que a rede tenda à generalização total conforme aumente o seu treinamento. Sendo um modelo monotônico, uma posição de memória da WiSARD que tenha sido alterada para '1', nunca será decrementada, voltando a configuração inicial. Ou seja, um padrão aprendido pela WiSARD nunca é esquecido.

Para minimizar esta dificuldade, a solução foi fazer com que o conteúdo de cada posição de memória guarde um número inteiro para ser usado como contador, de forma a descobrir quais subconjuntos de *pixels* foram mais treinados. Dessa forma, para que um neurônio seja considerado ativo não basta apenas que o conteúdo endereçado por ele tenha sido acessado, mas que tenha tido uma quantidade de acessos igual ou superior a um limiar previamente determinado, denominado de *bleaching* (alvejante em inglês) (Grieco et al. 2010).

Observando-se a saturação dos discriminadores ao longo da fase de treinamento, percebeu-se que ao longo do tempo suas posições de memórias tendiam a serem

idênticas, ainda que tenham tido taxas de acesso muito distintas. Dessa forma, surgiu a ideia de se utilizar um threshold que limitasse a influência de acessos a determinadas posições de memória que não fossem realmente relevantes para se formar o padrão de um discriminador, alvejando o treinamento das células do nó-RAM.

Tradicionalmente, o *bleaching* é inicializado com valor zero e é incrementado gradualmente, na medida em que existe empate entre os discriminadores. Quanto maior o *bleaching*, menos generalista será a rede. As Figuras 3.8 e 3.9 (Carneiro 2012) mostram a influência do *bleaching* na etapa de classificação:

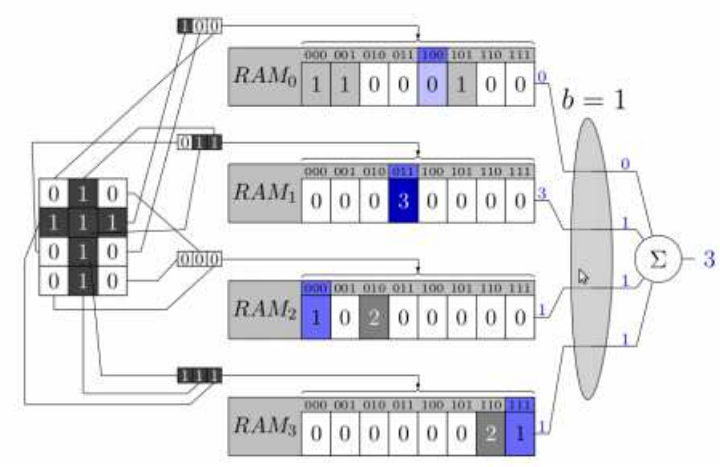

Figura 3.8: Bleaching b = 1.

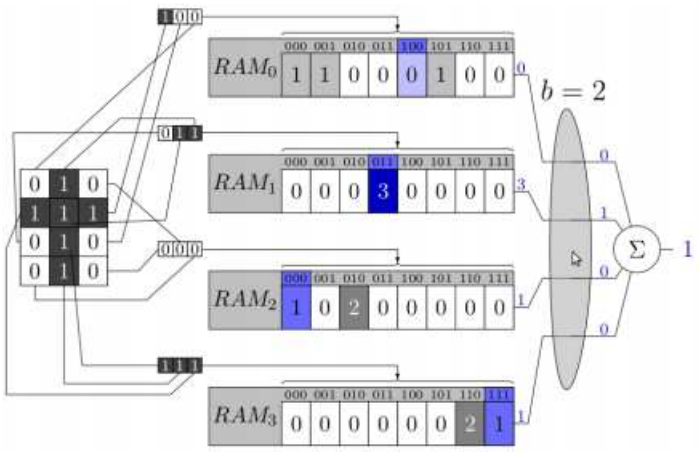

Figura 3.9: Bleaching b = 2.

3.3.4 DRASiW

A possibilidade de se visualizar o aprendizado de um discriminador é muito útil para a validação do treino de uma RNA, uma vez que se pode verificar se os padrões que estão sendo aprendidos correspondem aos desejados. Uma expansão à WiSARD que permite esse tipo de visualização foi relatada em (Gregorio 1997) e (Soares 1998).

Esse tipo de capacidade que pode ser adicionada à WiSARD é conhecida como DRASiW, por "inverter" o processo da etapa de treinamento. Seu resultado é chamado de "imagem mental". Para obtê-la, deve-se percorrer as RAMs de um discriminador, decodificando o conteúdo dos endereços de memória em posições da imagem. Isso só é possível porque o mapeamento pseudoaleatório é conhecido e é sempre o mesmo para toda a rede. A Figura 3.10 ilustra o uso da DRASiW a partir de um discriminador treinado a partir da imagens de dígitos.

Figura 3.10: Resultado do DRASiW a partir do treinamento de dígitos.

A Figura 3.11 (Bandeira 2010) mmostra como as imagens mentais são alteradas com a variação do *bleaching*. Enquanto na primeira imagem o *bleaching* é nulo, na terceira ele está no seu valor ótimo.

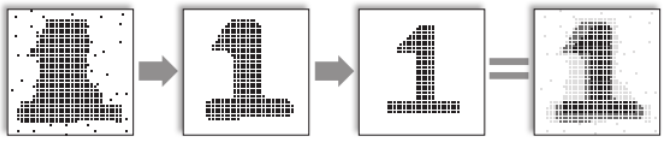

Figura 3.11: DRASiW com bleaching.

Capítulo 4

Metodologia

Basicamente dois problemas devem ser resolvidos pelo nosso modelo: detectar faces em imagens e classificar a emoção por ela exibida. Desta forma, nossa solução utiliza WiSARD para ambas as finalidades. Neste capítulo detalharemos nossa proposta de arquitetura, assim como as alternativas experimentadas para se alcançar a melhor configuração.

4.1 Arquitetura Proposta

Como delineado na Figura 4.1, a imagem ao ser submetida ao sistema é primeiramente pré-processada, tornando-se uma entrada binária. No segundo estágio do sistema, a imagem pré-processada é submetida a um detector de faces, que fará a extração da seção da imagem contendo a parte da face que contém AUs. Esta nova imagem é fornecida então ao terceiro estágio do sistema, um classificador WiSARD com sete discriminadores. Ocasionalmente a imagem poderá ser novamente pré-processada entre os estágios II e III. Diferentes abordagens foram testadas para os estágios I e II e serão descritas a seguir. Mais detalhes sobre a configuração dos parâmetros do estágio III serão dadas no próximo capítulo.

4.2 Pré-processamento

Uma vez que a WiSARD aceita apenas entradas binárias, é necessário antes da submissão de exemplos a rede, que eles sejam traduzidos adequadamente para uma lista de booleanos. Uma vez que neste trabalho, todos os discriminadores do classificador possuem a mesma retina, as imagens devem necessariamente serem redimensionadas para o seu tamanho ($N*n$ *bits*) antes de prosseguirem pelos diferentes níveis do sistema, caso contrário não haverá compatibilidade entre as informações da entrada e as posições de memória dos discriminadores da rede. Serão descritas a seguir

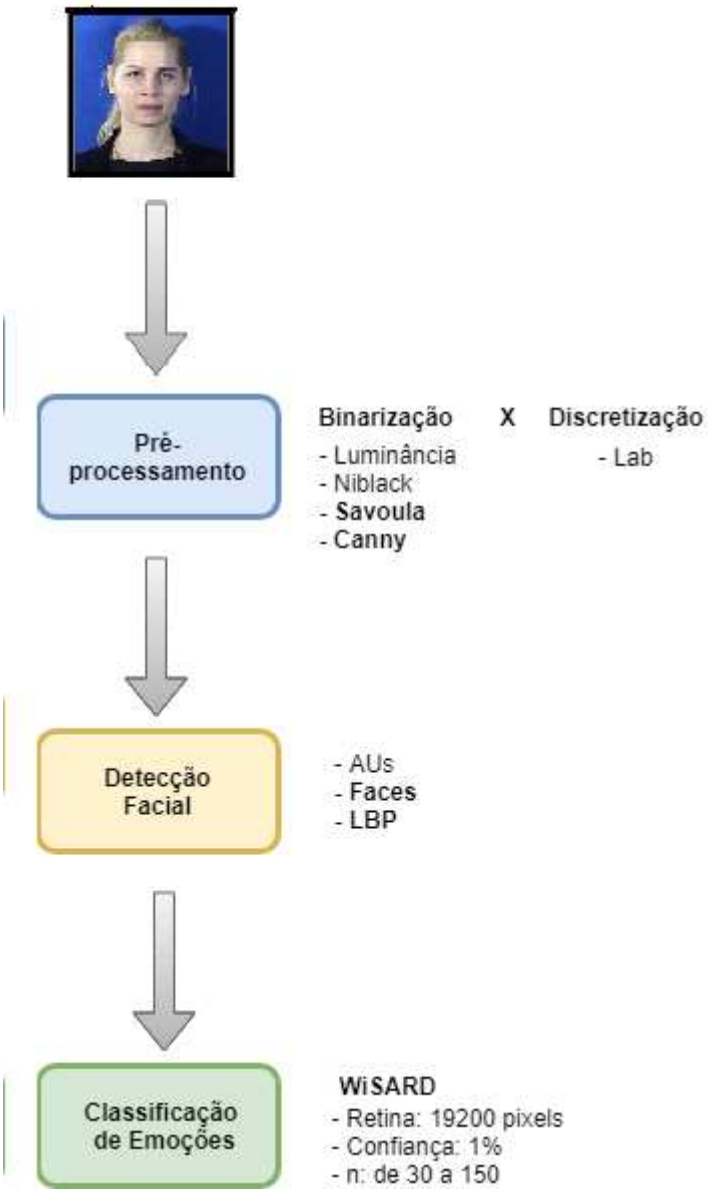

Figura 4.1: Arquitetura proposta. Estágio I: Pré-processamento. Estágio II: Detecção Facial. Estágio III: Classificação de Emoções.

as técnicas de pré-processamento para tradução de imagens coloridas em entradas binárias avaliadas neste trabalho.

