Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Juan Soto Vega

Reconhecimento Facial em Vídeo com uma amostra por pessoa utilizando *Stacked Supervised Auto-encoder*

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientadores: Prof. Raul Queiroz Feitosa
Prof. Patrick Nigri Happ

Rio de Janeiro

Agosto de 2016

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Pedro Juan Soto Vega

Reconhecimento Facial em Vídeo com uma amostra por pessoa utilizando *Stacked Supervised Auto-encoder*

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Raul Queiroz Feitosa
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Patrick Nigri Happ
Co-Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Álvaro Veiga Filho
Departamento de Engenharia Elétrica – PUC-Rio

Profª. Cristina Nader Vasconcelos
Universidade Federal Fluminense

Prof. Márcio da Silveira Carvalho
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 30 de agosto de 2016

Pedro Juan Soto Vega

Nasceu em Guantánamo, Cuba, em 1987. Em 2011 obteve seu diploma de graduação pela Universidade de Oriente na especialidade de Engenharia em Telecomunicações e Eletrônica. Atualmente é aluno de mestrado no programa de Engenharia Elétrica da PUC-RIO. Suas áreas de interesse são o processamento digital de imagens e visão computacional.

Ficha Catalográfica

Vega, Pedro Juan Soto

Reconhecimento facial em vídeo com uma amostra por pessoa utilizando stacked supervised auto-encoders / Pedro Juan Soto Vega ; orientador: Raul Queiroz Feitosa ; co-orientador: Patrick Nigri Happ. – 2016.

73 f. : il. (color.); 29,7cm

Dissertação (mestrado)-Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2016

Incluí referências bibliográficas.

1. Engenharia elétrica-Teses. 2. Reconhecimento de faces. 3. Redes neurais profundas. 4. Feitosa, Raul Queiroz II. Happ, Patrick Nigri III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV.

CDD: 621.3

À memória de meus avós
A minha mãe Irsia e meu tio Carlos Manuel
A minha noiva Brenda

Agradecimentos

Agradeço a todos que de forma direta ou indireta auxiliariam na execução dessa Dissertação de Mestrado e em especial:

Aos meus orientadores, Prof. Raul Queiroz Feitosa e Dr. Patrick Nigri Happ, pela oportunidade, paciência, dedicação e valiosas dicas.

À CAPES e à PUC-Rio pelos auxílios prestados, fundamentais para o desenvolvimento deste trabalho.

À minha mãe Irsia, por ser fonte de inspiração e ter me tornado quem sou.

Ao meu tio Carlos Manuel por seu incondicional apoio.

À minha noiva Brenda por toda sua ajuda, incentivo, paciência, amor e compreensão.

Aos meus colegas do Laboratório de Visão Computacional (LVC) da PUC-Rio, pela amizade e vital apoio.

A todos os professores e funcionários do Departamento.

A todos os amigos e familiares que de uma forma ou de outra me ajudaram para que este trabalho pudesse ser concluído.

Resumo

Vega, Pedro Juan Soto; Feitosa, Raul Queiroz; Happ, Patrick Nigri. **Reconhecimento Facial em Vídeo com uma amostra por pessoa utilizando Stacked Supervised Auto-encoder.** Rio de Janeiro, 2016. 73p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação propõe e avalia estratégias baseadas nos Stacked Supervised Auto-encoders (SSAE) para representação de imagens faciais em aplicações de vídeo vigilância. O estudo foca na identificação de faces a partir de uma amostra por pessoa na galeria (*single sample per person* - SSPP). Variações em termos de pose, expressão facial, iluminação e oclusão são abordadas de duas formas. Primeiro, o SSAE extrai atributos das imagens de faces que são robustos contra tais variações. Segundo, exploram-se as múltiplas amostras que podem ser coletadas nas sequências de vídeo de uma pessoa (*multiple samples per person probe* - MSPPP). Os métodos propostos foram avaliados e comparados usando os bancos de vídeos Honda/UCSD e VIDTIMIT. Adicionalmente, foi estudada a influência de parâmetros relacionados com a arquitetura do SSAE utilizando o banco de imagens estáticas Extended Yale B. Os resultados demonstraram que as estratégias que exploram as MSPPP em combinação com o SSAE podem superar o desempenho de outros métodos SSPP, como os Padrões Binários Locais (LBP), para reconhecimento de faces em vídeos.

Palavras-chave

Auto-encoder; Reconhecimento de faces; Vigilância.

Abstract

Vega, Pedro Juan Soto; Feitosa, Raul Queiroz (Advisor); Happ, Patrick Nigri (Co-advisor). **Single Sample Face Recognition from Video via Satcked Supervised Auto-encoder.** Rio de Janeiro, 2016. 73p. MSc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This work proposes and evaluates strategies based on Stacked Supervised Auto-encoders (SSAE) for face representation in video surveillance applications. The study focuses on the identification task with a single sample per person (SSPP) in the gallery. Variations in terms of pose, facial expression, illumination and occlusion are approached in two ways. First, the SSAE extracts features from face images, which are robust to such variations. Second, multiple samples per persons probes (MSPPP) that can be extracted from video sequences are exploited to improve recognition accuracy. The proposed methods were compared upon Honda/UCSD and VIDTIMIT video datasets. Additionally, the influence of the parameters related to SSAE architecture was studied using the Extended Yale B dataset. The experimental results demonstrated that strategies combining SSAE and MSPPP are able to outperform other SSPP methods, such as local binary patterns, in face recognition from video.

Keywords

Auto-encoder; Face recognition; Surveillance.

Sumário

Lista de figuras

Lista de tabelas

1
Introdução

São muito bem conhecidas as vantagens que o reconhecimento facial a partir de imagens digitais ou vídeos proporcionam para as sociedades atuais. Atualmente, a interação homem-computador é cada vez mais comum, sendo sem dúvida, a identificação e verificação de pessoas nos meios digitais um dos pilares fundamentais que suportam todo este desenvolvimento. No entanto, tais tarefas constituem ainda um desafio para pesquisadores das áreas de visão computacional e neurociências, os quais tentam fazer com que os computadores adquiram a mesma capacidade do cérebro para perceber e processar características intrínsecas das faces.

Mesmo com grandes progressos nesta linha de pesquisa, ainda buscam-se soluções para os mais diversos problemas que continuam afetando os sistemas de reconhecimento de faces, assim como para conseguir adaptá-los aos mais diversos cenários.

Entre tais cenários, um dos mais comuns é o reconhecimento facial tendo unicamente uma amostra por pessoa (*Single Sample per Person* - SSPP) (1) (2) na galeria. Estas amostras tipicamente apresentam pose frontal, boas condições de iluminação e expressão neutra do rosto, enquanto que as imagens das faces que devem ser identificadas apresentam variações em iluminação, pose, oclusões, expressões faciais diversas, etc. Isto constitui um desafio em especial para o reconhecimento em sequências de vídeo, onde as imagens são capturadas em condições não controladas, sem a colaboração dos alvos e com baixa qualidade dos vídeos.

Para enfrentar estas situações muitos modelos matemáticos têm sido propostos e aplicados na intenção de encontrar uma representação compacta das imagens que seja invariante sob tais condições. Os métodos baseados em análises de subespaços foram dos primeiros a serem aplicados neste contexto, no entanto não são adequados para aplicações em cenários de tipo SSPP, pois requerem

conjuntos de treinamento com um número de amostras por pessoa considerável para obter uma boa representação da face.

Por outro lado, os sistemas de reconhecimentos que utilizam métodos baseados na extração de atributos locais, como os Padrões Binários Locais (*Local Binary Pattern* - LBP) (3) e a Quantificação de Fase Local (*Local Phase Quantization* - LPQ) (4) apresentam excelentes resultados em certas condições adversas de iluminação. Entretanto, o desempenho decresce sob oclusões, fortes variações em pose e expressões faciais.

Outras alternativas, baseadas em redes neurais de múltiplas camadas, têm sido aplicadas com sucesso nos últimos anos para o aprendizado automático destas representações (5) (6) (7) (8) (9). Entre estas, os *denoising auto-encoders* (10) são exemplos destacados.

Motivados por tal sucesso, Gao e coautores (11) apresentam uma solução inovadora aplicada ao contexto de uma amostra por pessoa na galeria, propondo o *Stacked Supervised Auto-Encoder* (SSAE). O SSAE constitui uma arquitetura que considera as imagens faciais com pose não frontal, expressões faciais, oclusões e variações de iluminação como dados ruidosos, enquanto que consideram-se como dados limpos: imagens de faces com pose frontal, livre de oclusões, com expressão neutra e boas condições de iluminação. Diante disto, o SSAE tenta produzir descritores de imagens que são robustos contra tais variações.

1.1. Motivação

Os resultados reportados por Gao indicam que a aproximação proposta apresenta um desempenho superior em comparação com técnicas alternativas de representação de imagens para bancos de dados de imagens estáticas anotadas manualmente e com alinhamento de face. Contudo, caso um detector automático de faces como o Viola & Jones (12) seja aplicado, a eficiência do modelo decresce consideravelmente.

Sistemas de reconhecimento em vídeos tipicamente rastreiam a pessoa a ser identificada através de seus quadros. O conjunto de imagens coletadas ao longo do rastreamento em combinação com detectores automáticos de faces é denominado

nesta dissertação Múltiplas Amostras por Pessoa na Prova (*Multiple Samples per Person Probes* - MSPPP).

Diante deste cenário, esta dissertação visa estudar o desempenho do SSAE em imagens detectadas de forma automática, assim como a utilização das MSPPP para mitigar os efeitos decorrentes da baixa qualidade das imagens coletadas em aplicações de vídeo monitoramento.

1.2. Objetivos

O **objetivo geral** desta dissertação é avaliar o desempenho do SSAE para imagens de faces detectadas automaticamente em sequências de vídeo e examinar formas de explorar as Múltiplas Amostras por Pessoa da Prova, visando mitigar problemas que dificultam o reconhecimento facial em aplicações de vídeo monitoramento.

Os **objetivos específicos** deste trabalho são os seguintes:

- Avaliar a influência dos parâmetros relacionados com a arquitetura do SSAE em seu desempenho.
- Avaliar o SSAE a partir de imagens estáticas extraídas de sequências de vídeo.
- Avaliar o SSAE para imagens detectadas e anotadas automaticamente (11).
- Propor e avaliar técnicas que permitam explorar as Múltiplas Amostras por Pessoa na Prova disponíveis em sequências de vídeo.
- Identificar as dificuldades que impactam no desempenho do SSAE.

1.3. Estrutura da Dissertação

A presente dissertação está dividida em seis capítulos organizados da seguinte forma:

- O capítulo 2 apresenta um estudo do estado da arte em reconhecimento facial no cenário em que há uma amostra por pessoa na galeria.

- O capítulo 3 descreve os fundamentos teóricos subjacentes às redes neurais profundas, assim suas arquiteturas mais estudadas. Descreve-se em detalhe a arquitetura e o procedimento de treinamento do SSAE. Finalmente, expõe-se de forma resumida a técnica conhecida Padrões Binários Locais (LBP), que servirá como referência para avaliação, e descrevem-se as métricas de dissimilaridade utilizadas neste trabalho.
- O capítulo 4 introduz o reconhecimento de faces utilizando múltiplas amostras por pessoa na prova. Adicionalmente, são apresentadas técnicas que exploram a possibilidade de ter várias amostras por pessoa na prova.
- O capítulo 5 detalha a avaliação experimental seguida neste estudo, além de apresentar e discutir os resultados obtidos.
- O capítulo 6 finalmente resume as conclusões extraídas no desenvolvimento desta pesquisa e aponta para trabalhos futuros.

2
Trabalhos Relacionados

O reconhecimento de faces utilizando uma única imagem por pessoa como amostra na galeria constitui um dos cenários mais comuns para os sistemas de reconhecimento. Nos últimos anos, muitos trabalhos têm sido apresentados tendo como principal objetivo melhorar os resultados obtidos neste tipo de cenário. Neste capítulo são apresentados, de forma sucinta, um conjunto de trabalhos que representam o estado da arte em reconhecimento de faces com uma amostra por pessoa na galeria.

2.1.
Métodos baseados em Análise de Subespaços

Os métodos baseados em análise de subespaços como *EigenFaces* (13) e a Análise de Componentes Principais de duas dimensões (*Two Dimensional Principal Components Analysis* - 2DPCA) (14) estão entre as mais conhecidas técnicas de representação de imagens faciais. Ambos têm como base teórica a Análise de Componentes Principais (*Principal Components Analysis* - PCA) (15).

Nestas aproximações cada imagem facial é representada por um vetor de alta dimensão (gerado pela concatenação dos valores de cada pixel). Determina-se um subespaço definido pelas direções de mais alta variância. As principais vantagens destas abordagens estão relacionadas com a preservação detalhada de padrões de textura e informação e com a capacidade de capturar características globais das faces.

No entanto, estes modelos possuem duas desvantagens em cenários de uma amostra por pessoa (*Single Sample per Person*- SSPP) na galeria. A primeira diz respeito à relação entre a dimensionalidade das imagens e o número de amostras; e a segunda é dada pela existência de um único vetor representando cada classe (pessoa), o que dificulta a representação das variações de aparência nas faces de cada pessoa.

Com o objetivo de resolver estes inconvenientes e fazer com que os métodos sejam mais adequados em cenários SSPP, trabalhos têm sido propostos como os apresentados por Wu e coautores (16) e Chen e colaboradores (17). Nestes estudos, acrescentam-se ao conjunto de treinamento imagens sintéticas geradas através da aplicação de transformações geométricas, distúrbios por adição de ruído ou filtragem à imagens reais. Estes métodos também são usados nos trabalhos apresentados em (18) (19) (20) (21) (22). A Figura 2-1 apresenta um exemplo destas transformações.

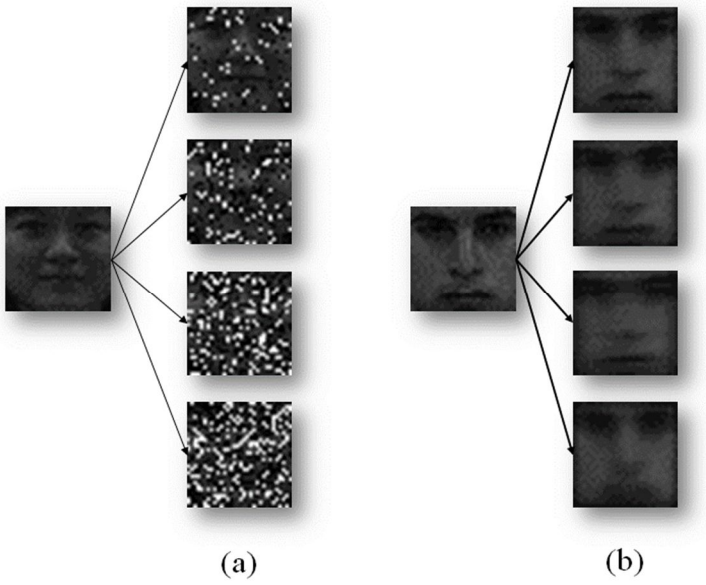

Figura 2-1: Geração de novas amostras a partir de: (a) adição de ruído (*salt & pepper*) e (b) pré-processamento da imagem com filtros (neste exemplo gaussianos).

Todas estas abordagens melhoram as taxas de reconhecimento dos métodos originais dos quais se derivam. No entanto, esta melhora é relativamente modesta.

2.2. Métodos baseados em Extração de Atributos Locais

Os métodos de extração de atributos locais geralmente utilizam uma imagem por pessoa para extrair medidas geométricas, como a largura da cabeça, distância entre os olhos, a posição da boca entre outras (2). Estas medidas são armazenadas para posteriormente serem comparadas com medidas extraídas de imagens faciais da pessoa cuja identidade se deseja determinar. Duas dificuldades

impactam o desempenho de tais abordagens: 1) os atributos geométricos são difíceis de extrair em certos casos, por exemplo, em pose não frontal e oclusão; 2) estes atributos isoladamente não capturam todas as variações que as imagens faciais de uma mesma pessoa podem apresentar.

Para enfrentar tais problemas duas direções de pesquisa foram propostas, sendo a primeira focada na forma de detectar atributos faciais robustos frente às diversas formas de variação. Exemplos disso são as propostas de Brunelli e Poggio (23) e Rowley e colaboradores (24) que detectam de forma automática as regiões dos olhos, o nariz e a boca utilizando os níveis de cinza das imagens. Nestes casos, usa-se a correlação como medida de similaridade entre as regiões de uma face conhecida e uma desconhecida.

A segunda linha de pesquisa tenta obter uma representação das características locais mais poderosas que as abordagens geométricas. Nesta direção destaca-se a contribuição de Manjunath e coautores (25) propondo um método para detectar e reconhecer faces baseado nas *wavelets* de Gabor (26) (27). O método detecta uma série de pontos chaves com os quais se constrói um grafo topológico de acordo com a seguinte regra: dois pontos chaves próximos espacialmente com uma certa distância mínima deverão ser conectados com uma aresta. Depois que o grafo é construído, o reconhecimento de face passa a ser formulado como um problema de correspondência entre grafos, para o qual existem diversas soluções na literatura. A efetividade do método foi avaliada em uma base de dados onde foram testadas imagens de 86 pessoas com variações de pose e expressões faciais. Nestas condições o algoritmo obteve uma taxa de 94% de reconhecimento, demostrando um bom desempenho. Uma limitação deste método é que os grafos são rígidos, ou seja, uma vez construído, o grafo não se altera, o que dificulta o ajuste do grafo a imagens distintas daquela a partir da qual o grafo foi construído.

Baseados nesta observação, Lades e coautores (28) propuseram uma topologia deformável para fazer a comparação dos grafos. Esta topologia é conhecida atualmente como EBGM (*Elastic Bunch Graph Matching*) (29). Constatou-se, contudo, que este método não apresenta bom desempenho em situações em que o ponto chave que suporta o grafo está oculto por alguma oclusão. A Figura 2-2 apresenta de forma resumida o processo de detecção dos pontos chave assim como de criação do grafo que modela o rosto. No extremos

esquerdo se tem uma imagem facial, e ao lado os *kernels* de convolução baseados nas *wavelets* de Gabor. O resultado da convolução dos *kernels* com a imagem facial é apresentado no centro da figura em termos da parte imaginária e da magnitude. O ponto chave é descrito em termos da frequência espacial e da orientação dominante na região em seu entorno. Finalmente o grafo é construído a partir dos pontos chave tais como os posicionados sobre os olhos, nariz e boca .

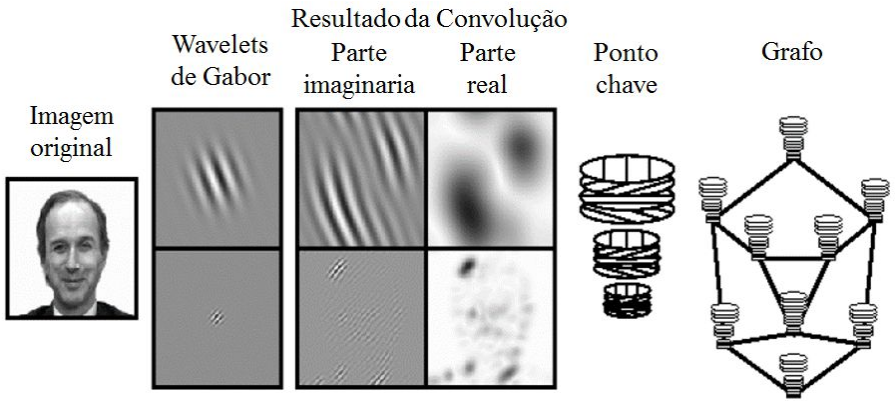

Figura 2-2: Representação do grafo baseado nas *wavelets* de Gabor. O ponto chave contém a informação de frequência e orientação dos seus vizinhos. (Figura modificada de (29))

2.3. Métodos baseados em Extração de Atributos de Aparências Locais

Nas abordagens baseadas em atributos de aparência locais, a imagem é dividida em pequenas regiões das quais são extraídos atributos que atendem a algum critério entre os já descritos, como as *wavelets* de Gabor, etc. Como o número destes atributos pode ser muito grande, alguns dos métodos envolvem uma redução de dimensionalidade utilizando PCA (15) ou Análise Discriminante Linear (*Linear Discriminant Analysis* - LDA) (30). Posteriormente, um classificador é aplicado para fazer a identificação.

Destacam-se aqui modelos baseados nos Padrões Binários Locais (*Local Binary Pattern* - LBP) (3) e em Quantificação de Fase Local (*Local Phase Quantization* - LPQ) (4). Ambos os métodos exploram atributos de textura, criando uma codificação da imagem facial baseando-se nas variações dos níveis de cinza (para o LBP) e da informação de fase (para LPQ) na vizinhança da cada um dos pixels da imagem. Em seguida essa representação da imagem facial é

dividida em regiões de dimensões fixas, das quais se extraem descritores formados a partir dos histogramas dos atributos de textura em cada região. Por fim, o descritor é submetido à comparação com descritores de imagens faciais de pessoas conhecidas.

O LBP e o LPQ apresentam excelentes resultados, sobretudo devido à invariância dos descritores quanto à iluminação. Contudo, tais métodos são muito sensíveis a variações em pose e oclusões. Este fato tem sido comprovado em muitos trabalhos do estado da arte e também se confirma neste trabalho. Uma descrição detalhada destes algoritmos é apresentada no capítulo 3 desta dissertação.

Considerando tudo o que foi descrito até este ponto, observa-se que para o cenário SSPP, os métodos baseados em extração de atributos locais oferecem soluções mais robustas e eficientes do que os que realizam uma análise do subespaço. A principal desvantagem dos últimos em comparação com os primeiros reside no número de amostras por pessoa que se faz necessário para criar uma boa representação. No entanto, os métodos baseados em atributos locais apresentam dificuldades diante de variações de pose, oclusão e ruído. Observa-se na Figura 2-3 a diversidade em termos de variações de iluminação, expressões faciais, pose, escala, oclusões e ruído presentes em aplicações práticas de reconhecimento faciais em vídeo.

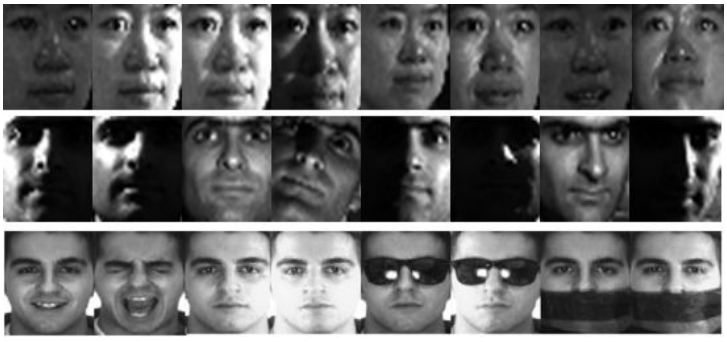

Figura 2-3: Exemplos de variações presentes nas imagens de face.

Trabalhos recentes apresentam excelentes resultados que superam as dificuldades apresentadas neste capítulo. Um aspecto comum a vários destes trabalhos é o uso de redes neurais de muitas camadas escondidas, conhecidas como Redes Neurais Profundas (*Deep Neural Networks,* DNNs) e que são a base

do que se conhece hoje na literatura como Aprendizado Profundo (*Deep Learning*), termo muito usado na área de Aprendizado de Máquinas. Por tudo isso, o capítulo seguinte descreve algumas destas arquiteturas.

3
Fundamentos Teóricos

Dada a importância da extração de atributos faciais e tendo como base as recentes contribuições nesta área, o presente capítulo descreve em detalhe os principais fundamentos teóricos dos algoritmos sobre os quais se apoia esta dissertação, assim como as implementações dos mesmos.

3.1.
Redes Neurais Profundas

As redes neurais artificiais, até alguns anos atrás, apresentavam certas limitações para tarefas de visão computacional, reconhecimento de padrões e outras áreas relacionadas. Problemas como a limitada quantidade de dados etiquetados para treinamento, inviabilizavam a representação de imagens em vários níveis.

Hinton e colaboradores (31) propuseram uma nova forma de treinar as redes neurais de múltiplas camadas, através de um pré-treinamento não supervisionado que produz uma inicialização dos pesos da rede. Este pré-treinamento mitiga os problemas do decaimento do gradiente e da baixa disponibilidade de dados etiquetados. No passo seguinte um treinamento supervisionado é aplicado para ajustar os pesos inicializados no pré-treinamento ao problema em questão. Este último procedimento é conhecido como ajuste fino dos parâmetros. Nesta proposta de Hinton e coautores (31) o pré-treinamento é realizado tendo como base uma variante das Máquinas de Boltzmann denominada Máquinas de Boltzmann Restritas (*Restricted Boltzman Machine* - RBMs) e o algoritmo de Divergência Contrastiva (*Contrastive Divergence* - CD) (32) para a atualização dos pesos em cada camada.

Os resultados obtidos em (31) demonstraram que é possível treinar múltiplas camadas em uma rede neural, corrigir os problemas supra citados e aprender uma representação robusta da imagem capaz de melhorar a eficiência dos sistemas de reconhecimento de padrões de forma geral.

Como consequência, outros tipos de arquiteturas baseadas nas redes neurais profundas foram propostos tendo como foco principal o aprendizado de representações. Exemplos disso são as pesquisas apresentadas por Bengio e coautores (33) (34) (35) baseadas em *Autoencoders*, Tang e colaboradores com as *Deep Lambertian Networks,* (DLN) (36), além de Kan e coautores que propuseram um modelo baseado em *autoencoders* que implementa um aprendizado progressivo no qual cada camada da rede é forçada a aprender uma representação específica (37). Dado que a presente dissertação baseia-se nos *auto-encoders*, as seguintes seções apresentam estas arquiteturas.

3.1.1. Auto-Encoders

Os *auto-encoders* são redes neurais não supervisionadas que visam criar uma representação compacta de suas entradas, a partir da qual é possível reconstruir a própria entrada tão exatamente quanto possivel. A arquitetura é tipicamente usada para fazer redução de dimensionalidade (38) (39). Os *auto-encoders* têm normalmente duas partes estruturais: um codificador e um decodificador (37) que são implementadas em torno de uma única camada escondida. A Figura 3-1 mostra um exemplo desta arquitetura.

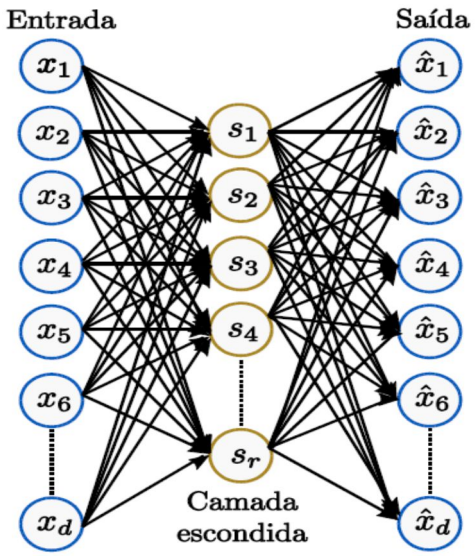

Figura 3-1: Exemplo de *auto-encoder*.

Observa-se que o codificador, denotado por f, mapeia a entrada $x \in \mathbb{R}^d$ para uma representação compacta $z \in \mathbb{R}^r$ através das ativações dos r neurônios da camada escondida, em que $r < d$. A função f pode ser escrita segundo a equação 3-1:

$$z = f(x) = s(Wx + b) \quad \textbf{3-1}$$

onde $W \in \mathbb{R}^{rxd}$ é uma matriz de coeficientes, $b \in \mathbb{R}^r$ representa o viés e $s(\cdot)$ uma função de ativação, comumente é não linear. Funções como a *sigmoide* e *tanh* são as mais usadas nestas arquiteturas.

O decodificador, representado por g tenta reconstruir o dado de entrada x a partir da representação z obtida pelo codificador:

$$\hat{x} = g(z) = s(\widehat{W}z + \hat{b}) \quad \textbf{3-2}$$

com $\widehat{W} \in \mathbb{R}^{dxr}$ contendo os coeficientes da transformação não linear e $\hat{b} \in \mathbb{R}^d$ o *bias* da reconstrução.

Dado que o objetivo do *auto-encoder* é obter uma reconstrução muito próxima dos dados de entrada, os parâmetros $W, b, \widehat{W}$ e $\hat{b}$ são ajustados através da minimização da função de perda:

$$\left[W^*, b^*, \widehat{W}^*, \hat{b}^*\right] = \underset{W,b,\widehat{W},\hat{b}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| x^{(i)} - g(f(x^{(i)})) \right\|_2^2 \quad \textbf{3-3}$$

onde $x^{(i)}$corresponde ao i-ésimo exemplo do conjunto de N amostras de treinamento.

Métodos baseados no gradiente descendente são utilizados para resolver a equação 3-3, a qual representa o erro quadrático entre os dados de entrada e a reconstrução feita pela rede. Desta forma, a rede é forçada a aprender uma representação mediante a redução da dimensionalidade dos dados de entrada, o que pode ser visto como uma variante do PCA, só que neste caso, não linear.

Seguindo a abordagem de aprendizado profundo, é possível construir estruturas mais complexas empilhando-se vários níveis, cada uma formada por *auto-encoders*. Esta nova configuração recebe o nome de *Stacked Auto-Encoders (SAE)*. Uma vez que um *auto-encoder* é treinado, pode se treinar outro recursivamente usando como dados de entrada as representações z de cada um dos dados de entrada originais do *auto-encoder* anterior. Assim que este novo *auto-encoder* for treinado, juntam-se as duas estruturas e o processo pode ser repetido

até que se atinja a profundidade desejada. Em teoria, melhores resultados se obtém à medida que se acrescentam mais níveis. Na prática, contudo, devido, por exemplo, ao número limitado de padrões de treinamento pouco ganho se obtém a partir de um certo número de níveis. A Figura 3-2 exemplifica o processo de treinamento dos SAE.

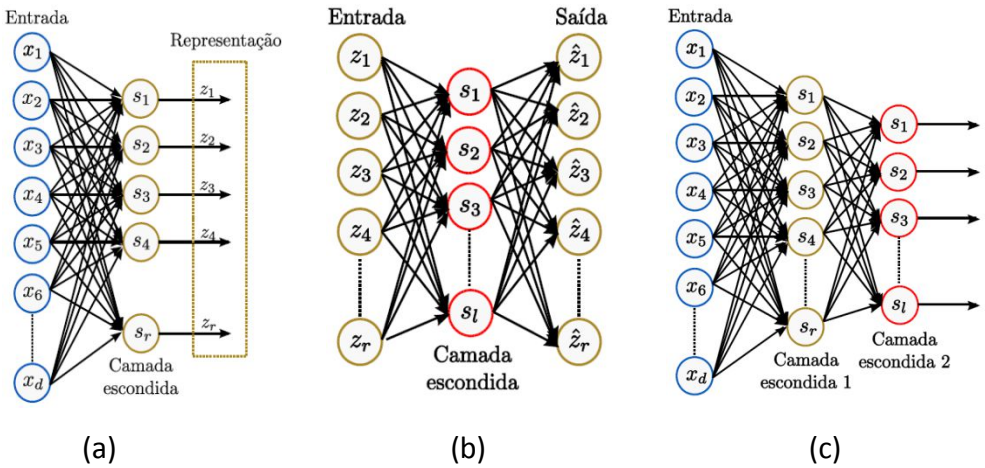

Figura 3-2: Treinamento do *Stacked Auto-Encoder*. (a): Representação obtida pelo treinamento do primeiro *auto-encoder*. (b) Treinamento do segundo *auto-encoder* no qual $l < r$. (c) *Satcked Auto-Encoder* treiando.

É importante destacar que a aproximação tradicional do SAE é orientada à redução da dimensionalidade dos dados de entrada ($l < r < d$) produzindo uma representação que pode ser vista como uma compressão com perdas aplicada em x (Figura 3-2). Isto não é suficiente para obter uma representação adequada dos dados, sobretudo se o que se deseja é reter informação relevante acerca de x na representação z. Uma possível solução seria ter as dimensões das camadas escondidas e dos dados de entrada iguais, quer dizer $l = r = d$. No entanto, o *auto-encoder* faria uma reconstrução perfeita dos dados, mapeando inclusive o ruído que idealmente deveria ser rejeitado. Por esta razão adicionam-se outras restrições na etapa de pré-treinamento, as quais produziram arquiteturas mais robustas como a que se apresenta na seguinte seção.

3.1.2.
Sparse Auto-Encoders

Considerando que:

1. O critério de treinamento baseado na redução de dimensionalidade não é suficiente para obter uma representação eficiente; e
2. Os neurônios no córtex cerebral representam um dado genérico de entrada de modo esparso, isto é, apenas uma pequena parcela dos neurônios envolvidos na representação é ativada.