4.2.1 Binarização

Binarização é o processo de se transformar uma imagem em uma imagem binária, ou seja, uma imagem onde só existe duas possibilidades de cores para os *pixels*. Em uma imagem binária cada *pixel* pode ser armazenado em um único *bit*, uma vez que

toda informação sobre sua cor pode ser representada por 0 ou 1.

Binarização pela luminância

Este tratamento reduz para duas a quantidade de cores a serem tratadas, transformando a imagem original em preto e branco a partir de um limite que, nesse caso, é baseado na luminância da fotografia, ou seja, na intensidade da luz refletida na foto. A luminância de cada *pixel* é calculada segundo a fórmula $L = 0,2126R + 0,7152G + 0,0722B$, onde R, G e B correspondem às intensidades de vermelho, verde e azul do *pixel* na escala RGB, respectivamente (Stokes et al. 1996). Dessa forma, o espaço cromático será reduzido apenas a duas cores, uma será aquela dos *pixels*, onde $L >= \alpha\overline{L}$ e a outra daqueles onde $L < \alpha\overline{L}$. A Figura 4.2 mostra exemplos de imagens binarizadas pela luminância:

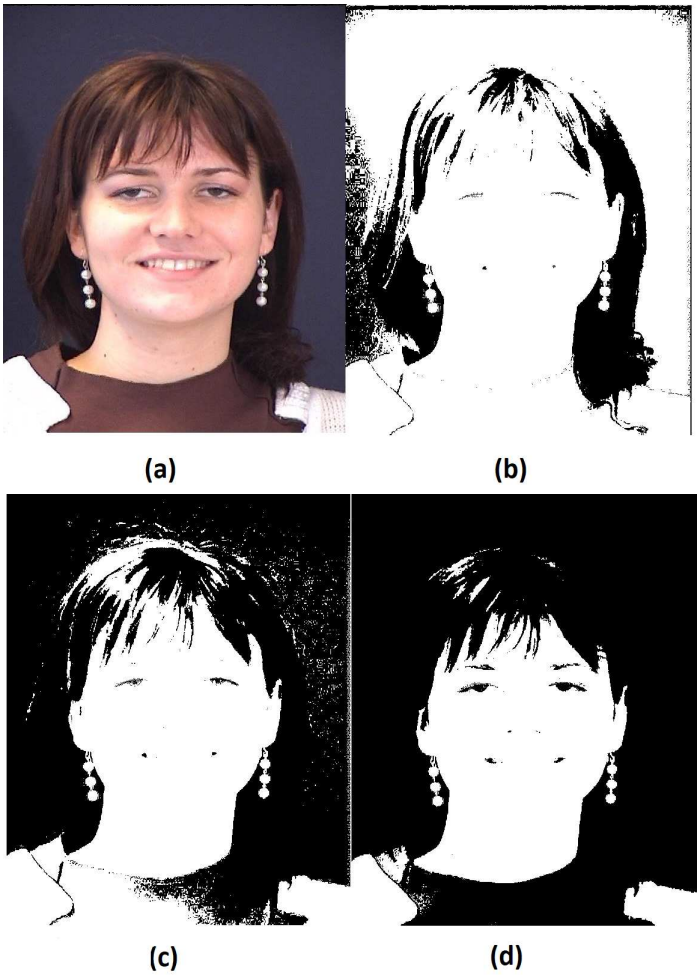

Figura 4.2: Binarização pela luminância: (a) imagem original; (b) $\alpha = 0,5$; (c) $\alpha = 0,7$; (d) $\alpha = 1$.

Binarização Niblack

Um dos mais clássicos e eficientes algoritmos de binarização de imagens em escala de cinza (Niblack 1986). Ele leva em conta a média local do histograma de intensidade de cinza nos *pixels* e seu desvio-padrão na vizinhança de um *pixel*, dentro de uma janela *bxb* pré-estabelecida. Além do tamanho da janela, o peso α que o desvio-padrão terá no cálculo do limiar da luminância também é configurável. Neste trabalho, ao invés de utilizarmos a intensidade de cinza, como é feito tradicionalmente, utilizamos a luminância da imagem em escala de cinza para o cálculo do limiar. A Equação 4.1 mostra como é calculado o limiar:

$$T = \bar{L} + \alpha * \sigma \tag{4.1}$$

A Figura 4.3 mostra exemplos de imagens binarizadas pela binarização Niblack:

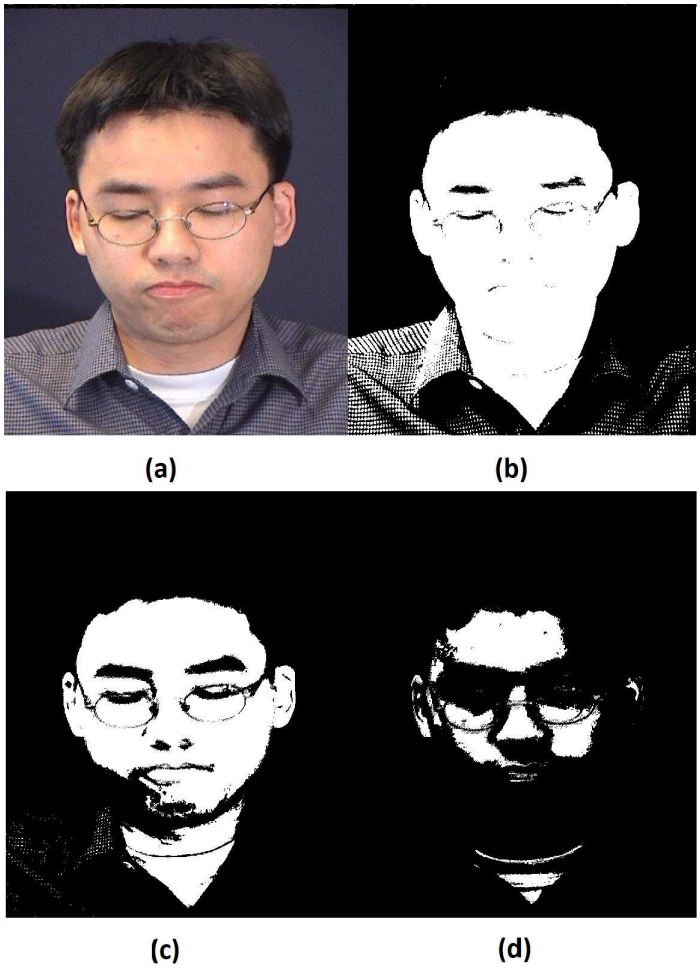

Figura 4.3: Binarização Niblack: (a) imagem original; (b) $\alpha = 0,1$; (c) $\alpha = 0,5$; (d) $\alpha = 0,9$.

Binarização Sauvola

Uma variação da binarização Niblack para imagens obtidas em circunstâncias de má iluminação e, consequentemente, com a presença de alta taxa de ruídos. (Savoula e Pietaksinen 2000) introduz uma variável R, com influência direta no desvio-padrão. Em seus próprios experimentos, Sauvola obteve experimentalmente $R = 128$ como melhor parâmetro. A Equação 4.2 mostra como é calculado o limiar:

$$T = \bar{L}_{local} + 1 + \alpha * \left(\frac{\sigma}{R} - 1\right) \tag{4.2}$$

A Figura 4.4 mostra exemplos de imagens binarizadas pela binarização Sauvola:

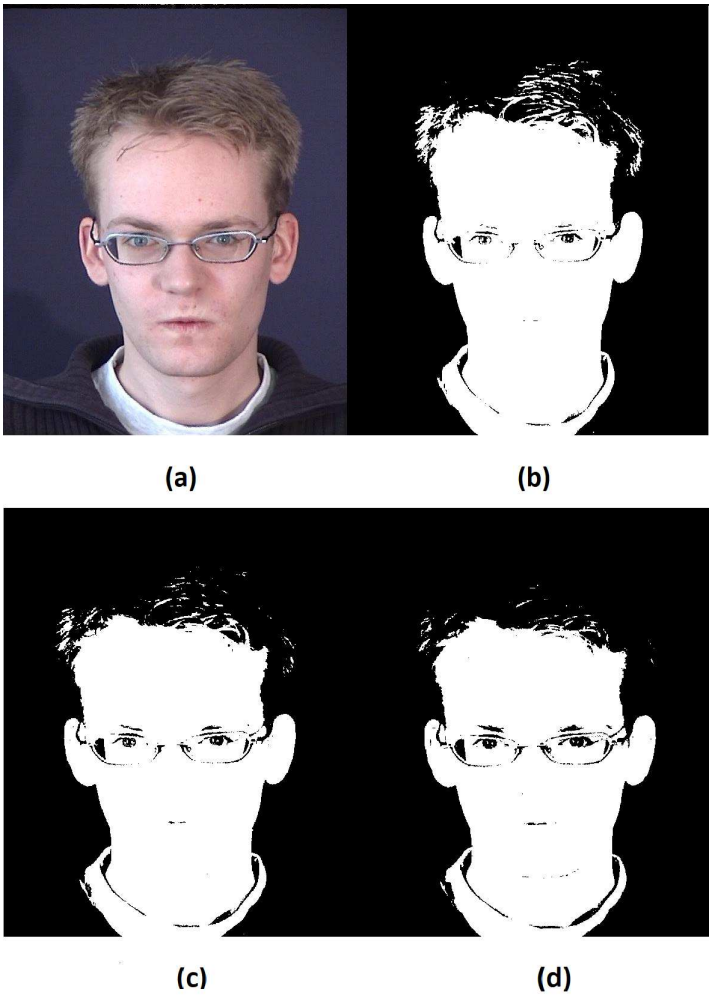

Figura 4.4: Binarização Sauvola: (a) imagem original; (b) $\alpha = 1$; (c) $\alpha = 10$; (d) $\alpha = 20$.

Binarização com Detector de Bordas Canny

Elaborado por John F. Canny (Canny 1986), esta binarização utiliza um algoritmo multi-estágios para identificar todas as bordas possíveis na imagem, de forma que

tais bordas estejam o mais próximas o possível das bordas da imagem original, com cada borda original gerando apenas uma única borda binarizada, de modo que qualquer ruído presente na imagem original não crie bordas binarizadas.

Este filtro é aplicado em cinco etapas:

1)Um filtro gaussiano é aplicado na imagem, de forma que ela seja suavizada e eventuais ruídos sejam removidos;

2)Calcula-se os gradientes de intensidade da imagem;

3)Algum método de supressão não-máxima é aplicado para eliminar bordas ambíguas;

4)Um limite duplo é aplicado para determinar bordas em potencial;

5)Todas as arestas fracas não conectadas a arestas fortes são removidas;

A Figura 4.5 mostra um exemplo de uma imagem binarizada pelo Detector de Bordas Canny:

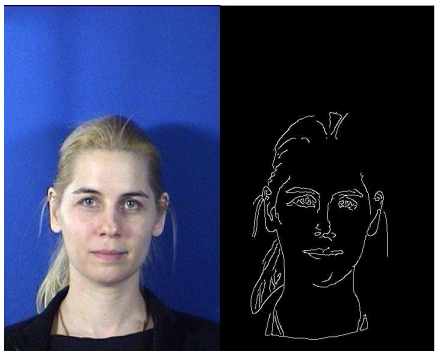

Figura 4.5: Binarização com Detector de Bordas Canny.

4.2.2 Discretização

Discretização refere-se à transformação do espaço de cores contínuo da fotografia original em um espaço discreto. Neste trabalho, esse espaço foi baseado no sistema Lab (Hunter 1948), que é baseado nos parâmetros L (luminosidade), a (espectro verde-vermelho) e b (espectro amarelo-azul). Segue a descrição do processo de discretização da imagem:

1. Para cada *pixel* da fotografia original, converte-se a cor de RGB para o

sistema “CIE 1931 XYZ color space” (Smith e Guild 1931) e então para Lab;

2. Divide-se os intervalos de L, de a e de b em, respectivamente, x, y e z espaços, sendo estes parâmetros previamente selecionados;

3. Todos os *pixels* são coloridos usando os valores centrais de L, a e b dos espaços nos quais sua cor original se encontra inserida;

4. Reverte-se o passo 1, ou seja, os *pixels* voltam a ficar em RGB, uma vez que muitas das cores do sistema Lab estão fora da gama de cores observadas pela visão humana. Neste passo a imagem já está discretizada;

5. Para cada *pixel* da imagem discretizada:

A) Obtém-se os níveis Lab;

B) Estes são convertidos para a codificação unária, de forma a preservar a distância de Hamming, e adicionados à entrada da rede.

A Figura 4.6 mostra exemplos de imagens discretizadas:
O tamanho da entrada binária obtida pela discretização será de $L * a * b$ *bits*.

4.3 Detecção de Faces

Uma vez que apenas a área da face que contém Action Units pode expressar emoções, todo o resto da face é desnecessário para o processo de classificação, fornecendo apenas informação ruidosa a rede. Logo, o segundo estágio do sistema detecta e extrai apenas esta região da imagem original, fornecendo-a ao estágio seguinte. Três abordagens para detecção facial baseadas em WiSARD foram testadas:

A) Treinar uma rede a reconhecer olhos e boca e depois extrair a região delimitada por estes itens;

B) Treinar uma rede a reconhecer toda a região facial;

C) Treinar uma rede com descritores LBP (Local Binary Pattern).

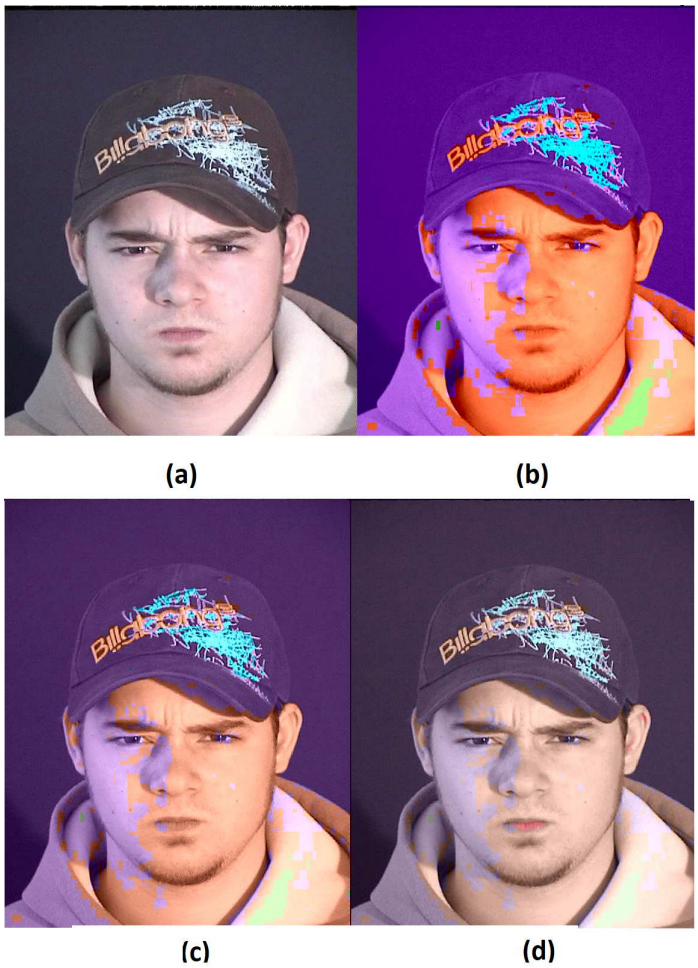

Figura 4.6: Discretização: (a) imagem original; (b)L = 1, a = 2, b = 2; (c)L = 2, a = 4, b = 4; (d)L = 4, a = 8, b = 8.

4.3.1 Detecção baseada em features

A WiSARD de detecção facial possui dois discriminadores, um deles treinado com imagens de olhos e outro com imagens de bocas. Então, quando uma imagem é submetida ao pré-processamento, ela é percorrida por uma janela deslizante, cujo tamanho é a média dos exemplares de olhos da base de treino da rede. Todas as áreas abarcadas pela janela são então pré-processadas e submetidas a classificação pelo discriminador treinado com olhos. As duas áreas da imagem que obtiverem maior placar r neste processo são eleitas como sendo correspondentes aos dois olhos da face, caso sejam simétricos.

Repete-se o processo com uma nova janela, igualmente do tamanho médio das imagens de boca da base de treino, e que percorre a imagem abaixo dos olhos já detectados. As áreas acessadas pela janela são submetidas ao outro discriminador e naturalmente a área de maior pontuação é selecionada como sendo a boca. Depois extrai-se toda região delimitada pelos olhos e boca, desde que estejam dispostas de acordo com o formato natural de uma face: olhos razoavelmente simétricos,

triangularizados com a região da boca, sendo esta situada abaixo da região dos olhos. Em caso de empate entre uma região ou outra, o desempate é feito considerando o placar do entorno. Todo este procedimento é abordado na Figura 4.7.