Olshausen e colaboradores (40) apresentaram uma aproximação na qual utilizam o conceito de esparsidade para obter uma representação mais eficiente e útil para a etapa de classificação, solucionando também o problema da perda de informação pela redução da dimensão do vetor de atributos, utilizando para isto a seguinte função de perda:

$$\left[W^*, b^*, \widehat{W}^*, \hat{b}^*\right] = \underset{W,b,\widehat{W},\hat{b}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| x^{(i)} - g(f(x^{(i)})) \right\|_2^2 - \lambda_{sp} J_{sp} \qquad \textbf{3-4}$$

onde λ_{sp} é uma constante positiva que determina a importância do segundo termo J_{sp}, conhecido como termo de esparsidade, relativo ao primeiro. O primeiro termo da função é o erro médio quadrático ou de preservação da informação. O termo J_{sp} mede a esparsidade do código z para uma imagem de entrada x_i, aplicando uma penalização forte ou fraca, dependendo de quão distribuídas são as ativações. Assim, representações com atividade significativa em vários coeficientes serão mais penalizadas do que aquelas em que a atividade se concentra em poucos coeficientes.

Das muitas opções que existem para definir o termo de esparsidade, uma das mais utilizadas é a divergência Kullback-Leibler (11) (41):

$$J_{sp} = \sum_{h} KL(\rho || \hat{\rho}_j) = \sum_{h} \sum_{j=1}^{r} \left(\rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \right) \qquad \textbf{3-5}$$

onde $\hat{\rho}_j$ é a média das ativações na unidade oculta j da camada escondida h computada sobre as N amostras de treinamento:

$$\hat{\rho}_j = \frac{1}{N} \sum_{i=1}^{N} z_j^{(i)} \qquad \textbf{3-6}$$

e ρ é o parâmetro de esparsidade, o qual toma valores próximos de zero para forçar que as médias das ativações em cada neurônio a também tomarem valores próximos de zero. Busca-se, desta forma, que a maioria das ativações na camada escondida tenda a zero, obtendo como consequência uma representação esparsa da imagem de entrada.

Da mesma forma que nos *auto-encoders*, uma vez que um primeiro bloco é treinado, é possível treinar outros recursivamente a fim de formar uma estrutura profunda capaz de extrair atributos com maior nível de complexidade.

3.1.3. Denoising Auto-Encoders

A partir de que o critério de reconstrução isoladamente não garante uma boa extração de atributos e, considerando a possibilidade de se obter uma representação de alta dimensionalidade baseada nas contribuições de Olshausen e Field (40) e Ranzato e coautores (42), os *denoising auto-encoders* (10) surgiram como uma tentativa de construir uma arquitetura capaz de aprender uma representação z onde sua dimensão fosse maior ou igual à dos dados de entrada $(r \geq d)$, a partir de um critério de reconstrução diferente.

O critério em questão se baseia no conceito de que "*uma boa representação é aquela que pode ser obtida de forma robusta a partir de uma entrada corrompida e que será útil para recuperar a entrada correspondente pura ou incorrupta"* (10).

Tendo então um conjunto de treinamento não corrompido x e sua versão ruidosa $\tilde{x} \sim q_{\mathcal{D}}(\tilde{x}|x)$, o *auto-encoder* é treinado para reconstruir a versão não corrompida do dado de entrada a partir da versão corrompida, resolvendo a equação 3-7:

$$\left[W^*, b^*, \widehat{W}^*, \hat{b}^*\right] = \underset{W,b,\widehat{W},\hat{b}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| x^{(i)} - g(f(\tilde{x}^{(i)})) \right\|_2^2 \qquad \textbf{3-7}$$

Observa-se que o *denoising auto-encoder* minimiza a mesma função de perda que o *auto-encoder* simples sem fazer redução de dimensionalidade. A diferença encontra-se no cálculo de z, que é realizado a partir de um mapeamento de entradas corrompidas nas correspondentes não corrompidas, o que já garante

por si só que não se aprenderá a função identidade. Um diagrama desta arquitetura é apresentado na Figura 3-3.

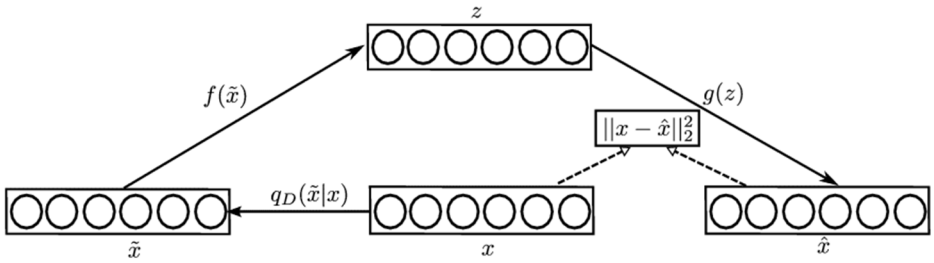

Figura 3-3: Arquitetura do *Denoising Auto-Encoder*.

Numa outra representação, a Figura 3-4 apresenta a formação de uma arquitetura mais complexa através do treinamento não supervisionado de dois blocos básicos dos *denoising auto-encoders*.

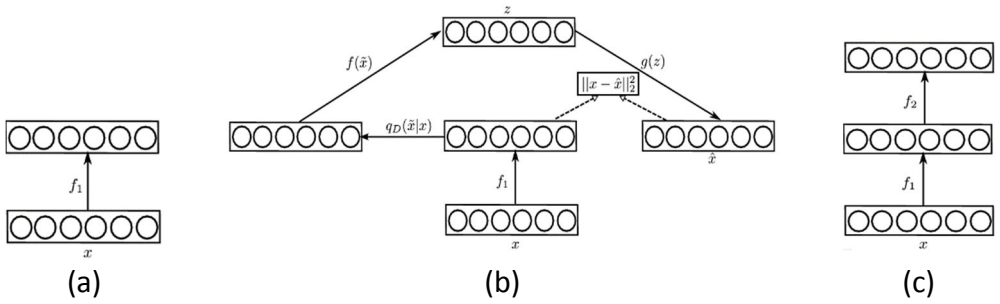

Figura 3-4: *Stacked Denoising Auto-Encoder*. (a): Assim que o primeiro *auto-encoder* é treinado (Veja Figura 3-3), a sua representação de saída é empregada como entrada do próximo bloco: (b). Por último é possível obter uma arquitetura com maior nível de complexidade: (c).

3.1.4. Stacked Supervised Auto-Encoders

Variantes ou extensões dos *denoising auto-enconders* foram propostas para problemas específicos de cenários diversos como o abordado no capítulo 2 do presente trabalho, ou seja, reconhecimento facial onde só se tem uma amostra por pessoa na galeria (*Single Sample per Person* - SSPP) com pose frontal, boas condições de iluminação e expressão neutra.

Um exemplo disto é a proposta de Gao e coautores (11), que aplica o conceito dos *denoising auto-encoders* ao cenário SSPP para criar um modelo supervisionado onde as imagens da galeria são tratadas como dados sem nenhum tipo de perturbação (dados sem ruído) e as suas variantes em termos de pose, iluminação, expressões faciais, oclusões, etc, como dados corrompidos. Da mesma forma que os *denoising auto-encoders*, este modelo denominado por Gao e colaboradores como *Stacked Supervised Auto-Encoders* (SSAE) é treinado para produzir a versão livre de ruído de qualquer imagem de entrada de uma pessoa.

Denotando cada uma das imagens corrompidas como $\tilde{x}^{(i)}$ e a sua versão não corrompida como $x^{(i)}(i = 1, \dots, N)$, os parâmetros $W, b, \widehat{W}, \hat{b}$ são determinados minimizando-se uma função de perda adequada, como por exemplo:

$$[W^*, b^*, \widehat{W}^*, \hat{b}^*] = \underset{W,b,\widehat{W},\hat{b}}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(||x^{(i)} - g(f(\tilde{x}^{(i)}))||_2^2 + \lambda ||f\left(x^{(i)}\right) - f(\tilde{x}^{(i)})||_2^2 \right) + \lambda_{wd} J_{wd} + \lambda_{sp} J_{sp} \right] \quad \textbf{3-8}$$

Na expressão, o termo $(||x^{(i)} - g(f(\tilde{x}^{(i)}))||_2^2)$, constitui o erro de reconstrução do *auto-encoder*. Por outro lado, o termo de preservação da similaridade $||f\left(x^{(i)}\right) - f(\tilde{x}^{(i)})||_2^2$, impõe que as representações de uma mesma pessoa $f(x^{(i)})$ e $f(\tilde{x}^{(i)})$, sejam semelhantes. A importância relativa deste termo é ajustada pelo coeficiente λ.

O termo de decaimento dos pesos J_{wd} é definido como a soma das normas de Frobenius:

$$J_{wd} = ||W||_F^2 + ||\widehat{W}||_F^2 \quad \textbf{3-9}$$

e tem como propósito garantir valores pequenos dos pesos, evitando-se, assim, *overfitting*. A importância relativa de J_{wd} é ajustada pelo coeficiente de regularização λ_{wd}.

Finalmente, o termo J_{sp}, ponderado por λ_{sp}, representa a restrição de esparsidade imposta na saída da camada escondida e é definido pela divergência Kullback-Leiber como:

$$J_{sp} = KL(\rho_0||\rho) + KL(\rho_0||\tilde{\rho}) \quad \textbf{3-10}$$

onde o índice j denota um neurônio particular da camada oculta, e:

$$\rho_j = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})_j \quad \textbf{3-11}$$

$$\tilde{\rho}_j = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{x}^{(i)})_j \quad \textbf{3-12}$$

$$KL(\rho_0||\rho) = \sum_j \left(\rho_0 \log\left(\frac{\rho_0}{\rho_j}\right) + (1-\rho_0)\log\left(\frac{1-\rho_0}{1-\rho_j}\right)\right) \quad \textbf{3-13}$$

Tendo como propósito construir uma rede neural profunda, o esquema proposto pode ser replicado em múltiplas camadas, treinando cada uma destas separadamente segundo a equação 3-8. Assim que a primeira camada é treinada, os atributos produzidos na sua saída para dados limpos e com ruído são utilizados como dados de entrada para treinamento da camada seguinte. Em relação ao número de camadas escondidas, os autores identificaram de forma empírica que o esquema não traz ganho significativo para mais de duas camadas.

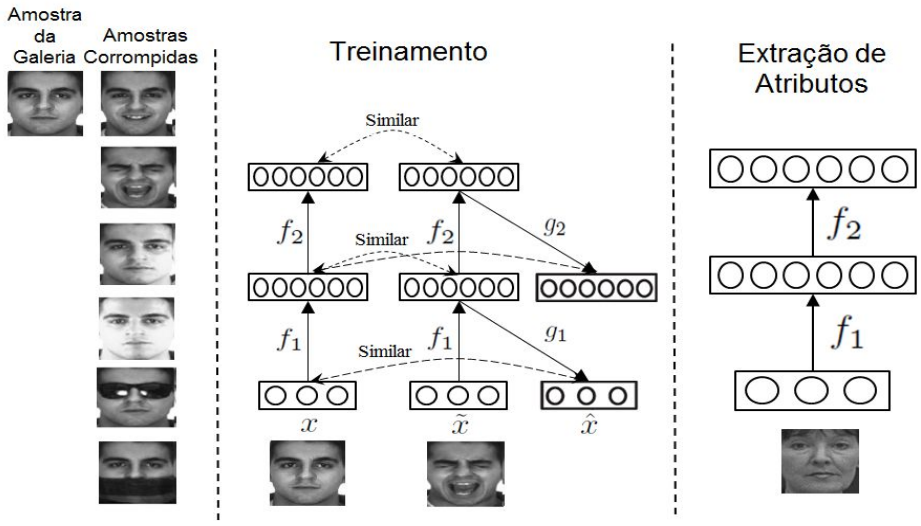

Figura 3-5: Arquitetura do SSAE.

A Figura 3-5 ajuda na compreensão da arquitetura e dos procedimentos de treinamento e de extração de atributos. As imagens na esquerda da figura representam as amostras de treinamento, nota-se que a imagem da galeria é frontal, bem iluminada, sem oclusão e com expressão neutra. Por outro lado, as imagens de prova (amostras corrompidas) apresentam um conjunto de variações importantes em pose, expressão da face, oclusões parciais e iluminação. No meio da figura é representado o processo de treinamento, onde os símbolos f_1 e f_2 referem-se a funções dos codificadores da primeira e da segunda camada escondida, respectivamente. Da mesma forma, g_1 e g_2 descrevem as funções dos

decodificadores em cada uma das camadas. Por fim, o esquema à direita diz respeito à extração de características de qualquer imagem de entrada a fim de produzir a sua representação.

Uma descrição detalhada do processo de treinamento é apresentada no Algoritmo 1. Observa-se que o conjunto de treinamento é dividido em imagens corrompidas e não corrompidas e cada uma destas é etiquetada de acordo com a pessoa representada na imagem. A função de perda é essencialmente a mesma da equação 3-8, com a pequena diferença de que agora existem múltiplas imagens corrompidas de uma mesma pessoa.

Algoritmo 1. Treinamento do SSAE

Entrada: Imagens da galeria $X_{tr} = \left\{x_{tr}^{(i)}\right\}$, para $i = 1, \dots, G_{tr}$
Imagens corrompidas $\tilde{\mathrm{X}}_{tr} = \left\{\tilde{x}_{tr}^{(j)}\right\}$, para $j = 1, \dots, N$
Etiquetas das imagens corrompidas $L_{tr} = \{l_{tr}^{(j)}\}$, para $j = 1, \dots, N$ onde $l_{train}^{(j)}$ pode assumir valores no conjunto $\{1, \dots, G_{tr}\}$
Número de camadas escondidas H
Coeficiente de similaridade λ
Coeficiente de regularização λ_{wd}
Coeficiente de esparsidade λ_{sp}
Inicialização das matrizes de pesos $W^{(h)}, \widehat{W}^{(h)}$ for $h = 1,\dots H$,
Inicialização dos *bias* $b^{(h)}, \hat{b}^{(h)}$ para $h = 1,\dots, H$

1: **For** $h = 1,\dots, H$ **do**
/Treinamento da rede
2: $\left[W^*, b^*, \widehat{W}^*, \hat{b}^*\right] \leftarrow \mathrm{argmin}_{W,b,\widehat{W},\hat{b}} \left\{\left[\frac{1}{N}\sum_{i=1}^{G}\sum_{j|l^{(j)}=i}\left(||x_{tr}^{(i)} - g(f(\tilde{x}_{tr}^{(i)}))||_2^2 + \lambda||f\left(x_{tr}^{(i)}\right) - f(\tilde{x}_{tr}^{(i)})||_2^2\right)\right] + \lambda_{wd}J_{wd} + \lambda_{sp}J_{sp}\right\}$
/Salvar os parâmetros treinados da camada atual
3: $\left\{W^{(h)}, b^{(h)}\right\} \leftarrow [W^*, b^*]$
4: **if** $h < H$ **do**
/Cálculo das representações da camada atual
5: **For** $i = 1, \dots, G_{tr}$ **do**
6: $x_{tr}^{(i)} \leftarrow f(x_{tr}^{(i)})$
7: **end For**
8: **For** $j = 1, \dots N$ **do**
9: $\tilde{x}_{tr}^{(i)} \leftarrow f(\tilde{x}_{tr}^{(i)})$
10: **end For**
11: **end if**
12: **end For**
13: **Saída:** Modelo SSAE $\theta = \left\{W^{(1)}, b^{(1)}, \dots, W^{(H)}, b^{(H)}\right\}$

Uma vez que a arquitetura proposta por Gao e colaboradores (11) é utilizada como base de estudo para o presente trabalho, a equação 3-8 foi modificada acrescentando-se o termo de regularização dos pesos na função de perda com o objetivo de estudar sua contribuição na obtenção de um modelo com uma boa generalização do erro.

3.2. Padrões Binários Locais

Os Padrões Binários Locais (*Local Binary Patterns* - LBP) introduzidos por Ojala e colaboradores (3) são um dos descritores de textura mais utilizados na atualidade. Embora simples, possuem alto nível de descriminação e eficiência computacional sendo esta a razão pela qual é extensivamente utilizado em diferentes aplicações. Adicionalmente, o LBP constitui uma referência para muitas outras aproximações incluindo a estudada nesta dissertação.

O LBP codifica cada *pixel* da imagem como uma sequência de *bits* a partir das diferenças entre os níveis de cinza de cada *pixel* e de sua respectiva vizinhança. O resultado é representado por uma matriz da mesma dimensão da imagem de entrada, cujos elementos são os números binários formados pelas sequências de *bits* assim calculadas.

De forma mais detalhada, a imagem é codificada pela comparação de cada pixel x_c com seus vizinhos x_i. Esta codificação é definida pela equação 3-14 onde LBP(x_c) é o número que representa a textura associada à vizinhança do pixel x_c.

$$LBP(x_c) = \sum_{i=0}^{7} D(x_c - x_i)2^i \qquad \textbf{3-14}$$

$$D(x_c - x_i) = \begin{cases} 1, x_c \leq x_i \\ 0, x_c < x_i \end{cases} \qquad \textbf{3-15}$$

Uma vez feito isto, a nova codificação é dividida em pequenos blocos, a partir dos quais é computado um vetor de atributos com base nos seus histogramas. A Figura 3-6 representa de forma resumida o processo.

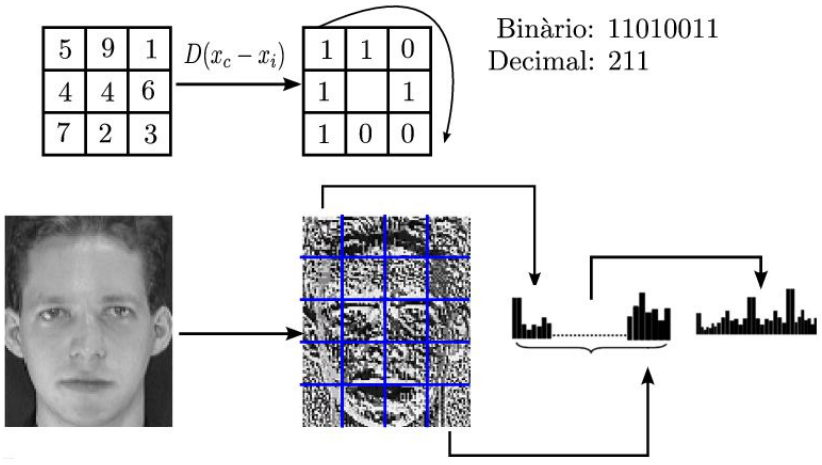

Figura 3-6: Arquitetura do LBP.

3.3. Funções de similaridade

É comum o reconhecimento facial baseie-se em medidas de similaridade para identificar ou verificar a identidade de uma pessoa. Estes graus podem ser obtidos com base nos atributos extraídos por algoritmos especializados, como o LBP, por exemplo. Neste sentido, variadas técnicas são empregadas para estabelecer estas medidas de similaridade, desde a utilização de simples medidas de distâncias entre os vetores de atributos até o uso de classificadores complexos.

A presente seção apresenta e explica de forma sucinta as métricas e classificadores aqui utilizados.

3.3.1. Funções de similaridade baseadas em distância

O problema de estimar um grau de similaridade entre dois vetores de atributos pode ser reduzido a um simples problema de cálculo da distância entre estes. Neste sentido, muitas são as alternativas que podem ser escolhidas para o cálculo da distância entre dois pontos num espaço n-dimensional. No entanto, tem se constatado que na literatura existe uma grande utilização das distâncias euclidiana (43), cosseno (44) e χ^2 (3), razão pela qual as adotamos neste trabalho.

Sejam dois vetores de atributos $Z^{(1)} = (z_1^{(1)}, z_2^{(1)}, \dots, z_n^{(1)})$ e $Z^{(2)} = (z_1^{(2)}, z_2^{(2)}, \dots, z_n^{(2)})$. As seguintes distâncias são definidas:

$$d_E\left(Z^{(1)}, Z^{(2)}\right) = \sqrt{\sum_{i=1}^{n} (z_i^{(1)} - z_i^{(2)})^2} \qquad \textbf{3-16}$$

$$d_C\left(Z^{(1)}, Z^{(2)}\right) = \cos(\theta) = \frac{\vec{Z}^{(1)} \cdot \vec{Z}^{(2)}}{||Z^{(1)}||||Z^{(2)}||} \qquad \textbf{3-17}$$

$$d_{\chi^2}\left(Z^{(1)}, Z^{(2)}\right) = \sum_i \frac{(z_i^{(1)} - z_i^{(2)})^2}{z_i^{(1)} + z_i^{(2)}} \qquad \textbf{3-18}$$

onde $d_E(\cdot), d_C(\cdot), d_{\chi^2}(\cdot)$ são as distâncias euclidiana, cosseno e χ^2 respectivamente. Em particular, a distância $d_{\chi^2}(\cdot)$ é mais utilizada em situações

onde os vetores são gerados a partir de histogramas, sendo o LBP um dos exemplos onde é utilizada.

3.3.2. Funções de similaridade em função da esparsidade

Com base nas características esparsas dos vetores de atributos extraídos pelos modelos apresentados nas seções 3.1.2 e 3.1.4 do presente capítulo, Gao e colaboradores (11) propõem a utilização do modelo de classificação de representações esparsas (*Sparse Representation Classification* - SRC). O objetivo subjacente a esta proposta é aproveitar as características discriminativas decorrentes da esparsidade (45) representando a amostra de teste como uma combinação linear das amostras de treinamento (amostras da galeria).

Para compreender a ideia original do método é preciso considerar o caso no qual se têm mais do que uma imagem por pessoa na galeria, por exemplo, n_i imagens para o *i*-ésimo sujeito. Considerando ainda que cada imagem seja representada por um vetor $x \in \mathbb{R}^d$ formado pela concatenação das colunas da imagem, é possível criar então uma matriz $A_i = [x_{i,1}, x_{i,2}, \dots, x_{i,n_i}] \in \mathbb{R}^{d \times n_i}$, na qual cada coluna representa uma amostra do *i*-ésimo sujeito da galeria.

Admite-se que qualquer nova imagem de teste $y \in \mathbb{R}^d$ pertencente ao sujeito i, vai estar no espaço definido pelas amostras de treinamento associadas a este sujeito, podendo ser expressa por

$$y = \alpha_{i,1}x_{i,1} + \alpha_{i,2}x_{i,2} + \cdots + \alpha_{i,n_i}x_{i,n_i} \quad \textbf{3-19}$$

onde $\alpha_{i,j} \in \mathbb{R}, j = 1, 2, \dots, n_i$.

Dado que na maioria das aplicações práticas a identidade da amostra de teste y incialmente é desconhecida, é preciso reformular a representação de y em função de todas as amostras de treinamento.

Assim, constrói-se uma matriz A justapondo-se as n amostras pertencentes aos k sujeitos do conjunto de treinamento, ou seja,

$$A = [A_1, A_2, \dots, A_k] = [x_{1,1}, x_{1,2}, \dots, x_{k,n_k}] \quad \textbf{3-20}$$

e y então, é redefinida em função de A segundo a equação 3-21:

$$y = A\alpha_0 \quad \in \mathbb{R}^d \quad \textbf{3-21}$$

onde $\alpha_0 = [0, \dots, 0, \alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,n_i}, 0, \dots, 0]^T$ é um vetor de coeficientes no qual todos os elementos são zeros, exceto aqueles associados ao *i*-ésimo sujeito.

Como as entradas do vetor α_0 codificam a identidade da amostra de teste y, a identificação pode ser obtida solucionando-se o sistema de equações lineares $y = A\alpha$. Obviamente, se $d > n$ o sistema de equações é sobredeterminado, o que convencionalmente é resolvido calculando o mínimo da norma euclidiana:

$$(\ell^2): \quad \hat{\alpha}_2 = \min ||\alpha||_2 \qquad \text{sujeito a} \quad A\alpha = y \tag{3-22}$$

No entanto, a solução $\hat{\alpha}_2$ não é especialmente informativa para o reconhecimento de y, dado que geralmente $\hat{\alpha}_2$ contém um grande número de entradas com valores absolutos significativamente maiores que zero. Esta é então a motivação para tentar encontrar uma solução esparsa para o sistema $y = A\alpha$ através da otimização de:

$$(\ell^1): \quad \hat{\alpha}_1 = \min ||\alpha||_1 \qquad \text{sujeito a} \quad A\alpha = y \tag{3-23}$$

onde $|| \cdot ||_1$ denota a norma ℓ^1.

É importante destacar que a busca de uma solução esparsa sobre um sistema sobredeterminado constitui um problema complexo do ponto de vista computacional, o que limita a utilização de outras alternativas como a norma ℓ^0, a qual seria ideal na solução do problema em questão.

Tendo agora uma nova imagem de teste y pertencente ao sujeito i do conjunto de treinamento, o cálculo da correspondente representação esparsa $\hat{\alpha}_1$ de y via 3-23, torna possível a identificação do sujeito, levando em consideração que idealmente as entradas de $\hat{\alpha}_1$ diferentes de zero estarão associadas com as colunas de A que representam o sujeito i. Contudo, ruído e erros intrínsecos ao modelo fazem com que algumas entradas de $\hat{\alpha}_1$, associadas com outros sujeitos que não o i, apresentem valores diferentes de zero, de modo que a identificação por esta via pode não ser inequívoca.

Para contornar esta dificuldade, adota-se como critério para a classificação de y uma medida de quão boa é a reprodução da amostra de teste a partir dos coeficientes associados com todas as amostras de treinamento de cada sujeito. A seguir, o algoritmo 2 resume o procedimento de classificação.

Algoritmo 2. Classificação baseada em Representação Esparsa (SRC).

Entrada: Matriz de amostras de treinamento $A = [A_1, A_2, \dots, A_k] \in \mathbb{R}^{d \times n}$
Amostra de teste $y \in \mathbb{R}^d$

1: Normalização das colunas de A com norma unitária ℓ^2
/Resolver o problema de minimização ℓ^1
2 $\hat{\alpha}_1 = \min_\alpha ||\alpha||_1$: sujeito a $A\alpha = y$
/Calcular os resíduos
3: **For** $i = 1, \dots, k$ **do**
4: $r_i(y) = ||y - A\vartheta_i(\hat{\alpha}_1)||_2$
5: **end For**
6: **Saída:** identidade $(y) = \min_i r_i(y)$

No algoritmo anterior, é utilizada a função $\vartheta_i: \mathbb{R}^n \to \mathbb{R}^n$ que seleciona os coeficientes associados com o i-ésimo sujeito da galeria. Assim, para $\hat{\alpha}_1 \in \mathbb{R}^n, \vartheta_i(\hat{\alpha}_1) \in \mathbb{R}^n$ é um novo vetor em que todos os elementos são nulos, com exceção dos associados ao sujeito i, cujos valores coincidem com os elementos correspondentes de $\hat{\alpha}_1$.

Voltando para o cenário SSPP, a mesma metodologia pode ser aplicada, mas com pequenas modificações para adaptá-la ao novo contexto. Dado que a matriz de treinamento $A \in \mathbb{R}^{d \times k}$ muda a sua dimensão, passando a ter uma amostra para cada sujeito, o vetor $\hat{\alpha}_1 \in \mathbb{R}^k$ apresentará apenas um coeficiente não nulo para cada classe ou sujeito. De igual forma, $\vartheta_i(\hat{\alpha}_1) \in \mathbb{R}^k$ selecionará o coeficiente associado com o sujeito i para posteriormente calcular o resíduo $r_i(y)$ com o qual será identificada a amostra de teste y.

4
Reconhecimento de faces utilizando Múltiplas Amostras por Pessoa na Prova

O reconhecimento de faces baseado nas imagens coletadas em sequências de vídeo constitui hoje um dos problemas mais desafiadores em visão computacional. A baixa qualidade das imagens que compõem o vídeo em combinação com as frequentes falhas dos algoritmos de detecção (12) na localização dos pontos fiduciais da face, impactam negativamente os métodos de reconhecimento.

Neste contexto, muitas técnicas têm sido propostas com o objetivo de tirar vantagem das múltiplas amostras que podem ser coletadas de uma pessoa nos vídeos. Por exemplo, Manuwar e coautores (41) propõem um esquema de votação por maioria no qual consideram as imagens de face detectadas em todos os quadros. Com o mesmo objetivo de explorar as Múltiplas Amostras por Pessoa no momento da prova (MSPPP), Xiaoming e colaboradores (46) utilizam as Cadeias Ocultas de Markov (*Hidden Markov Models* - HMM) para modelar o comportamento dinâmico das faces de um mesmo indivíduo ao longo dos quadros do vídeo, enquanto Kuang-Chi e coautores (47) empregam aproximações Bayesianas.

Tendo como base as aproximações acima mencionadas, um conjunto de funções de decisão é proposto neste capítulo, as quais, em combinação com os métodos de extração de atributos LBP e SSAE, também exploram a possibilidade de haver muitas amostras para a pessoa a ser identificada.

4.1.
Reconhecimento de faces em vídeos com SSAE

Os resultados alcançados por Gao e colaboradores (11) com o SSAE para imagens estáticas anotadas manualmente, encorajaram o presente estudo do comportamento do método para imagens de faces detectadas e enquadradas

automaticamente. O Algoritmo 3 apresenta em detalhe o procedimento de reconhecimento proposto nesta dissertação.

Algoritmo 3. Reconhecimento de face em vídeos com SSAE

Entrada: Conjunto de imagens da galeria: $X_{gal} = \left\{x_{gal}^{(i)}\right\}$, para $i = 1, \dots, G_{gal}$
Modelo SSAE $\theta = \left\{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\right\}$

1: **For** $i = 1, \dots, G_{gal}$ **do**
/Cálculo da representação SSAE das faces da galeria
2: $z_{gal}^{(i)} \leftarrow s(W^{(2)} s\left(W^{(1)} x_{gal}^{(i)} + b^{(1)}\right) + b^{(2)})$
/Inicialização do conjunto de dissimilaridade
3: $\delta_{prova}^{(i)} = \emptyset$
4: **end For**
5: **For** cada nova amostra $\tilde{x}_t$ **do**
/Calcular a representação SSAE
6: $z_t \leftarrow s(W^{(2)} s\left(W^{(1)} \tilde{x}_t + b^{(1)}\right) + b^{(2)})$
7: **For** $i = 1, \dots, G_{gal}$ **do**
8: /Calcular a dissimilaridade com a entrada ***i*** da galeria
9: $\delta_{prova}^{(i)} \leftarrow \delta_{prova}^{(i)} \cup D\left(z_t, z_{gal}^{(i)}\right)$
10: **end For**
/Seleção da identidade
11: $L_{prova} = \phi\left(\left\{\delta_{prova}^{(i)}\right\}\right)$
12: **Saída:** L_{prova}
13: **end For**

Para SSAE com duas camadas ocultas, a representação $z_{gal}^{(i)}$ da imagem de face da galeria $x_{gal}^{(i)}$ é calculada por:

$$z_{gal}^{(i)} \leftarrow s\left(W^{(2)} s\left(W^{(1)} x_{gal}^{(i)} + b^{(1)}\right) + b^{(2)}\right) \quad \textbf{4-1}$$

para $i = 1, \dots, G_{gal}$, onde G_{gal} é o número de sujeitos que compõem a galeria.

Assim que uma nova face, denotada daqui para diante no texto com o subscrito *prova,* é detectada pela primeira vez em uma sequência de vídeo, um conjunto $\delta_{prova}^{(i)}$ é criado para cada entrada i da galeria. Este conjunto acumula as dissimilaridades entre esta entrada e a face de prova.

O reconhecimento pode ser executado para cada novo quadro do vídeo ou só depois que todos os quadros onde aparece a mesma pessoa tenham sido coletados. Em qualquer caso, se $\tilde{x}_t$ é uma face capturada no quadro *t*, sua representação z_t pode ser calculada pela mesma função do lado direito da equação 4-2, especificamente:

$$z_t \leftarrow s\left(W^{(2)} s\left(W^{(1)} \tilde{x}_t + b^{(1)}\right) + b^{(2)}\right) \quad \textbf{4-2}$$

A dissimilaridade entre a imagem $\tilde{x}_t$ e uma das imagens da galeria é calculada pela função $D(\cdot)$, representada por alguma das funções de dissimilaridade estabelecidas na seção 3.3 e cujos argumentos são as representações da amostra z_t e da imagem da galeria $z_{gal}^{(i)}$ associadao ao i-ésimo sujeito. Finalmente estes valores de dissimilaridade acumulados em $\delta_{prova}^{(i)}$, são a base do procedimento de identificação da prova a partir das técnicas apresentadas na seção seguinte.

4.2. Funções de decisão

Baseado nos valores de dissimilaridade acumulados ao longo da sequência de vídeo para cada sujeito da galeria, a função $\phi\left(\left\{\delta_{prova}^{(i)}\right\}\right)$determina a identidade da pessoa presente na cena a partir de uma das seguintes formulações: votação por maioria, melhor escore, mediana dos escores, que são descritas a seguir.