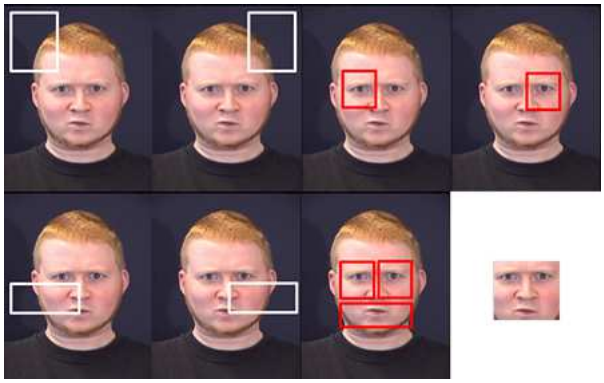

Figura 4.7: Uma janela percorre a imagem procurando os principais candidatos a olhos e boca e, baseado na posição deles, detecta a face na imagem.

4.3.2 Detecção baseada em faces

A WiSARD possui apenas um discriminador, treinado a partir de faces recortadas manualmente. Todo o processo é muito similar a primeira abordagem, como ilustrado pela Figura 4.8, só que muito mais rápido, por possuir duas iterações a menos.

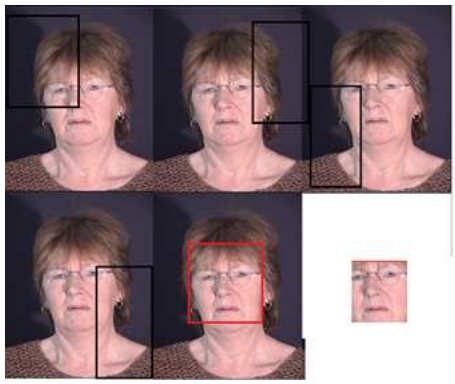

Figura 4.8: Uma janela percorre a imagem procurando o principal candidato a face.

4.3.3 Detecção baseada em descritores LBP

Utiliza-se aqui novamente uma WiSARD com um único discriminador, treinado a partir de descritores LBP. Para obtê-los, divide-se a imagem em N blocos, e em seguida, para cada bloco, analisa 9 *pixels* por vez (uma janela de 3X3). Dentro desta janela, seleciona-se o *pixel* central e compara-se ele com os seus vizinhos, no sentido horário ou anti-horário. O resultado desta comparação será um valor binário.

Caso sua luminância seja superior ao do vizinho, essa comparação contribuirá com o *bit* 1 para o resultado, enquanto o inverso disso com o *bit* 0. Desta forma, após percorrer todas as janelas 3X3 possíveis em um bloco, cria-se um histograma de frequência dos resultados das suas comparações.

Este histograma nada mais será que um vetor de 256 posições, já que o valor mais alto que aparecerá em uma janela será "11111111", caso o *pixel* central possua luminância superior a todos os vizinhos. Depois de criar todos os N histogramas, une-se todos eles em um único vetor, que será a entrada da rede. O algoritmo está representado pela Figura 4.9.

Este tipo de descritor é particularmente útil para detectar variações de textura, sendo robusto contra oclusões e variações da iluminação, além de ser razoavelmente barato computacionalmente e poder ser combinado com outros tipos de descritores, como o HOG. Esta técnica, que compatilha do mesmo cerne teórico do famoso algoritmo de Viola-Jones (que particiona a imagem, atribuindo uma semântica particular a cada seção através de um algortimo multi-estado), foi apresentada originalmente em (Ojala et al. 1994) e ainda é muito utilizada, devido a sua velocidade se comparada com outras abordagens baseadas em Haar Cascade, que por trabalharem com ponto flutuante são computacionalmente custosos e sacrificam a velocidade em nome de acurácia, indo contra um das motivações centrais deste trabalho.

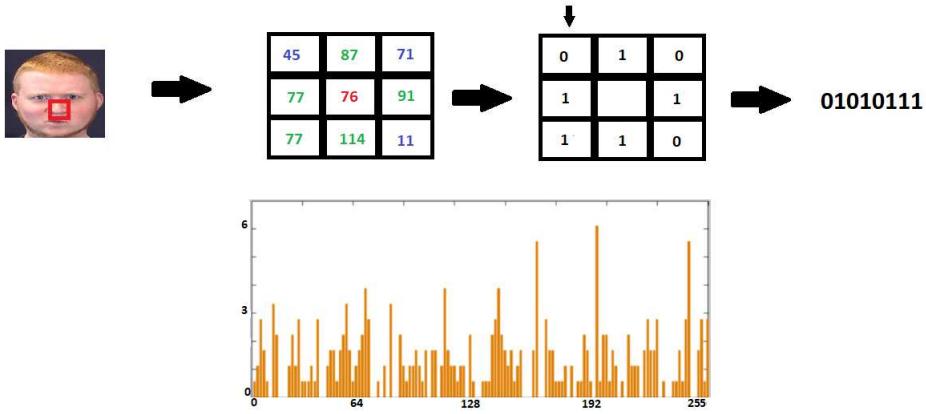

Figura 4.9: Análise de uma janela 3X3 para formação do descritor LBP do seu *pixel* central.

Capítulo 5

Experimentos

Neste capítulo serão relatados os resultados da WiSARD em validações cruzadas nos principais *datasets* de classificação de emoções faciais. Será analisada também a variação da acurácia da WiSARD em relação aos parâmetros responsáveis por sua configuração (o número de RAMs, de entradas nas RAMs e a confiança do resultado esperado). Também serão avaliados o desempenho da rede em termos de tempo de processamento, a busca pela configuração ótima do detector de faces, assim como as imagens mentais geradas pela rede de melhor acurácia. Comparações com o estado-da-arte e outros resultados significativos encontrados na literatura também estão disponíveis.

5.1 Datasets

Os *datasets* escolhidos para este trabalho são aqueles que tem tido maior destaque internacionalmente, sendo frequentemente utilizados como *benchmark* dos algoritmos mais recentes.

5.1.1 Cohn-Kanade Extended Dataset

Base criada pelo Grupo de Análise Afetiva da Universidade de Pittsburgh (Kanade et al. 2000). É formada por 500 sequências de imagens de 100 sujeitos, com idade entre 18 e 30 anos. 65% dos indivíduos é do gênero feminino. 82% dos indivíduos são caucasianos, 15% afro-americanos e 3% latinos e asiáticos. Todas as imagens são completamente frontais e são todas posadas.

Cada série de imagens nesta base é formada por 23 fotografias, sendo que, em cada uma delas, pelo menos uma Action Unit está ativa proeminentemente, podendo haver combinações de várias delas em uma única imagem. A primeira imagem de cada série exibe a emoção neutra ou predominantemente neutra, e ao longo das imagens, alguma outra emoção vai se tornando predominante, até atingir o ápice

da sua expressividade. Cada imagem é anotada com informações sobre a emoção exibida e com as AUs nela expressa. A Figura 5.1 dá exemplos das imagens deste dataset.

Figura 5.1: Exemplares da Cohn-Kanade Extended Dataset (CKP).

5.1.2 MMI Database

Produzido por (Pantic et al. 2005). Possui 52 indivíduos, de 19 a 62 anos, sendo 52% do gênero masculino. A base é dividida entre europeus, asiáticos e sul-americanos, e as imagens foram obtidas com iluminação natural, nos mais diversos *backgrounds*. Existem neste *dataset* imagens posadas e espontâneas, assim como vídeos, ambos frontais e laterais.

Foram registradas 79 sequências de expressões de cada indivíduo, tanto em vídeo quanto em fotografia. Tais sequências começam e terminam na emoção neutra, podendo ser entremeadas por alguma outra emoção ou não. Todos os arquivos são anotados com informações do voluntário, AUs (no caso de vídeo há também informações relacionados aos *frames* onde são manifestados) e emoção exibidas. A Figura 5.2 dá exemplos das imagens deste dataset.

5.2 Experimentos prévios

As implementações da WiSARD foram feitas tanto em Java, quanto em Python 3, para efeitos de comparação da performance. Experimentos prévios foram realizados a fim de estimar dados relevantes da rede e calibrar os parâmetros da rede utilizada na Validação Cruzada.

5.2.1 Parâmetros

Ao se instanciar uma rede, dois parâmetros são fundamentalmente importantes: o tamanho dos nós-RAMs, ou seja, a quantidade de *bits* de endereçamento utilizados

Figura 5.2: Exemplares da MMI Database. Como se pode perceber, as imagens tem formatos distintos.

pelos discriminadores e o valor da confiança da rede.

5.2.2 Retina

A mesma retina foi utilizada para todos os discriminadores do classificador, de forma que as imagens foram redimensionadas para o tamanho de 120 X 160 *pixels* na fase de pré-processamento, de forma que os discriminadores possuem 19200 linhas de endereçamento.

Tamanho das RAMs

Quanto maiores forem os nós-RAMs, menor será a quantidade de neurônios na rede, uma vez que a quantidade de $N * n$ *bits* endereçados pela rede não poderá mudar, devido a imutabilidade da retina. Ou seja, quanto maior forem os RAMs, mais generalista será a rede, e vice-versa. Nos experimentos prévios, variou-se o valor de n para encontrar aquele que produzia a rede mais acurada. Ao observar o comportamento de n, automaticamente já o estamos fazendo para N, que será obtido da divisão do tamanho da retina pela quantidade de *bits* de endereçamento dos nós-RAM.

Confiança

Verificou-se que, conforme a confiança da resposta da rede se aproximava de 5%, rapidamente a rede se tornava incapaz de responder adequadamente, classificando todas as emoções como neutras, uma vez que essa é a resposta padrão da rede caso todos os discriminadores obtenham $r = 0$, evento normalmente causado pela ação

do *bleaching* devido a sucessivos empates na fase de classificação. Então, para os testes abaixo descritos fixou-se a confiança em 1%.

Bleaching

As redes implementadas nos experimentos aqui descritos tem seu *bleaching* incrementado em apenas uma unidade para cada empate na fase de classificação.

5.2.3 Detector de Faces

Um subconjunto com 100 imagens do MMI Database sorteadas aleatoriamente foi utilizado para se obter a melhor configuração para o detector de faces. Testou-se aqui as três abordagens já discutidas para a detecção facial: (A) rede treinada com olhos e bocas (dois discriminadores); (B) rede treinada com faces (um discriminador); (C) rede treinada com descritores LBP (um discriminador).

Nas abordagens (A) e (B), a imagem era limiarizada através da sua luminância média (com $\alpha = 0,7$), de forma a se obter faces de 120X160. A seção da imagem detectada como sendo uma face era depois extraída da imagem original, para que pudesse receber outro tipo de pré-processamento na fase seguinte do processo de classificação emocional.

A avaliação da performance dos detectores de face (Tabela 5.1) foi realizada empiricamente, considerando-se como uma face detectada corretamente, quando toda região possuidoras de AUs havia sido capturada pelo detector e a mesma se encontrava centralizada na imagem. Percebe-se que a abordagem que utiliza a WiSARD com dois discriminadores era especialmente sensível a presença de oclusões, óculos e barba. A abodargem baseada em LBP teve desempenho bem superior aos outros detectores.

Uma vantagem da abordagem (B) foi em relação a performance, demorando um tempo médio de 0,41s para detectar uma face (média obtida a partir de 10000 exemplos), enquanto o modelo (A) obteve 0,66s e o (C) obteve 0,91s. Vale ressaltar que a abordagem (B) empatou com a velocidade média das implementações comerciais de detectores faciais rodando em um mesmo CPU dos testes realizados, segundo o *benchmark* exposto em (Soyata e Powers 2016). A Figura 5.3 dá exemplos de faces detectadas pelo modelo sem peso, treinado a partir de imagens de faces (abordagem B).

5.2.4 Binarização X Discretização

A Figura 5.4 mostra os resultados experimentais em uma Validação Cruzada com 10 blocos no MMI Database (base completa), comparando a influência de ambas as

Tabela 5.1: Quantidade de faces extraídas de forma completamente correta pelos diferentes detectores.

Tamanho da rede	Detector A (%)	Detector B (%)	Detector C(%)
20	8	10	35
50	10	11	40
80	17	21	51
100	21	25	55
120	27	36	67
150	41	52	78
175	37	50	73
200	35	50	71
250	31	47	71

Figura 5.3: Faces detectadas com a WiSARD treinada com imagens de faces.

técnicas na acurácia da rede, em um gráfico *bits* de endereçamento X acurácia.

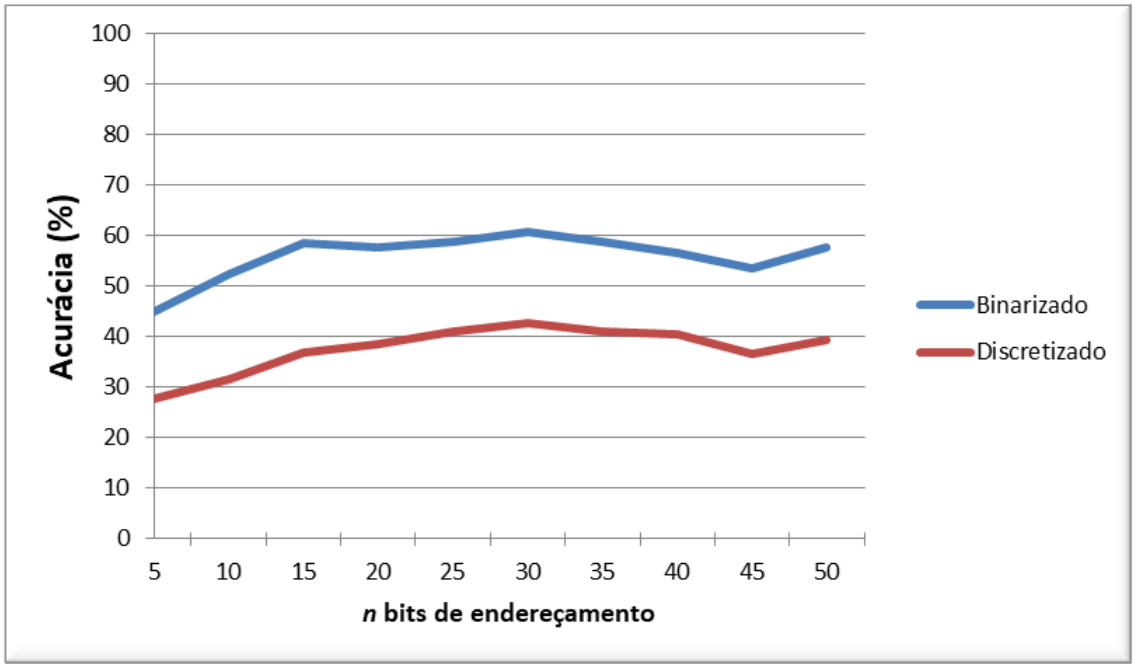

Figura 5.4: Comparação da acurácia da rede utilizando binarização e discretização como técnica de pré-processamento.

Nota-se que, nesta comparação, foi utilizado o tipo mais básico de binarização, aquela baseada meramente em luminância média (com $\alpha = 0,7$, o melhor valor obtido experimentalmente para o escalar do limiar neste tipo de pró-processamento).

A discretização utilizada nesta comparação foi feita com os melhores valores de Lab encontrados experimentalmente, sendo L = 4, a = 2, b = 2.

Enquanto a binarização filtra bordas relevantes de uma imagem, objetivando preservar contornos geométricos semanticamente substanciais, a discretização preserva informações pertinentes ao conteúdo cromático da imagem original, ao mesmo tempo que reduz significativamente o tamanho da entrada.

No entanto, como a única informação realmente decisiva para a classificação de uma emoção facial é a presença de certas Action Units, que por sua vez são determinadas unicamente pelo contorno das *features* faciais, sem qualquer ligação com a cor dos *pixels* que a preenchem, a binarização acaba por ser a técnica de pré-processamento mais sintética e menos ruidosa para este tipo de finalidade.

5.2.5 Desempenho da rede

Na implementação em Java, o tempo de treinamento médio da rede por imagem (medido com 10000 imagens em um computador com processador Intel i7 e Windows 10) foi de 0,01s e o tempo médio de classificação foi de 0,07s. Na implementação em Python, utilizando o interpretador Pypy, o tempo de treinamento médio da rede foi de 0,04s e o tempo médio de classificação foi de 0,21s. Nesta medição não foi computado o tempo de detecção de faces.

5.3 Validação Cruzada

A Validação Cruzada em ambos *datasets* foi feita com 10 blocos. Como a binarização pela luminância e a Niblack são a base para a implementação da binarização Sauvola (que reune as vantagens destes métodos com seu tratamento anti-ruído), elas não foram testadas nesta validação. A binarização de Sauvola utilizada nestes experimentos foi baseada em janelas de 30X80 *pixels*. Para detecção de faces, foram utilizados tanto o detector B, quanto o C nesta validação.

5.3.1 Cohn Kanade Extended

Todas as 5876 imagens anotadas com emoção foram utilizadas. O estado-da-arte para este *dataset* (Zafer et al. 2013) obteve 100% de acurácia, usando apenas imagens seletas e sem qualquer tipo de Validação Cruzada. A Figura 5.5 mostra a comparação do desempenho das entradas pré-processadas pela binarização de Sauvola e pelo detector de bordas Canny com a variação da quantidade de entradas nas RAMs (n). A Tabela 5.2 compara a solução aqui apresentada com o estado-da-arte e outros resultados relevantes, a Tabela 5.3 e 5.4 são as matrizes de confusão obtidas no

experimento aqui relatado, tanto para a detecção de faces baseada em amostragem de faces, tanto quanto para a solução baseada em descritores LBP, respectivamente.