4.2.1. Votação por maioria

A cada quadro t um voto é atribuído ao sujeito da galeria mais semelhante à prova $\tilde{x}_t$. Ao final, atribui-se à *prova* a identidade mais votada. Formalmente, o voto $v_t^{(i)}$ do quadro *t* para o sujeito *i* da galeria é definido como

$$v_t^{(i)} = \begin{cases} 1, \text{se } i = \underset{j}{\operatorname{argmin}}\left(D\left(z_t, z_{gal}^{(j)}\right)\right) \\ 0, \text{caso contrário} \end{cases}, \quad \text{para } i = 1, \ldots, G_{gal} \qquad \textbf{4-3}$$

A identidade da prova L_{prova} até o quadro *t* é dada pela equação

$$L_{prova} = \phi\left(\left\{\delta_{prova}^{(i)}\right\}\right) = \operatorname{argmax}_i \sum_t v_t^{(i)}. \qquad \textbf{4-4}$$

4.2.2. Melhor escore

Primeiramente, determina-se para cada sujeito da galeria o menor valor de dissimilaridade entre todas as amostras coletadas até o quadro corrente. Formalmente,

$$S_t^{(i)} = \min_t D\left(z_t, z_{gal}^{(i)}\right), \text{ para } i = 1, \ldots, G_{gal} \quad \textbf{4-5}$$

onde $S_t^{(i)}$ contém o menor valor de dissimilaridade da entrada i da galeria até o quadro t.

A identidade da *prova* é atribuída ao sujeito da galeria que obteve o melhor escore (menor dissimilaridade) entre as menores dissimilaridades até o quadro *t*, ou seja,

$$L_{prova} = \phi\left(\left\{\delta_{prova}^{(i)}\right\}\right) = \min_i\left(S_t^{(i)}\right) \quad \textbf{4-6}$$

4.2.3. Mediana dos escores

A identidade da *prova* é atribuída ao sujeito da galeria cuja mediana das dissimilaridades computadas até o quadro *t* é mínima. Formalmente, determina-se para cada sujeito da galeria a mediana $M_t^{(i)}$ dos valores de dissimilaridade até o quadro t:

$$M_t^{(i)} = \underset{t}{\text{median}}\, D\left(z_t, z_{gal}^{(i)}\right), \text{ para } i = 1, \ldots, G_{gal} \quad \textbf{4-7}$$

E, da mesma forma que nas formulações anteriores, a identidade da imagem de prova incide sobre o sujeito da galeria com o menor valor da mediana:

$$L_{prova} = \phi\left(\left\{\delta_{prova}^{(i)}\right\}\right) = \min_i\left(M_t^{(i)}\right) \quad \textbf{4-8}$$

Como já foi expressado no início do capitulo, o objetivo de cada uma destas aproximações é aproveitar as múltiplas amostras que um sujeito pode apresentar ao longo do vídeo. Espera-se que esta multiplicidade de amostras possa mitigar os efeitos das variações na aparência das faces ao longo do vídeo.

É importante ressaltar que, para o caso do operador LBP, o Algoritmo 3 também pode ser utilizado, assim como as estratégias propostas, dado que o LBP também utiliza uma medida de dissimilaridade na identificação ou verificação da imagem de face.

5
Análise Experimental

Neste capítulo, descreve-se de forma detalhada o procedimento experimental adotado nesta dissertação. Primeiro, são expostas as características dos bancos de dados usados no desenvolvimento desta pesquisa. Em seguida, descreve-se a metodologia experimental empregada para avaliar a eficiência dos métodos propostos. Por fim, os resultados obtidos são apresentados, analisados e interpretados.

5.1.
Descrição dos Bancos de Dados

Neste trabalho foram utilizados quatro bancos de dados: dois de imagens estáticas (CMU-PIE - CMU *Pose, Illumination and Expression* (PIE)) (48) e Extended Yale B (49)) e dois de sequências de vídeos (Honda/UCSD (47) e VIDTIMIT (50)). Todos os bancos estão disponíveis publicamente. Suas características gerais são apresentadas a seguir.

5.1.1.
CMU-PIE Database

A base de dados CMU-PIE[1] é composta por 41.368 imagens de 68 sujeitos, cada um contendo entre 600 e 615 imagens (48) com uma resolução de 640×480 pixels. Esta base encontra-se organizada em duas partições, a primeira contém variações de pose e iluminação e a segunda apresenta variações de pose combinadas com expressões faciais.

As variações de iluminação e pose da primeira partição foram captadas em dois momentos, um com as luzes da sala ligadas e o outro com as luzes desligadas, combinado ainda com os efeitos do flash das câmeras fotográficas. No caso da combinação de pose com expressão facial, foram registradas condições

[1]Disponível em: http://www.ri.cmu.edu/research_project_detail.html?project_id=418

com a face neutra, sorrisos, olhos fechados e falando. O caso particular de pessoas usando óculos também foi considerado.

De forma geral, para cada sujeito, as imagens representam um total de 13 poses diferentes, aliadas a 43 condições de iluminação e 4 expressões faciais diferentes. A Figura 5-1 apresenta exemplos das imagens faciais de um indivíduo sob estas condições. Na primeira coluna à esquerda mostram-se exemplos de imagens frontais, com expressão neutra do rosto e boas condições de iluminação. O restante exemplifica algumas das variações que o banco contém.

Figura 5-1: Exemplo das variações presentes no banco de dados CMU-PIE.

5.1.2. Extended Yale B Database

A base de dados Extended Yale B[2] (49) contém um total de 16.768 imagens de 38 indivíduos. De forma semelhante ao CMU-PIE (48), as imagens representam 64 variações nas condições de iluminação e 9 tipos de pose. A principal diferença em relação à base anterior é que esta não contém expressões faciais.

Os 10 primeiros sujeitos, os quais pertencem à base original Yale B, apresentam 64 imagens cada e contém variações combinadas de iluminação e pose, sendo utilizados somente nas avaliações. Os 28 restantes, contêm entre 500 e 560 imagens, sob as mesmas condições e foram destinados para o treinamento

[2]Disponível em: http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

do *auto-encoder*. Exemplos das imagens podem ser observados na Figura 5-2. A primeira coluna à esquerda mostra exemplos de imagens frontais, com expressão neutra do rosto e boas condições de iluminação. O restante exemplifica algumas das variações que o banco contém, com predominância dos diferentes padrões de iluminação. Observa-se também a diminuição de exemplos com variações na pose, em comparação com CMU-PIE (48).

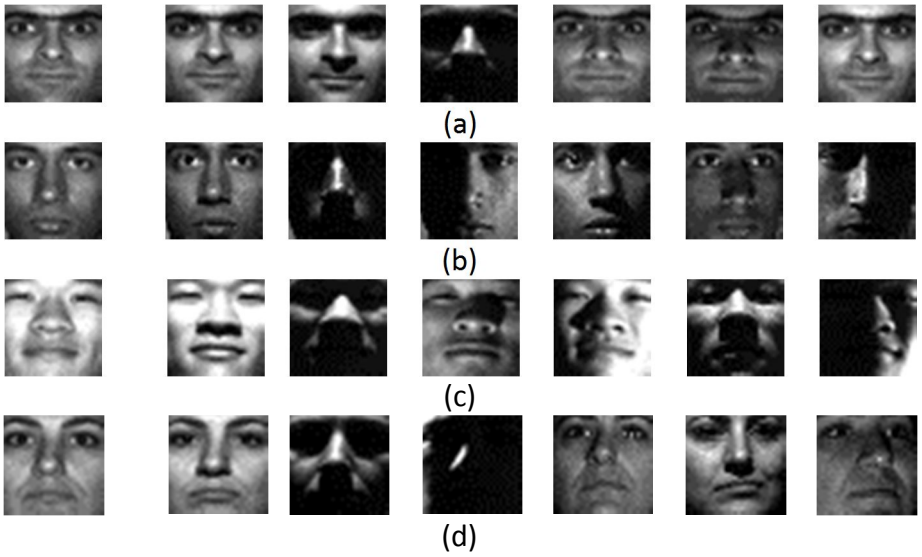

Figura 5-2: Exemplos de imagens faciais de base de dados Extended Yale B.

5.1.3. Honda/UCSD Video DataBase

Honda/UCSD[3] (47) constitui um dos bancos de vídeos mais utilizados na literatura em relação ao reconhecimento de faces. O mesmo dispõe de 59 vídeos distribuídos por 19 pessoas. De forma geral, o número de vídeos por pessoa varia de 2 a 5 e o número de quadros em cada vídeo de 92 a 645. Assim, o número de imagens de face coletadas para cada sujeito varia de 92 a 1149.

Todos os vídeos do banco apresentam rotações significativas da cabeça, grandes mudanças na escala, oclusões parciais da face e consideráveis variações de iluminação. Estes fatores fazem de Honda/UCSD (47) a base de dados mais desafiadora entre as utilizadas neste trabalho. A Figura 5-3 apresenta exemplos desta base, onde a primeira coluna à esquerda mostra exemplos de imagens

[3]Disponível em: http://vision.ucsd.edu/~leekc/HondaUCSDVideoDatabase/HondaUCSD.html

frontais, com expressão neutra do rosto e boas condições de iluminação. As outras representam uma única amostra entre as imagens coletadas dos diferentes vídeos.

Figura 5-3: Imagens coletadas na base de vídeos Honda/UCSD.

Outro detalhe a ser considerado nas bases de vídeos está relacionado ao comprimento das sequências de quadros de cada sujeito. A Figura 5-4 apresenta esta relação para o presente banco de dados.

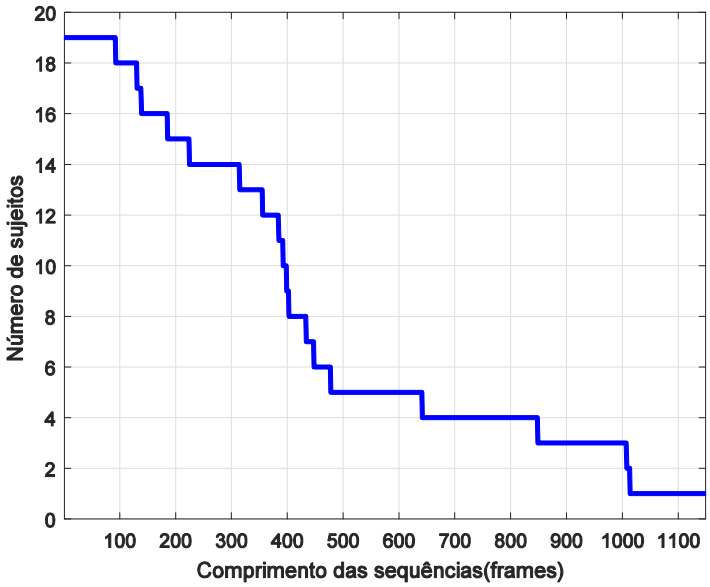

Figura 5-4: Número de sujeitos em função do comprimento das sequências de vídeo em Honda/UCSD.

O gráfico mostra a quantidade de sujeitos da base de dados com sequências de vídeos de comprimento n. Por exemplo, existem 14 sujeitos com pelo menos 300 amostras na base de dados, assim como há apenas 3 com vídeos pelo menos 1000 amostras.

5.1.4. VIDTIMIT Audio-Video Dataset

VIDTIMIT[4] (50) compreende vídeos de 43 pessoas com 3 sequências cada, gravadas em momentos diferentes. O número médio de quadros por vídeo é de aproximadamente 100 com uma resolução de 512×384 pixels. Desta forma, aproximadamente 300 imagens de faces são coletadas para cada sujeito. Em cada vídeo, a pessoa balança a cabeça para à esquerda, para à direita, de volta ao centro, para cima, para baixo, e finalmente de volta para o centro.

Da mesma forma que nos exemplos anteriores, a Figura 5-5 apresenta exemplos coletados nesta base. Já a Figura 5-6 apresenta o número de indivíduos da base de dados em função do número de amostras disponíveis por indivíduo para a base VIDTIMIT.

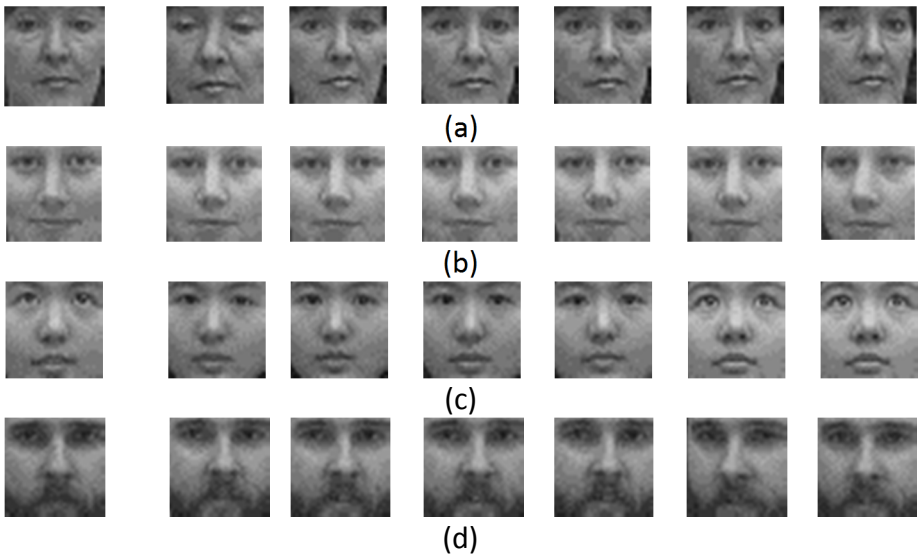

Figura 5-5: Imagens coletadas na base de vídeos VIDTIMIT.

[4]Disponível em: http://conradsanderson.id.au/vidtimit/

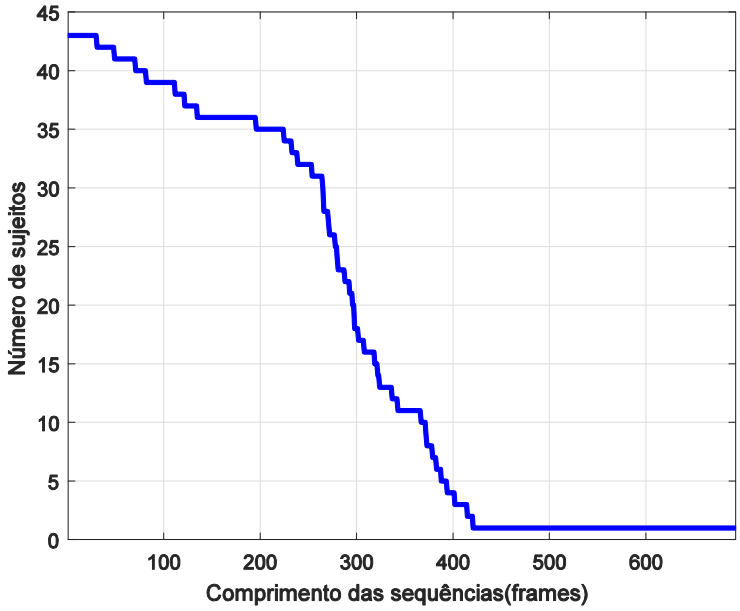

Figura 5-6: Número de sujeitos em função do comprimento das sequências de vídeo em VIDTIMIT.

Em VIDTIMIT as variações em pose não são muito significativas em comparação com as existentes em Honda/UCSD. Além disto, expressões faciais e condições de iluminação são mais controladas e não há oclusões.

5.2. Experimentos

Em linhas gerais, em cada base de dados as imagens faciais foram detectadas automaticamente usando o algoritmo Viola & Jones (12), empregando para isto os treinamentos que OpenCV (*Open Source Computer Vision Library*) disponibiliza para detectar rostos, olhos, nariz e a boca.

Em seguida, cada uma das faces detectadas foi normalizada geometricamente para a dimensão de 32×32 pixels a partir das coordenadas dos olhos detectados pelo Viola & Jones. A normalização é feita de tal forma que o centro dos olhos se mantenha sempre nas mesmas coordenadas nas imagens normalizadas.

Nos experimentos relacionados com o *Stacked Supervised Auto-encoder* (SSAE), foi utilizada uma arquitetura de duas camadas escondidas com 1024 unidades cada uma, de acordo com o sugerido em (11). As camadas foram treinadas usando o algoritmo *Backpropagation* (32) e o método de gradiente

conjugado Polack-Ribiere (51). Este último substitui o algoritmo *Limited Memory Broyden Fletcher Goldfarb Shanno* (L-BFGS) utilizado em (11), devido à menor complexidade do primeiro em relação ao número de parâmetros. Adicionalmente, foi utilizada como função de ativação a *sigmoide* ao invés *tanh* sugerida em (11). Esta escolha decorreu da observação empírica deque o treinamento com a *sigmoide* apresentou uma convergência mais rápida, bem como as redes resultantes registaram resultados ligeiramente superiores.

Todos os experimentos foram realizados a partir de um protótipo criado em MATLAB. As implementações do algoritmo *Backpropagation* e do método de gradiente conjugado Polack-Ribiere em MATLAB, foram obtidas de Salakhutdinov e coautores[5] (32). A versão de MATLAB empregada foi a R2016a (9.0.0.341) de 64 bits, executando sobre Windows 7 em um computador Intel(R) Core(TM) i7-3960x CPU @ 3.30 GHz (12 CPUs) com 24 GB de memória RAM.