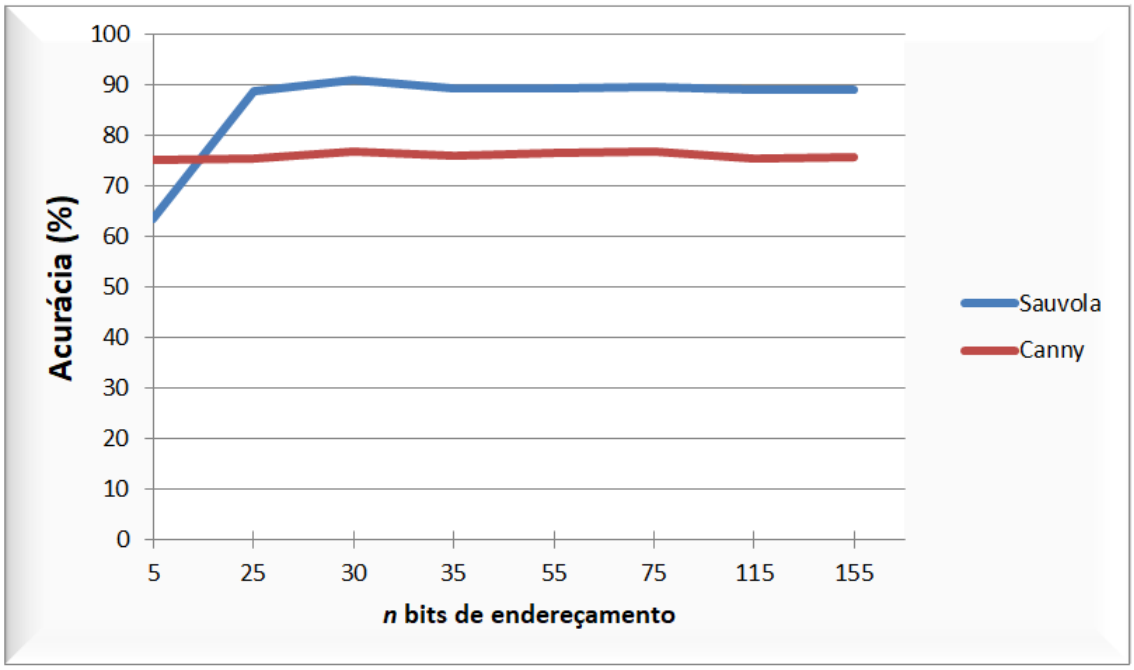

Figura 5.5: Resultados da validação cruzada no CKP, com entradas pré-processadas pela binarização de Sauvola e pelo detector de bordas Canny. O desempenho vencedor foi a rede cujos nós-RAMs endereçam 50 posições de memória e as entradas foram pré-processadas com binarização Sauvola.

Tabela 5.2: O atual estado-da-arte em reconhecimento de emoções no CKP. A acurácia da WiSARD em DAF1 foi 90,01%, com um desvio-padrão de 0,6%, e em DAF2 foi 97,3%, com desvio-padrão de 0,7; Legenda - DAF1: Detecção Automática de Faces utilizando o detector B; DAF2: Detecção Automática de Faces utilizando o detector C; DMF: Detecção Manual de Faces; VC: Validação Cruzada; LOO: leave one out (em cada iteração deixa um exemplo apenas para ser classificado).

Autor	Metódo	Dataset	Validação	Acurácia (%)
(Zafer et al. 2013)	NCC	Parcial	LOO	**100**
(Burkert et al. 2015)	NCC	Completo	VC 10-blocos	99,6
(Lopes et al. 2016)	NCC	Completo	VC 8-blocos	96,76
(Happy and Routray 2015)	SVM	Parcial	VC 10blocos	94,09
(Kotsia et al. 2008)	Multiclass SVM	Completo	LOO	91,6
WiSARD (proposto)	DAF1	Completo	VC 10-blocos	**90,01**
WiSARD (proposto)	DAF2	Completo	VC 10-blocos	**97,01**
WiSARD (proposto)	DMF	Completo	VC 10-blocos	**100**

A maioria das imagens erroneamente classificadas nesse *dataset* retornou a emoção "neutra" como resultado. Isso ocorreu quase inteiramente em imagens cujo rosto foi mal detectado e, portanto, a parte da imagem selecionada para representá-lo não obteve um placar suficientemente satisfatório em qualquer discriminador na

Tabela 5.3: A matriz de confusão de uma validação cruzada com 10 blocos com WiSARD (DAF1) utilizando o dataset CKP.

	Neutro	Felicidade	Tristeza	Medo	Raiva	Repulsa	Surpresa
Neutro	**0,927**	0,012	0,013	0,005	0,019	0,009	0,012
Felicidade	0,086	**0,906**	0,004	0	0	0,002	0,001
Tristeza	0,156	0,003	**0,84**	0	0	0	0
Medo	0,091	0,002	0	**0,891**	0	0,005	0,01
Raiva	0,143	0	0,008	0	**0,841**	0,003	0,001
Repulsa	0,138	0,002	0	0,002	0,007	**0,848**	0,002
Surpresa	0,109	0,001	0,001	0,004	0,001	0,004	**0,88**

Tabela 5.4: A matriz de confusão de uma validação cruzada com 10 blocos com WiSARD (DAF2) utilizando o dataset CKP.

	Neutro	Felicidade	Tristeza	Medo	Raiva	Repulsa	Surpresa
Neutro	**0,981**	0,07	0,006	0,003	0,001	0,001	0,001
Felicidade	0,02	**0,977**	0,002	0,001	0	0	0
Tristeza	0,02	0,004	**0,976**	0	0	0	0
Medo	0,04	0	0	**0,95**	0	0,01	0
Raiva	0,03	0	0,002	0	**0,967**	0,001	0
Repulsa	0,039	0	0	0	0,01	**0,951**	0
Surpresa	0,049	0	0	0	0	0,08	**0,943**

fase de classificação, de modo que, gradualmente, o *bleaching* reduziu a pontuação de todos os discriminadores para 0, fazendo com que a rede retornasse a classificação padrão "neutro". Quando a detecção facial manual foi aplicada, a precisão de 100% foi alcançada. Pode-se ver que a detecção facial, apesar de um problema distinto, é de relevância crucial para a classificação das emoções.

5.3.2 MMI Database

Todas as imagens e vídeos desta base que têm emoções anotadas foram usados na validação. Foram extraídos 50 *frames* de cada vídeo da base, com os 15 primeiros associados à emoção "neutra" e os outros com sua própria anotação. A precisão obtida foi comparada com (Burkert et al. 2015) e com (Wang e Yin 2007), que também usaram a base para o reconhecimento das emoções, em vez de seu uso tradicional na detecção de AUs. A Figura 5.6 mostra a comparação da acurácia da rede na classificação, quando as entradas são pré-processadas pela binarização de Sauvola e pelo detector de bordas Canny, com a variação da quantidade de entradas nas RAMs (*bits* de endereçamento). A Tabela 5.5 compara a solução aqui apresentada com o estado-da-arte e outros resultados relevantes e a Tabela 5.6 é a matriz de confusão obtida no experimento aqui relatado para a detecção de faces baseada em amostragem de faces.

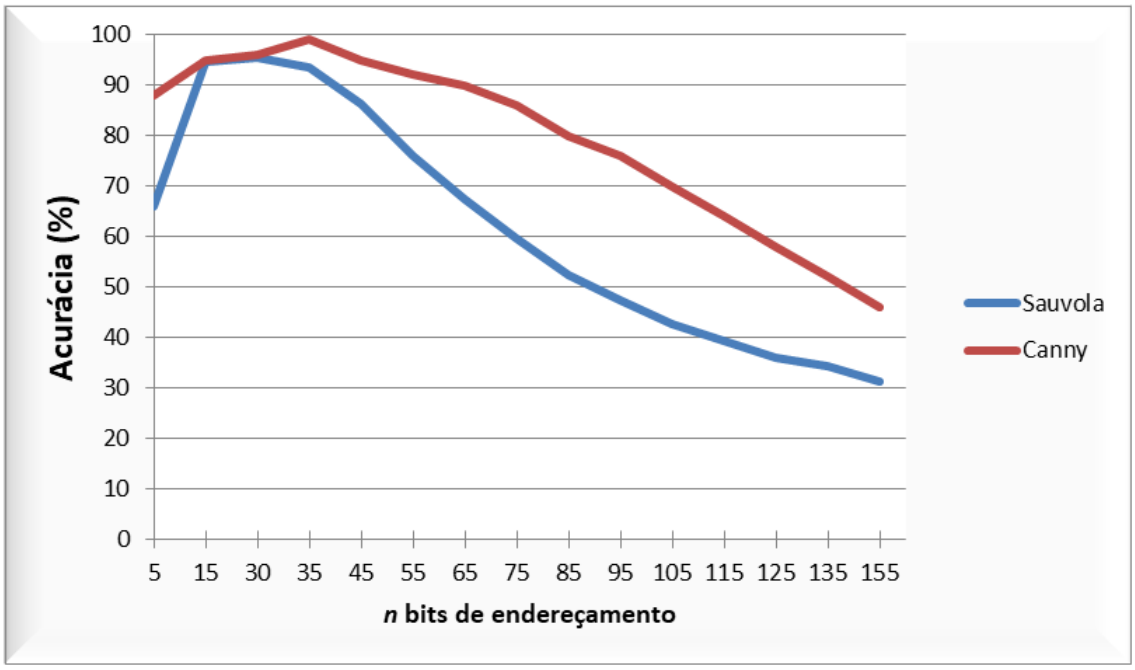

Figura 5.6: Resultados da validação cruzada no MMI, com entradas pré-processadas pela binarização de Sauvola e pelo detector de bordas Canny. O desempenho vencedor foi a rede cujos nós-RAMs endereçam 50 posições de memória e as entradas foram pré-processadas com detector de bordas Canny.

Tabela 5.5: Atual estado-da-arte em reconhecimento de emoções no MMI. A acurácia da WiSARD foi 99,3%, com um desvio-padrão de 0,1%.