A seguir se descrevem os experimentos realizados e os resultados coletados em cada um deles. Para cada experimento apresentam-se os objetivos estabelecidos e os detalhes do procedimento experimental.

5.2.1. Experimento 1

Objetivo: Avaliar a sensibilidade do SSAE em relação à solução inicial dos pesos.

Protocolo experimental: Este experimento foi realizado usando a base de dados Extended Yale B. Para o treinamento da arquitetura foram utilizadas as amostras correspondentes a 28 sujeitos totalizando 8.648 imagens, enquanto que para o teste foram empregados 10 sujeitos com 64 imagens cada. A redução da quantidade de imagens no treinamento provém das falhas na detecção da face por parte do algoritmo Viola & Jones.

O parâmetro de similaridade λ da função de perda descrita em 3-8 foi definido com valor de 5, enquanto o coeficiente de regularização λ_{wd} tomou o valor de 10^{-2}. O parâmetro de esparsidade λ_{sp} e o coeficiente ρ_0, ambos relacionados com a divergência Kulback-Leibler, foram definidos em 10^{-4}

[5]Disponível em: http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html

e 10^{-3}, respectivamente. Estes valores estão próximos aos valores ótimos, como se verá mais adiante na seção que descreve o Experimento 2.

No caso do treinamento, para cada sujeito do conjunto, uma imagem frontal com boas condições de iluminação e sem oclusões foi tomada para a galeria. O resto das imagens foram consideradas como dados corrompidos (vide Algoritmo 1). Para o teste, a escolha da imagem dos sujeitos da galeria foi feita da mesma forma, mas neste caso foi composta por cem sujeitos, sendo 10 da Extended Yale B (conjunto de teste), 19 de Honda/UCSD, 43 de VIDTIMIT e 28 selecionados aleatoriamente de CMU-PIE e Extended Yale B (conjunto de treinamento). Dos sujeitos testados, a imagem adotada para formar a galeria foi retirada do conjunto de prova.

As dissimilaridades entre as representações obtidas pelo SSAE foram calculadas pelas quatro métricas apresentadas na seção 3.3. Neste experimento, cada imagem dos sujeitos foi tratada como uma amostra a ser reconhecida. A taxa de reconhecimento de cada sujeito foi computada separadamente e então, a taxa de reconhecimento global foi obtida a partir da média das taxas de cada sujeito.

A combinação treinamento-teste foi executada 50 vezes, empregando em cada rodada uma inicialização diferente dos pesos. Conforme descrito em (11), (6) e (34), estes valores iniciais dos pesos foram aleatoriamente amostrados entre$\left[-\sqrt{\frac{6}{d_x+d_h}}, \sqrt{\frac{6}{d_x+d_h}}\right]$, onde d_xe d_h são as dimensões do vetor de entrada e da camada escondida, respectivamente. Para os viés b e $\hat{b}$, vetores nulos foram tomados como valores iniciais.

Resultados e discussão: A Figura 5-7 apresenta os resultados obtidos neste experimento fazendo uso de um diagrama de caixa (*boxplot*).

Primeiramente, observa-se que embora as taxas de reconhecimento sejam muito baixas, os resultados obtidos tiveram uma dispersão aproximada de 2% para todas as métricas avaliadas, além de um número de valores atípicos (*outliers*) muito pequeno em cada uma destas. Por outro lado, o grau de simetria dos resultados ao redor da mediana e entre os valores máximo e mínimo de cada caixa foi elevado.

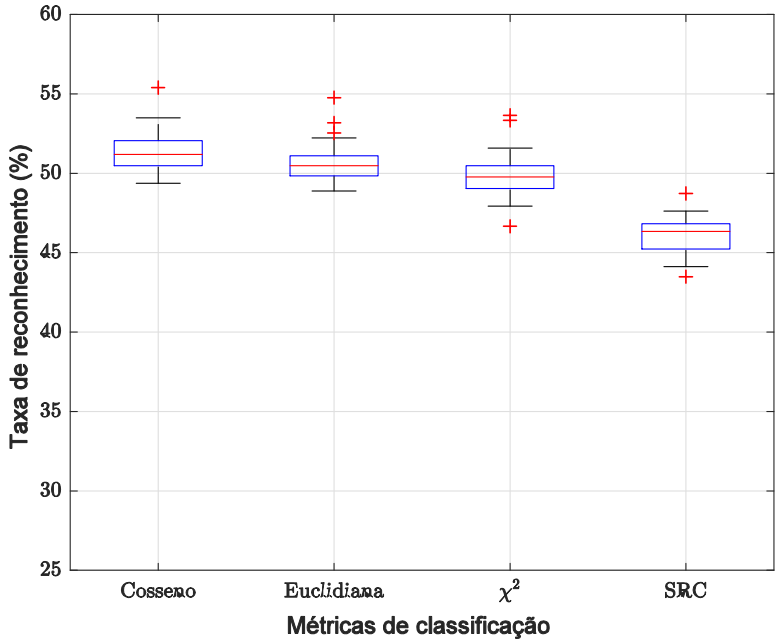

Figura 5-7: Dispersão da taxa de reconhecimento para cada uma das métricas de classificação.

Dado que os resultados em termos de taxa de reconhecimento são consequência direta do conjunto de parâmetros $\{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$ encontrados em cada rodada do algoritmo, e com base nas considerações quanto à dispersão destes resultados, fica demonstrado que o método de treinamento é pouco sensível à escolha dos pesos/viés.

Finalmente, a partir de uma comparação do desempenho de cada uma das métricas de dissimilaridade, constatou-se a distância cosseno apresentou os melhores resultados, enquanto o SRC teve o pior desempenho.

5.2.2. Experimento 2

Objetivo: Avaliar a sensibilidade do modelo em relação aos valores dos parâmetros da função objetivo.

Protocolo experimental: Este experimento foi realizado utilizando novamente a base de dados Extended Yale B. A distribuição dos conjuntos de treinamento e teste foi a mesma que no experimento anterior: 8.648 imagens de 28 sujeitos no treinamento e 640 imagens de 10 sujeitos no teste.

Para cada sujeito do conjunto de treinamento foi adotada a imagem frontal com boas condições de iluminação e sem oclusões para a galeria. As demais imagens foram consideradas como dados corrompidos (vide Algoritmo 1). No teste, a galeria foi formada por cem sujeitos, cada um com uma imagem escolhida sob o mesmo procedimento descrito anteriormente, sendo 10 de Extended Yale B (conjunto de teste), 19 de Honda/UCSD, 43 de VIDTIMIT e 28 selecionados aleatoriamente de CMU-PIE e Extended Yale B (conjunto de treinamento). Dos sujeitos testados, a imagem considerada para formar a galeria foi retirada do conjunto de prova.

As dissimilaridades entre as representações obtidas por SSAE foram calculadas pelas métricas apresentadas na seção 3.3. Da mesma forma que no experimento anterior, cada imagem dos sujeitos foi tratada como uma amostra a ser reconhecida, sendo computada a taxa de reconhecimento de cada sujeito de forma separada. Em seguida, a taxa de reconhecimento global foi obtida a partir da média das taxas por sujeito.

Segundo sugerido em (11), (6) e (34), os valores iniciais dos pesos foram aleatoriamente amostrados entre $\left[-\sqrt{\frac{6}{d_x+d_h}}, \sqrt{\frac{6}{d_x+d_h}}\right]$, onde d_x e d_h são as dimensões do vetor de entrada e da camada escondida respectivamente. Para os viés b e $\hat{b}$, vetores nulos foram tomados como valores iniciais e 10^{-3} foi o valor de ρ.

Finalmente, um conjunto de sete possíveis valores foi considerado para cada um dos parâmetros λ, λ_{wd} e λ_{sp} da função objetivo descrita por 3-8. Desta forma, um total de 343 configurações foram treinadas e testadas. A Tabela 5-1 relaciona estes valores para cada parâmetro.

λ	λ_{wd}	λ_{sp}
0	0	0
10^{-1}	10^{-7}	10^{-7}
1	10^{-6}	10^{-6}
5	10^{-5}	10^{-5}
10^{1}	10^{-4}	10^{-4}
10^{2}	10^{-3}	10^{-3}
10^{3}	10^{-2}	10^{-2}

Tabela 5-1: Conjunto de possíveis valores dos parâmetros da função de perda.

Resultados e discussão: Das 343 configurações de parâmetros treinadas e testadas, foi escolhida aquela que obteve a taxa de reconhecimento mais alta usando as quatro métricas empregadas.

Uma vez tendo sido determinada a configuração de maior desempenho, isto é, $\lambda = 10^{-1}, \lambda_{wd} = 10^{-7}$ e $\lambda_{sp} = 10^{-4}$, analisou-se o comportamento da taxa de reconhecimento variando-se os valores de cada parâmetro por separado, mantendo os outros dois, fixos nos valores ótimos. A Figura 5-8 mostra estes resultados.

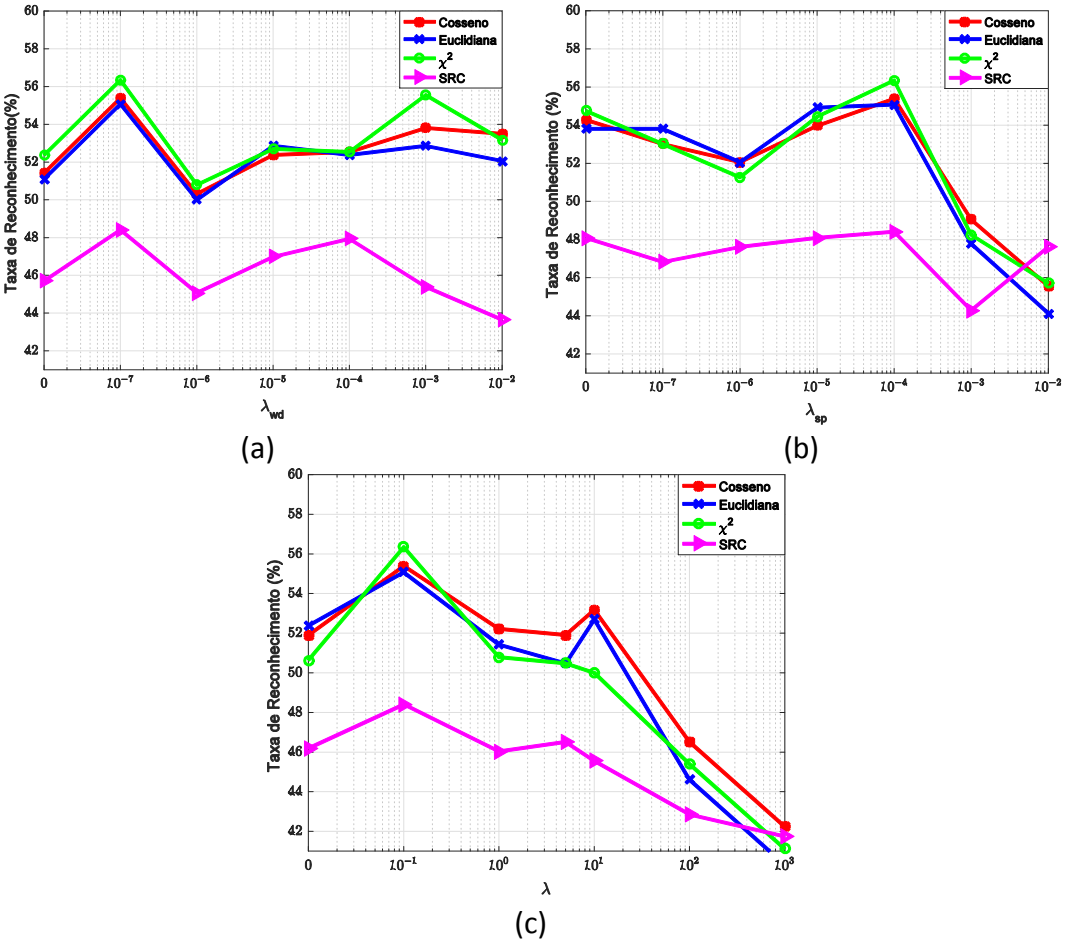

Figura 5-8: Taxa de reconhecimento em função dos parâmetros (a) λ_{wd}, (b) λ_{sp} e (c) λ.

Cabe notar que o máximo desempenho não foi significativamente superior ao que se obteve com as outras configurações testadas. Ademais, os resultados foram consistentes com o reportado por Gao e coautores (11), como se discute a seguir.

Em primeiro, o valor do parâmetro de esparsidade λ_{sp} foi exatamente igual ao reportado em (11), obtendo o melhor desempenho em 10^{-4} e o pior no

valor 10^{-2}, provando assim a importância do termo de esparsidade na função de perda.

Em segundo, nota-se que o termo de regularização λ_{wd} foi significativamente inferior ao termo de esparsidade, sugerindo que o termo da função de perda que favorece pesos baixos não é muito relevante. Esta conclusão é coerente com (11) que não inclui este termo na função de custo.

Em relação ao termo de similaridade λ, observa-se que o melhor desempenho é obtido em 10^{-1}, enquanto o pior é para 10^{3}, tendo um comportamento monotonamente decrescente entre estes valores.

Por outro lado, notou-se que o ganho apresentado pelo SSAE para o valor de $\lambda = 10^{-1}$em relação à $\lambda = 0$ não é muito significativo, o que leva a questionar a importância deste termo na função de custo. No entanto, é demonstrado em (11), que o termo de similaridade não impacta de forma considerável o desempenho do SSAE em condições de pouca variabilidade da pose, situação que também acontece na base de dados utilizada neste experimento.

Por último, nota-se que o padrão do experimento anterior é repetido, dado que o pior resultado, do ponto de vista das métricas de dissimilaridade, foi obtido pelo SRC, ao mesmo tempo que as restantes obtiveram desempenhos ligeiramente melhores em relação ao experimento 1.

5.2.3. Experimento 3

Objetivo: Determinar a influência do número de amostras de treinamento sobre a taxa de reconhecimento.

Protocolo experimental: Assim como nos experimentos 1 e 2, para cada sujeito do conjunto de treinamento a imagem frontal com boas condições de iluminação e sem oclusões foi adotada para a galeria. O resto das imagens foram consideradas como dados corrompidos (veja o Algoritmo 1). Para o teste, a galeria novamente esteve composta por cem sujeitos cada um com uma imagem nas mesmas condições das experiências anteriores, sendo 10 de Extended Yale B (conjunto de teste), 19 de Honda/UCSD, 43 de VIDTIMIT e 28 selecionados aleatoriamente de CMU-PIE e Extended Yale B (conjunto de treinamento). A

imagem dos sujeitos testados, tomada para formar a galeria, foi igualmente retirada do conjunto de prova.

As dissimilaridades entre as representações obtidas pelo SSAE foram calculadas pelas métricas apresentadas na seção 3.3. Aplicando o mesmo procedimento que nos experimentos anteriores, cada imagem de prova foi tratada como uma amostra para ser reconhecida, a partir da qual foi computada a taxa de reconhecimento de cada sujeito em separado, para logo em seguida obter a taxa de reconhecimento global a partir da média das taxas por sujeito.

Os valores iniciais dos pesos foram aleatoriamente amostrados entre$\left[-\sqrt{\frac{6}{d_x+d_h}}, \sqrt{\frac{6}{d_x+d_h}}\right]$, onde d_xe d_h são as dimensões do vetor de entrada e da camada escondida, respectivamente. Para os viés b e $\hat{b}$, vetores nulos foram tomados como valores iniciais. Segundo os resultados do experimento 2, o parâmetro de similaridade λ, foi definido em 10^{-1}, enquanto o coeficiente de regularização λ_{wd} em 10^{-7}. O parâmetro de esparsidade λ_{sp} e o coeficiente ρ_0, ambos relacionados com a divergência Kulback-Leibler, foram definidos em 10^{-4} e 10^{-3} respectivamente.

Novamente foi utilizada a base de dados Extended Yale B para este experimento. A distribuição dos conjuntos de treinamento e teste foi de 8.648 imagens de 28 sujeitos no treinamento e 640 imagens de 10 sujeitos no teste. Dado que este experimento pretende avaliar o desempenho do SSAE em função do número de amostras de treinamento, subconjuntos de 1000, 2000, 4000 e 8000 amostras foram utilizados para tal fim.

Resultados e discussão: Os resultados são apresentados na Figura 5-9.

Tal qual esperado, a efetividade do SSAE em termos da taxa de reconhecimento foi monotonamente crescente em todo o intervalo de avaliação. Em termos simples, esta observação deixa claro que quanto maior for o número de amostras no treinamento, melhor será o desempenho da rede ou de qualquer outra metodologia de aprendizado. Esta conclusão, evidentemente não surpreende, mas os resultados dão ideia de que o método tem potencial de alcançar taxas substancialmente mais elevadas se treinado com bases de dados maiores.