Autor	Metódo	Dataset	Validação	Acurácia (%)
(Burkert et al. 2015)	CNN	Completo	VC 10 blocos	**98,63**
(Wang e Yin 2007)	LDA	Completo	LOO	93,33
(Wang e Yin 2007)	QDC	Completo	LOO	92,78
(Wang e Yin 2007)	NBC	Completo	LOO	85,56
WiSARD (proposto)	DAF1	Completo	VC 10 blocos	**99,3**
WiSARD (proposto)	DAF2	Completo	VC 10 blocos	**99,3**
WiSARD (proposto)	DMF	Completo	VC 10 blocos	**99,4**

Tabela 5.6: A matriz de confusão de uma validação cruzada de 10 blocos com a WiSARD (DAF1) utilizando o MMI dataset.

	Neutro	Felicidade	Tristeza	Medo	Raiva	Repulsa	Surpresa
Neutro	**0,994**	0,004	0,002	0	0	0	0
Felicidade	0,016	**0,973**	0	0	0,001	0,001	0,008
Tristeza	0,002	0	**0,993**	0	0	0,004	0
Medo	0,017	0	0	**0,979**	0	0	0,003
Raiva	0,011	0,002	0	0	**0,972**	0,014	0
Repulsa	0,008	0	0	0	0	**0,987**	0,005
Surpresa	0,01	0,004	0,001	0,001	0	0,004	**0.976**

A maioria das imagens mal classificadas aqui são aquelas que estão na transição entre o estado neutro original e a emoção exibida no vídeo de onde esse *frame* foi retirado, de modo que a emoção ainda não atingiu um grau de expressividade suficiente.

5.3.3 Críticas aos datasets

Ambos *datasets* apresentam imagens com anotações duvidosas. Isso se reflete especialmente no MMI, onde parte das anotações foi gerada pela comunidade. Alguns exemplos de imagens erroneamente classificadas pela nossa solução no MMI Database são ilustrados pela Figura 5.7 e descritos abaixo:

(A) *frames* finais de S032-010 na sessão 1818, anotados com a emoção "surpresa", exibem um indivíduo com lábios levemente abertos e levantados, demonstrando "felicidade";
(B) *frames* finais de S028-002 na Sessão 1867, anotados com a emoção "repulsa", exibem um indivíduo com uma boca ligeiramente aberta e nenhuma linha de expressão efetiva é exibida acima dos olhos, fazendo com que a imagem pareça "neutra";
(C) os *frames* intermediários de S045-011 na sessão 1927 foram anotados com a emoção "raiva", mas a maioria deles exibe um indivíduo com olhos fechados e uma cabeça ligeiramente inclinada, de modo que pareça estar experimentando "tristeza".

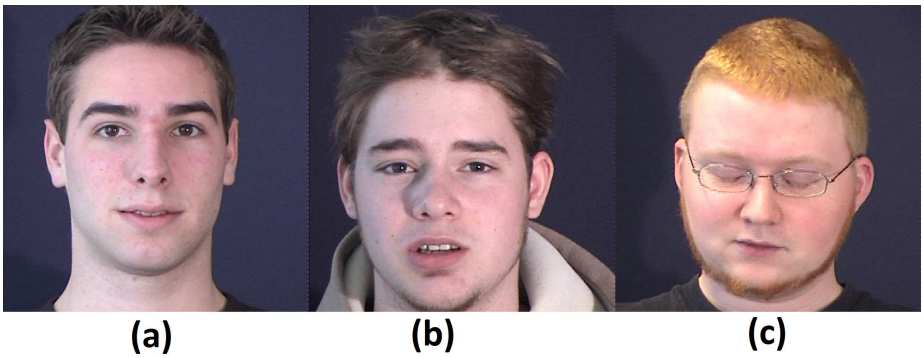

Figura 5.7: Imagens do MMI Database, cujas rotulações não condizem com a emoção exibida: (a) "surpresa", (b) "repulsa", (c) "raiva"

5.3.4 Binarização Sauvola X Detector de bordas Canny

Como se observou nas últimas seções, a binarização Sauvola se mostrou mais eficiente no *dataset* CKP, enquanto a abordagem de Canny foi melhor no MMI. Uma explicação possível para este comportamento é que o CKP possui muito mais ruído que o MMI, e que nestes casos o alto valor de desvio-padrão exigido pelo detector de bordas Canny para eliminar bordas espúrias acabou suavizando exageradamente

a imagem, levando ela a perder informações relevantes. Nos demais casos, este pré-processamento foi superior a binarização Sauvola.

5.4 Observando AUs nas imagens mentais

A fim de obter resultados qualitativos da capacidade de aprendizagem da WiSARD, imagens mentais das emoções faciais básicas foram geradas a partir de uma amostra aleatória de 20% do *dataset* MMI. As imagens mentais geradas pela DRASiW neste experimento são exibidas da Figura 5.8 até a Figura 5.14.

Figura 5.8: Imagem mental do discriminador "Neutro".

Figura 5.9: Imagem mental do discriminador "Felicidade".

Figura 5.10: Imagem mental do discriminador "Tristeza".

Figura 5.11: Imagem mental do discriminador "Medo".

Uma observação informal e empírica foi realizada nessas imagens mentais, conjuntamente com a tabela EMFACS, e procurou-se identificar nelas as AUs mais intensas e, se de fato, elas correspondiam aquelas esperadas por aquelas emoções específicas. Um estudo mais completo e criterioso está sendo desenvolvido neste sentido, com o uso de uma rede neural sem peso treinada para detectar AUs.

Figura 5.12: Imagem mental do discriminador “Raiva”.

Figura 5.13: Imagem mental do discriminador “Repulsa”.

Figura 5.14: Imagem mental do discriminador “Surpresa”.

Como exibido pela Figura 5.15, a imagem mental do discriminador “Neutro” possui com nitidez as AUs 41 (relaxamento parcial do músculo levantador da pálpebra superior), 42 (ativação do músculo orbicular do olho) e 44 (ativação simultânea dos músculos corrugador do supercílio e orbicular do olho), todos presentes na emoção “neutra”, e notável ausência de qualquer AU da família 1-28, que necessariamente explicitariam uma emoção não-neutra.

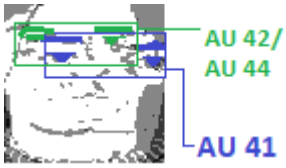

Figura 5.15: AUs verificadas na imagem mental do discriminador “Neutro”.

Como exibido pela Figura 5.16, a imagem mental do discriminador “Felicidade” possui com nitidez a AU 6 (suspensão da musculatura da bochecha), que só é ativado nas emoções “felicidade” e “tristeza”, AU 13 (ativação do músculo levantador do ângulo da boca), ativado só na emoção “felicidade”, e AU 26 (simultânea contração do músculo masseter e relaxamento dos músculos temporal e pterigóideo interno), comunal a todos as emoções.

Como exibido pela Figura 5.17, a imagem mental do discriminador “Tristeza” possui com nitidez a AU 6, que só é ativado nas emoções “felicidade” e “tristeza”.

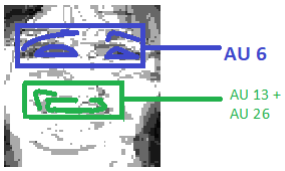

Figura 5.16: AUs verificadas na imagem mental do discriminador “Felicidade”.

Figura 5.17: AUs verificadas na imagem mental do discriminador “Tristeza”.

Como exibido pela Figura 5.18, a imagem mental do discriminador “Medo” possui com nitidez a AU 2 (contração do músculo frontal), que só é ativado nas emoções “medo” e “surpresa”, e a AU 20 (ativação simultânea do músculo risório com o músculo plástima), ativado apenas na emoção “medo”.

Figura 5.18: AUs verificadas na imagem mental do discriminador “Medo”.

Como exibido pela Figura 5.19, a imagem mental do discriminador “Raiva” possui com nitidez a AU 5 (ativação simultânea do músculo levantador da pálpebra superior com o músculo tarsal superior), ativado nas emoções “raiva”, “medo” e “surpresa”.

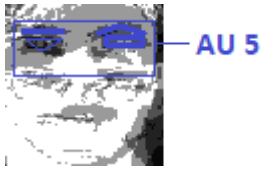

Figura 5.19: AUs verificadas na imagem mental do discriminador “Raiva”.

Como exibido pela Figura 5.20, a imagem mental do discriminador “Repulsa” possui com nitidez a AU 9 (contração do músculo levantador do lábio superior e da asa do nariz), que só é ativado nas emoção “repulsa”. A AU 2 também aparece com

relativa nitidez. Tal AU está presente nas emoções “medo” e “surpresa”, mas não na emoção “repulsa”. Isso justifica o fato de “surpresa” ter sido a única classificação errada para a emoção “repulsa” neste *dataset*, com exceção da emoção “neutra”, nos casos onde o *bleaching* causou empate absoluto entre os discriminadores.

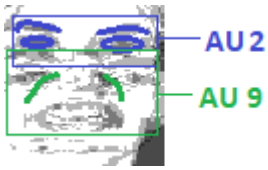

Figura 5.20: AUs verificadas na imagem mental do discriminador “Repulsa”.

Como exibido pela Figura 5.21, a imagem mental do discriminador “Surpresa” possui com nitidez a AU 1 (erguimento da parte interna da sombrancelha), ativado nas emoções “tristeza”, “medo” e “surpresa”, e a AU 26, comunal a todas as emoções.