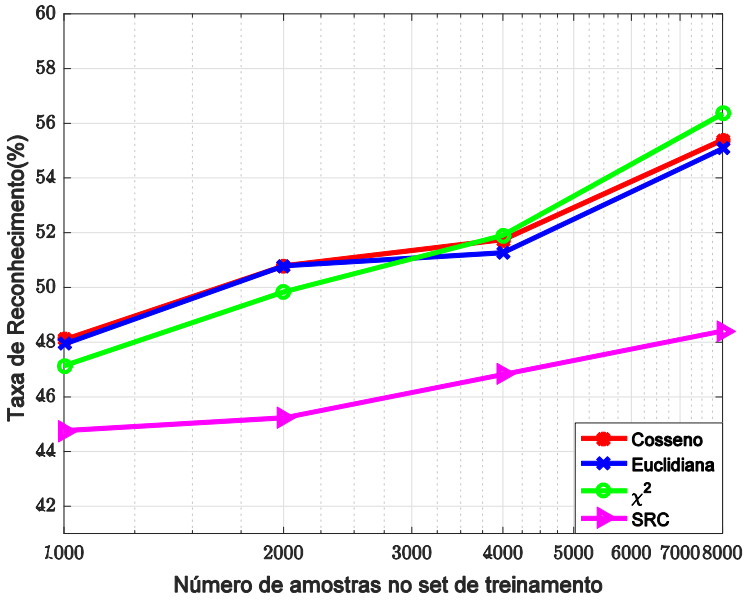

Figura 5-9: Desempenho do SSAE em função do número de amostras de treinamento.

Por fim, embora o resultado de todas as métricas tenha sido baixo de uma maneira geral, o SRC, assim como nas experiências 1 e 2, foi ainda pior. Uma possível explicação para este fraco desempenho reside na pouca quantidade de amostras que a galeria contém (uma amostra por sujeito), o que constitui um fator negativo para este classificador, pois quanto maior for a quantidade de amostras por sujeito, maior poder de discriminação terá o classificador.

5.2.4. Experimento 4

Objetivos: Este experimento apresenta os seguintes objetivos

a) Determinar a taxa de acerto da proposta MSPP em função do número de amostras da *prova*.

b) Avaliar o desempenho do método MSPPP para diversas métricas.

Protocolo experimental: Para este experimento as bases de dados CMU-PIE e Extended Yale B foram utilizadas como conjuntos de treinamento. O CMU-PIE esteve composto por 19.277 imagens de 68 sujeitos e o Extended Yale B teve a mesma distribuição dos experimentos anteriores com 8.648 imagens de 28 sujeitos. A redução da quantidade de imagens dos conjuntos de treinamento é oriunda de falhas na detecção da face por parte do algoritmo Viola & Jones.

As bases Honda/UCSD e VIDTIMIT também foram empregadas no treinamento, mas de forma alternada. Ambas possuem a mesma composição descrita nas seções 5.1.3 e 5.1.4, respectivamente. De cada sujeito dos conjuntos de treinamento foi selecionada a imagem frontal com boas condições de iluminação e sem oclusões para a galeria. O resto das imagens foram consideradas como dados corrompidos (veja o Algoritmo 1).

Os valores iniciais dos pesos foram aleatoriamente amostrados entre$\left[-\sqrt{\frac{6}{d_x+d_h}}, \sqrt{\frac{6}{d_x+d_h}}\right]$, onde d_xe d_h são as dimensões do vetor de entrada e da camada escondida, respectivamente. Para os viés b e $\hat{b}$, vetores nulos foram tomados como valores iniciais. Segundo os resultados do experimento 2, o parâmetro de similaridade λ, foi definido em 10^{-1}, enquanto o coeficiente de regularização λ_{wd} em 10^{-7}. O parâmetro de esparsidade λ_{sp} e o coeficiente ρ_0, ambos relacionados com a divergência Kulback-Leibler, foram definidos em 10^{-4} e 10^{-3} respectivamente.

A avaliação foi conduzida apenas sob as bases de vídeos. Em cada caso, foram considerados os parâmetros estimados usando os dois bancos de imagens estáticos e o de vídeo restante. Cada teste foi baseado na mesma galeria, a qual compreende cem sujeitos: 19 de Honda/UCSD, 43 de VIDTIMIT e 28 selecionados aleatoriamente de CMU-PIE e Extended Yale B. Adicionalmente, as imagens destes cem sujeitos foram retiradas dos conjuntos de prova e as dissimilaridades entre as representações obtidas pelo SSAE foram computadas usando as métricas descritas em 3.3.

O LBP (3) também foi utilizado neste experimento, sendo empregada a seguinte configuração de parâmetros. Os códigos LBP para cada pixel foram obtidos sobre um círculo de raio 1 (um) pixel com centro no próprio pixel. Em seguida, a imagem resultante foi dividida em blocos não sobrepostos de dimensão 8×8 pixels, a partir dos quais foi calculado um vetor de atributos com base em seus histogramas. Finalmente, a distância χ^2 foi utilizada para medir a dissimilaridade entre estes vetores.

Resultados e discussão: Os resultados apresentados a seguir estão organizados em função dos seguintes enfoques: reconhecimento baseado no quadro e reconhecimento baseado na sequência.

- **Reconhecimento baseado no quadro**

Nesta análise, o desempenho do SSAE foi avaliado para cada uma das imagens detectadas nos quadros, considerando cada uma destas como uma única imagem de prova a ser reconhecida. Desta forma, a taxa de reconhecimento foi calculada para cada sujeito de forma separada e o desempenho global computado a partir da média das taxas dos sujeitos testados. As Tabelas 5-2, 5-3 e 5-4 mostram os resultados obtidos.

Método	Honda/UCSD	VIDTIMIT
SSAE+*Euc.*	34	62
SSAE+*Cos.*	34	62
SSAE+χ^2	34	62
SSAE+*SRC*	30	59

Tabela 5-2: Taxa de reconhecimento (%) para o enfoque baseado no quadro com o SSAE treinado em CMU-PIE.

Método	Honda/UCSD	VIDTIMIT
SSAE+*Euc.*	33	58
SSAE+*Cos.*	33	58
SSAE+χ^2	33	59
SSAE+*SRC*	28	56

Tabela 5-3: Taxa de reconhecimento (%) para o enfoque baseado no quadro com o SSAE treinado em Extended Yale B.

Método	Honda/UCSD	VIDTIMIT
SSAE+*Euc.*	22	58
SSAE+*Cos.*	22	58
SSAE+χ^2	21	59
SSAE+*SRC*	30	56

Tabela 5-4: Taxa de reconhecimento (%) para o enfoque baseado no quadro com o SSAE treinado em VIDTIMIT (testado no Honda/UCSD) e Honda/UCSD (testando no VIDTIMIT).

De acordo com os dados das tabelas, as taxas obtidas para a base de dados VIDTIMIT foram baixas, sendo ainda piores para Honda/UCSD. Resultados tão decepcionantes possuem duas razões: a típica baixa qualidade das imagens coletadas dos vídeos e a utilização de procedimentos automáticos para a detecção das faces e sua normalização. Em contraste com este procedimento, (11) apresenta relatos nos quais a precisão do SSAE foi medida utilizando imagens de rostos

manualmente anotadas e enquadradas. De fato, os autores de (11) mencionam muito brevemente o baixo desempenho do SSAE quando aplicado sobre imagens coletadas pelos algoritmos de detecção mais eficientes da literatura.

Os resultados obtidos sobre o Honda/UCSD foram piores em comparação com VIDTIMIT, pois naquela base há maiores variações em iluminação, pose e oclusões.

Em relação às métricas de dissimilaridade, nenhuma diferença substancial é observada nos resultados obtidos pelas distâncias cosseno, euclidiana e χ^2. O desempenho do SRC para a base VIDTIMIT foi semelhante ao das outras métricas, enquanto que para Honda/UCSD alternou entre os melhores e piores valores, todos, porém, consideravelmente baixos.

- **Reconhecimento baseado na sequência**

Diferente do enfoque anterior, o reconhecimento baseado na sequência estuda e expõe as melhorias que podem ser obtidas ao explorar as múltiplas amostras de uma pessoa na prova (MSPPP), coletadas a partir de sequências de vídeo.

Com esta finalidade, a função $\phi(\cdot)$ foi primeiramente aplicada para todas as subsequências de comprimento L para uma única pessoa usando o Algoritmo 3. A seguir, a taxa de reconhecimento média para o comprimento L foi calculada para essa pessoa. Este mesmo procedimento foi executado para cada sujeito a fim de se calcular posteriormente a taxa de reconhecimento média em relação a todos os sujeitos para o comprimento L. Por fim, este processo foi realizado para todos os valores possíveis de L entre 1 (equivalente ao reconhecimento baseado no quadro) e o tamanho da maior sequência de vídeo.

A Figura 5-10 mostra a taxa de reconhecimento em função do comprimento das sequências para as três funções de decisão propostas neste trabalho, usando a distância cosseno como métrica de dissimilaridade.

Os gráficos das Figura 5-10.a, 5.10.c e 5.10.e (coluna da esquerda na Figura 5.10) referem-se aos testes realizados sobre a base de dados Honda/UCSD, enquanto que as Figuras 5-10.b, 5.10.e e 5-10.f (coluna da esquerda) estão relacionadas com os testes realizados sobre a base VIDTIMIT. Cada linha da Figura 5-10 corresponde ao treinamento do modelo SSAE realizado sobre uma

diferente base de dados. Na primeira linha (5-10.a e 5-10.b) a base utilizada foi a CMU-PIE, na segunda linha (5-10.c e 5-10.d) foi a Extended Yale B e na terceira linha (5-10.e e 5-10.f) foram a Honda/UCSD e VIDTIMIT, respectivamente.

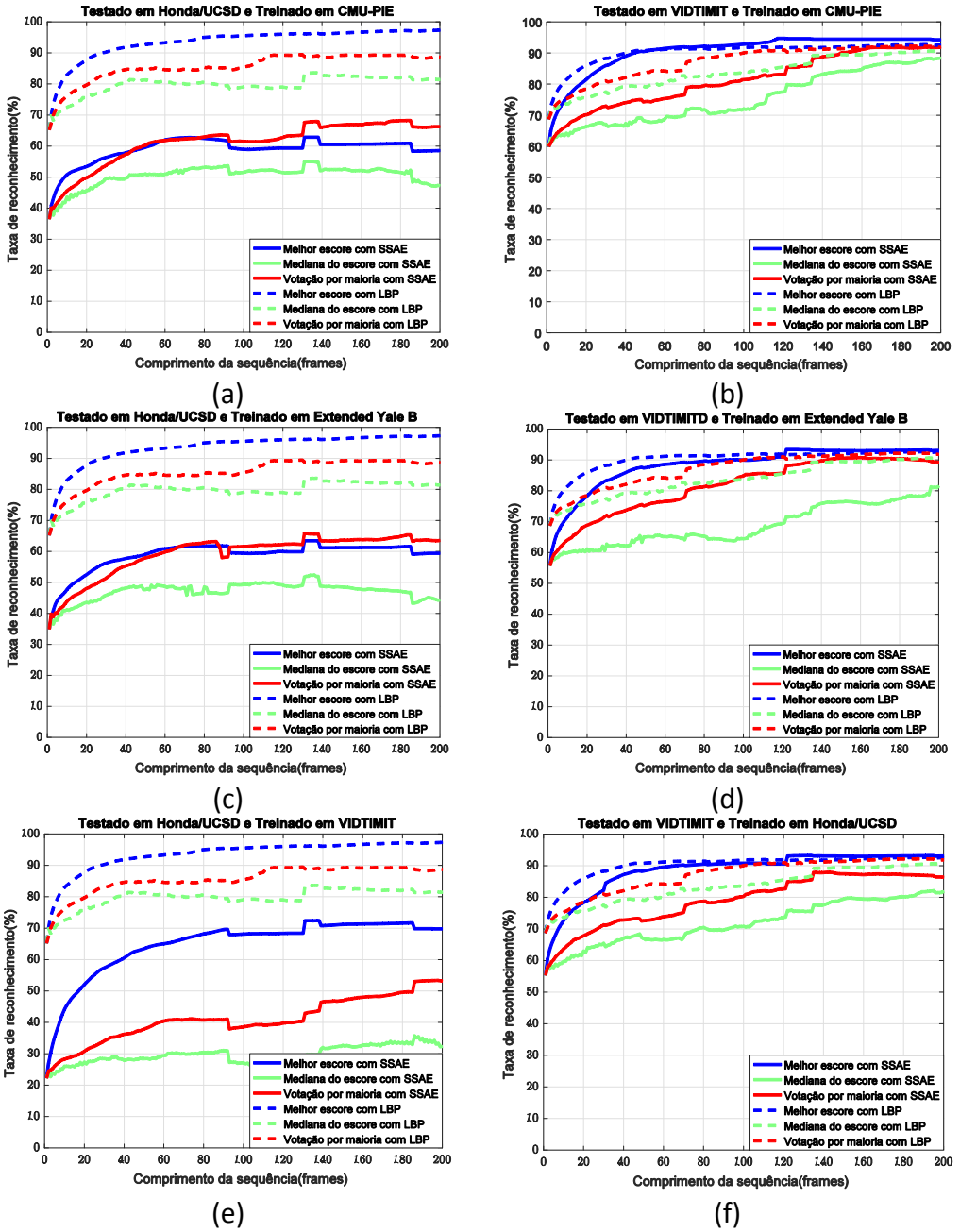

Figura 5-10: Desempenho das funções de decisão ao longo das sequências de vídeo.

A primeira observação sobre estes gráficos é que o reconhecimento baseado na sequência geralmente é muito superior ao baseado no quadro. Isto demonstra que o problema da baixa qualidade das imagens na maioria dos sistemas de

reconhecimento de face pode ser parcialmente aliviado a partir da exploração das múltiplas amostras coletadas para uma pessoa.

Nota-se ainda que a acurácia em todos os casos aumenta com o comprimento da sequência, ou seja, à medida que mais amostras são adicionadas à prova, as taxas de reconhecimento tendem a ser maiores.

Chama a atenção o fato das curvas não serem suaves; a maioria destas apresentam mudanças abruptas para alguns comprimentos das sequências. Estas mudanças ocorrem em valores de L em que se altera o número de sujeitos utilizados para se calcular as taxas médias de reconhecimento.

Por exemplo, suponha que um determinado indivíduo tenha ℓ amostras disponíveis para teste e que a taxa de reconhecimento para $L= \ell$ tenha sido calculada sobre um total de n_ℓ indivíduos. O valor da taxa de reconhecimento para $L= \ell +1$ será calculada sobre n_ℓ -1 indivíduos, o que pode provocar as mudanças abruptas na curva de $L= \ell$ para $L= \ell +1$. É interessante notar que estas mudanças ocorrem sempre nos mesmos valores de L.

Assim, taxas para comprimentos de sequência menores são calculadas empregando um maior número de sujeitos do que para as de maior comprimento.

A comparação das Figuras 5-4 e 5-6 com 5-10.a e 5-10.b, respectivamente, permite observar que as alterações nas curvas acontecem exatamente no comprimento em que o número dos sujeitos disponíveis para o cálculo da taxa de reconhecimento diminui. Dado que Honda/UCSD apresenta menor número de sujeitos que VIDTIMIT, 19 e 43 respectivamente, a mudança na primeira vai ser mais abrupta do que na última. No entanto, observa-se que o padrão é o mesmo em todas as curvas.

Outra importante observação é que de todas as funções de decisão avaliadas, a correspondente ao *melhor escore* obteve um desempenho superior em comparação com a *mediana* e a *votação por maioria*, sem exceção.

Outro aspecto importante é que o SSAE foi muito mais sensível a variações na qualidade da imagem do que o LBP, o que é claramente observado nos testes realizados em Honda/UCSD, que apresenta mais variações do que VIDTIMIT. Isto pode ser devido à robustez intrínseca do LBP para diferentes padrões de iluminação contidos nestas bases de dados combinado com o grau de frontalidade que a normalização geométrica impõe ao utilizar a detecção dos olhos como

referencia. Contudo, para vídeos bem-comportados em termos destas variações, como os de VITIMIT, o SSAE foi superior ao LBP para o caso de sequências maiores do que 120 quadros.

5.2.5. Experimento 5

Ao longo deste estudo notou-se que os erros de identificação ocorriam em sua maioria quando a dissimilaridade entre a prova e o registro mais similar da galeria era elevada. Esta métrica, ou seja, dissimilaridade relativamente ao sujeito a quem o método atribui a identidade da prova é, portanto, um indicativo da incerteza quanto à resposta provida pelo método de reconhecimento. Neste experimento, considerou-se um modo de operação em que o sistema pode produzir a resposta "desconhecido", caso o sujeito da galeria mais similar à prova, seja ainda muito distinto desta. Sendo assim, o objetivo e o protocolo seguido neste experimento são descritos a seguir.