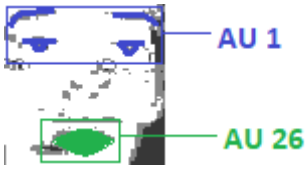

Figura 5.21: AUs verificadas na imagem mental do discriminador “Surpresa”.

Capítulo 6

Conclusão

Neste trabalho foi apresentado uma solução orientada a redes neurais sem peso para a questão da classificação de emoções em faces. Parte do processo de pesquisa envolveu o teste de diferentes formas de pré-processamento de imagens em entradas binárias, por serem estas as únicas compatíveis com o modelo de rede neural aqui estudado, e, verificou-se que técnicas de binarização eram proeminentes em relação a discretização do espaço de cores. Uma vez que o indicador de emoção em uma face é o conjunto de músculos ativos nela, este tipo de informação era melhor representado por técnicas de binarização, especialmente aquelas que contém filtros contra ruídos.

Testou-se também um processo para detecção de faces baseado totalmente em WiSARD. A técnica se mostrou muito rápida e particularmente eficiente em faces totalmente posicionadas frontalmente e com boa luminosidade, mas muito sensível a imagens com luminosidade ruim e com rotações na face, de forma semelhante ao que é historicamente relatado com redes com peso que não utilizam técnicas de ajuste de *features*.

Ao ser submetida aos principais *datasets* encontrados na literatura, a solução proposta provou ser eficaz em termos de acurácia e obteve resultados fortemente competitivos em relação ao atual estado-da-arte. Como o código das demais soluções relatadas não estão disponíveis, nem existem aplicativos públicos que as implementem, não foi possível comparar o desempenho delas em termos de velocidade.

A rede aqui sugerida obteve resultados extremamente significativos nos *datasets* onde foi aplicada, perdendo apenas um pouco de acurácia devido a falhas na detecção facial. Quando os mesmos exemplares que haviam sido classificados erroneamente tiveram a face extraída manualmente, o desempenho da rede se mostrou totalmente equiparado, demonstrando assim a influência da detecção de faces na solução do problema e a capacidade de classificação da WiSARD.

A implementação da WiSARD usada nestas validações, assim como os recursos de pré-processamento e os casos de teste estão todos disponíveis em: https://github.com/Lusquino/WisardEmocoes. Destaca-se aqui a velocidade de

aprendizado da WiSARD, sendo este o ponto forte desta solução, permitindo que ela seja utilizada em sistemas embarcados e outros que necessitem ser treinados em tempo-real, como ambientes online.

Por fim, a possibilidade de se representar graficamente o conhecimento contido no discriminador WiSARD através do processo DRASiW se mostrou muito eficaz na validação dos resultados da rede, uma vez que as imagens mentais geradas apresentavam algumas AUs bem definidas das classes por elas representadas. Os erros mais comuns de classificação também foram explicados por este processo, uma vez que certas AUs estranhas a emoção proprietária das imagens mentais também foram identificadas e sua classificação no EMFACS coincide com os resultados errôneos da rede, demonstrando assim o poder da DRASiW para extração de regras e representação do conhecimento adquirido pela rede.

Alguns trabalhos futuros são:

- Melhoria da detecção de faces através da identificação de *landmarks*, descritores HOG ou de alguma abordagem inspirada que torne a rede tolerante a rotação, translação e variações na iluminação;
- Detecção e classificação de AUs, assim como a classificação de emoções utilizando as regras da EMFACS;
- Detecção e classificação de micro-expressões em vídeos, a partir da análise temporal das emoções e da presença de AUs contraditórias.

Referências Bibliográficas

[1] 1606-1656. URL http://www.library.northwestern.edu/spec/hogarth/.

[2] 2018. URL https://en.wikipedia.org/wiki/Neuron.

[3] A. Alattar and S. Rajala. Facial features localization in front view head and shoulders images. *IEEE International Conference on Accoustics, Speech and Signal Processing*, page 3557–3560, 1999.

[4] Igor Aleksander. Self-adaptive universal logic circuits. *IEE Electronic Letters*, 2:321, Agosto 1966.

[5] Igor Aleksander, W. V. Thomas, and P. A. Bowden. WISARD: A radical step foward in image recognition. *Sensor Review*, 4:120–124, Julho 1984.

[6] Igor Aleksander, M. De Gregorio, F. M. G. França, P. M. V. Lima, and H. Morton. A brief introduction to weightless neural systems. *ESANN*, page 299–305, Abril 2009.

[7] Lawrence Cruvinel Bandeira. Nc-wisard: Uma interpretação sem pesos do modelo neural neocognitron. Master's thesis, UFRJ, Rio de Janeiro, 2010.

[8] V. Bettadapura. Face expression recognition and analysis: The state of the art. 2012.

[9] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *CoRR*, abs/1509.05371, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1509.html#BurkertTADL15.

[10] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8:679–714, 1986.

[11] D. O. Cardoso, Danilo S. Carvalho, Daniel S. F. Alves, Diego F. P. Souza, Hugo C. C. Carneiro, Carlos E. Pedreira, Priscila M. V. Lima, and Felipe M. G. França. Financial credit analysis via a clustering weightless neural classifier. *Neurocomputing*, 183:70–78, 2016.

[12] Douglas O. Cardoso. Uma arquitetura para agrupamento de dados em fluxo contínuo baseada em redes neurais sem pesos. Master's thesis, UFRJ, Rio de Janeiro, 2012.

[13] Hugo César Castro Carneiro. A função do Índice de síntese das linguagens na classificação gramatical com redes neurais sem peso. Master's thesis, UFRJ, Rio de Janeiro, 2012.

[14] J.F. Cohn, Z. Ambadar, and P. Ekman. *The handbook of emotion elicitation and assessment, Oxford University Press Series in Affective Science*. J.A. Coan and J.B. Allen, Oxford, 2005.

[15] R. Cowie, E. Douglas-Cowie, K. Karpouzis, G. Caridakis, M. Wallace, and S. Kollias. *Multimodal User Interfaces*. Springer Berlin Heidelberg, Heidelberg, 2008.

[16] Charles Darwin. *The Expression of the Emotions in Man and Animals*. J. Murray, London, 1904.

[17] Leopoldo A. D. Lusquino Filho, Felipe Maia Galvão França, and Priscila Machado Vieira Lima. Near-optimal facial emotion classification using a wisard-based weightless system. *ESANN*, Abril 2018.

[18] Maximo De Gregorio. On the reversibility of multi-discriminator systems. Technical Report 125/97, 1997.

[19] B. P. A. Grieco, Priscila M. V. Lima, Massimo De Gregorio, and Felipe M. G. França. Producing pattern examples from "mental"images. *Neurocomputing*, 73:1057–1064, 2010.

[20] S. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans on Affective Computing 6(1)*, page 1–12, Janeiro 2015.

[21] S. Haykin. *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, Upper Saddle River, New Jersey, Julho 1998.

[22] Richard Sewal Hunter. Photoelectric color-difference meter. *Journal of Optical Society of America*, page 661, Julho 1948.

[23] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Proc. IEEE Intl Conf. Face and Gesture Recognition (AFGR 00)*, page 46–53, 2000.

[24] C. Koch and T. Poggio. *Synaptic Function*. John Wiley and Sons, New Jersey, 1987.

[25] Irene Kotsia, Ioan Buciu, and Ioannis Pitas. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing 26(7)*, page 1052–1067, Julho 2008.

[26] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, page 2278–2324, 1998.

[27] Andre Teixeira Lopes, Edilson de Aguiar, Alberto F. De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 2016. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2016.07.026. URL `http://www.sciencedirect.com/science/article/pii/S0031320316301753`.

[28] W. Niblack. *An Introduction to Image Processing*. Prentice Hall, Englewood Cliffs, New Jersey, 1986.

[29] T. Ojala, M. Pietikäinen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR 1994)*, page 582–585, 1994.

[30] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Webbased database for facial expression analysis. *IEEE Intl Conf. Multimedia and Expo (ICME 05)*, page 317–321, Julho 2005.

[31] Rosalind Piccard. Affective computing. MIT Technical Report no.321 (Abstract), 1995.

[32] Michael A. Sayette, Jeffrey F. Cohn, Joan M. Wertz, Michael A. Perrott, and Dominic J. Parrott. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, 25, no. 3:167–185, 2001.

[33] C. M. Soares. Uma implementação em software do classificador wisard. *SBRN*, 5:225–229, 1998.

[34] Michael Stokes, Matthew Anderson, Srinivasan Chandrasekar, and Ricardo Motta. A standard default color space for the internet - srgb., 1996. URL `https://www.w3.org/Graphics/Color/sRGB.html`.

[35] Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology.*, 2017. doi: 10.17605/OSF.IO/HV28A. URL `https://psyarxiv.com/hv28a/`.

[36] Ming-Hsuan Yang, David J. Kriegman, and Narenda Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, Janeiro 2002.

[37] A. Zafer, R. Nawaz, and J. Iqbal. Face recognition with expression variation via robust ncc. *IEEE 9th Int Conf in Emerging Technologies (ICET)*, page 1–5, Dezembro 2013.