Objetivo: Este experimento apresenta os seguintes objetivos

a) Avaliar como a taxa de reconhecimento se comporta, caso se imponha um máximo ao valor de similaridade para o reconhecimento.
b) Avaliar a relação entre taxa de reconhecimento e de não reconhecimento no cenário em questão.

Protocolo experimental: adotou-se o mesmo protocolo experimental utilizado no Experimento 4. A diferença com o experimento anterior reside tão somente na maneira como se calcularam a métricas de desempenho.

O sistema de reconhecimento só fornece a identidade da prova, se a dissimilaridade desta com o sujeito arrolado na galeria que lhe é mais similar não exceder um determinado limiar. Caso tal condição não seja satisfeita, o sistema produz como resposta "desconhecido". Assim, as taxas de reconhecimento foram computadas excetuando-se as respostas "desconhecido". Além da taxa de reconhecimento, mediu-se ainda a taxa de não reconhecimento, definida como a proporção de respostas "desconhecido" produzidas pelo sistema.

As medidas foram contabilizadas para diferentes valores de limiares de similaridade, mais especificamente, os indicados nas tabelas 5-5 e 5-6. Estes limiares foram escolhidos com base nos histogramas dos valores de dissimilaridade. Neste experimento adotou-se tão somente a função de decisão do melhor escore dado pela distância cosseno, para o SSAE, e pela distância χ^2 para o LBP. As figuras 5-11, 5-12 e 5-13 apresentam estes resultados.

Limiares	Honda/UCSD	VIDTIMIT
Limiar 1	9.9×10^{-6}	2.1×10^{-5}
Limiar 2	1.1×10^{-5}	2.3×10^{-5}
Limiar 3	1.3×10^{-5}	2.5×10^{-5}
Limiar 4	1.5×10^{-5}	2.7×10^{-5}
Limiar 5	1.7×10^{-5}	2.9×10^{-5}
Limiar 6	1.9×10^{-5}	3.1×10^{-5}
Limiar 7	2.1×10^{-5}	3.3×10^{-5}

Tabela 5-5: Conjunto de limiares utilizados para o SSAE em combinação com a métrica do cosseno.

Limiares	Honda/UCSD	VIDTIMIT
Limiar 1	148	134
Limiar 2	154	140
Limiar 3	160	146
Limiar 4	166	152
Limiar 5	172	158
Limiar 6	178	164
Limiar 7	184	170

Tabela 5-6: Conjunto de limiares utilizados para o LBP em combinação com a métrica χ^2.

A Figura 5-11 apresenta os testes realizados na base de dados Honda/UCSD, onde a primeira coluna (Figuras 5-11.a, 5-11.c e 5-11.e) contém as taxas de reconhecimento e a segunda coluna (Figuras 5-11.b, 5-11.d e 5-11.f) as taxas de pessoas não reconhecidas. Nas linhas encontram-se representadas cada uma das bases de dados utilizadas no treinamento do SSAE. A mesma distribuição vale para a Figura 5-12 que apresenta os resultados dos testes em relação a base VIDTIMIT. A Figura 5-13, por sua vez, apresenta os gráficos para o LBP, neste caso para ambos os bancos de dados.

Primeiramente, observa-se que o resultado de maneira geral continua superior ao obtido no enfoque baseado nos quadros. Neste caso, contudo, tanto o

SSAE como o LBP apresentam desempenhos ainda melhores para algumas valores do limiar de dissimilaridade. De um modo geral, ao se aumentar o limiar, tanto as taxas de reconhecimento quanto a de não reconhecidos diminuem. Nota-se, por outro lado, que em alguns casos, a taxa de reconhecimento se estabiliza a partir de certo comprimento da (Figuras 5-12 e 5-13).

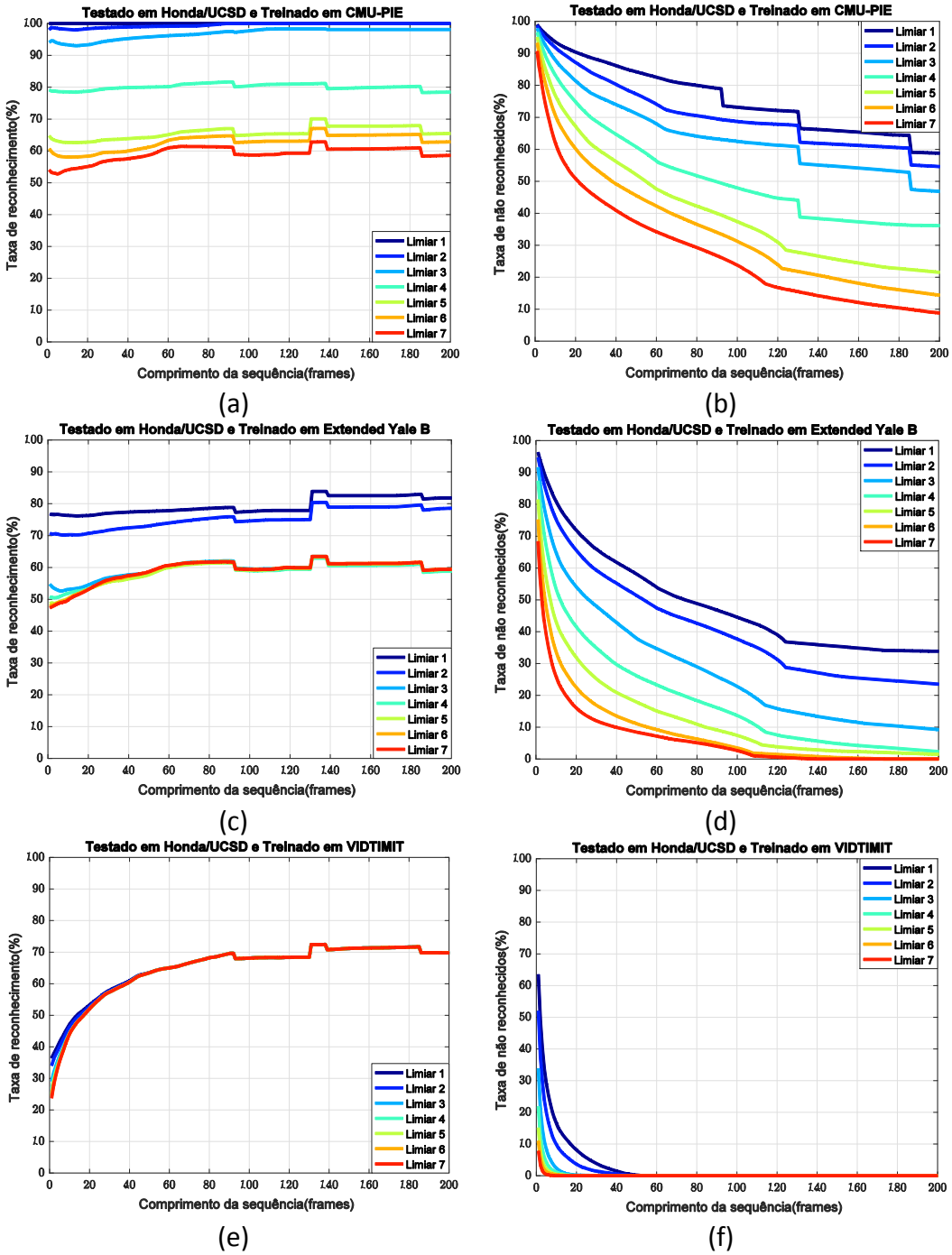

Figura 5-11: Desempenho do SSAE em função dos limiares no banco de dados Honda/UCSD.

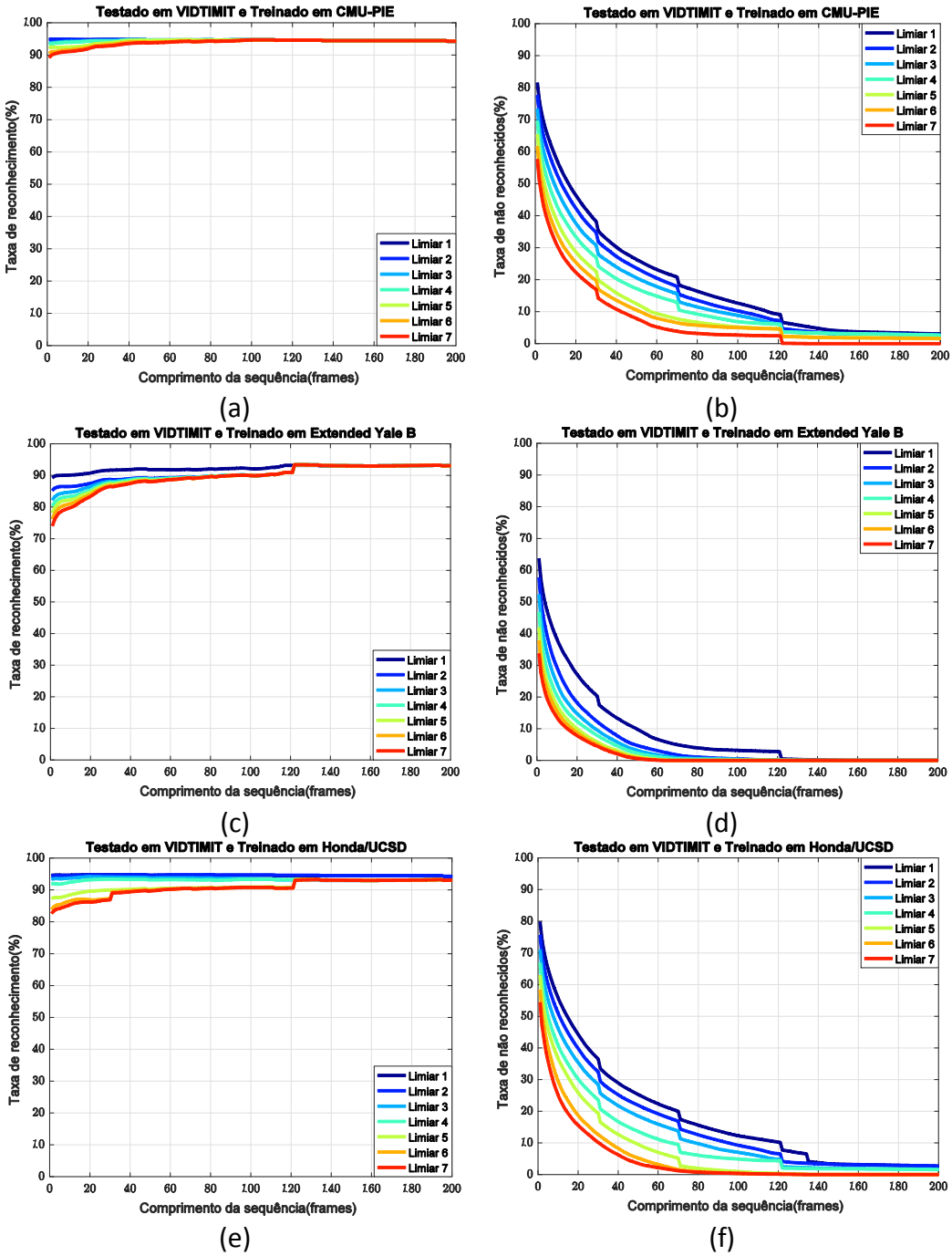

Figura 5-12: Desempenho do SSAE em função dos limiares no banco de dados VIDTIMIT.

Nota-se também que o comportamento da taxa de não reconhecidos em função do comprimento da sequência foi monotonamente decrescente. Isto demonstra que as probabilidades de reconhecimento de uma pessoa aumentam na mesma medida que mais amostras são coletadas para compor a *prova*. Em relação

ao efeito degrau das curvas, estes resultados apresentam o mesmo comportamento observado no Experimento 4, sendo ainda mais visível nas curvas das taxas de não reconhecidos.

Mesmo utilizando conjuntos de limiares diferentes para cada banco de vídeos, é perceptível que as variações entre as curvas que representam cada limiar são muito menores em VIDTIMIT do que em Honda/UCSD. Isto está novamente relacionado com as grandes variações contidas na base Honda/UCSD.

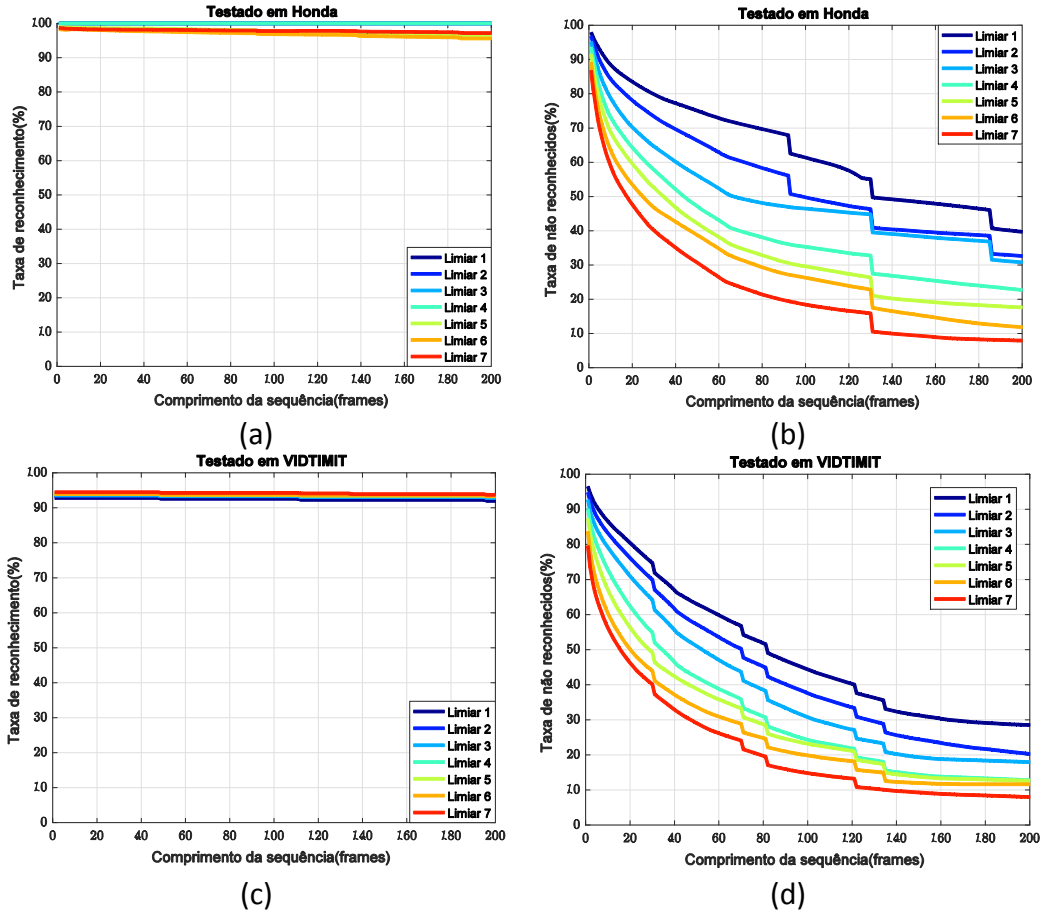

Figura 5-13: Desempenho do LBP em função dos limiares.

Cabe por fim mencionar, que as taxas de reconhecimento obtidas com o LBP (Figura 5-13 à esquerda) foram as melhores registradas. No entanto, as taxas de não reconhecimento (Figura 5-13 à direita) foram piores (mais altas) do que as obtidas com o SSAE. Há que se considerar neste caso, que os valores de limiar adotados para o LBP e para o SSAE são diferentes e, os resultados, não são diretamente comparáveis.

Por fim, entre os resultados produzidos pelo SSAE, os melhores foram obtidos quando treinado com a base CMU-PIE. Isto se deve principalmente ao número de amostras existentes no conjunto de treinamento.

6
Conclusões e Trabalhos Futuros

Neste capítulo apresentam-se as principais conclusões deste trabalho. Adicionalmente, indicam-se os trabalhos futuros a serem desenvolvidos na sequência desta pesquisa.

6.1.
Discussão

Neste trabalho, avaliou-se o desempenho do *Stacked Supervised Auto-Encoder* (SSAE) em sequências de vídeos usando uma amostra por pessoa na galeria (SSPP). Foram estudados ainda os fundamentos teóricos em que se apóia este modelo, assim como as principais abordagens propostas nos últimos anos para o cenário SSPP. O algoritmo LBP foi escolhido como referência nas avaliações do SSAE neste cenário.

Nos experimentos realizados, primeiro foi estudado o desempenho do SSAE em função dos parâmetros de regularização. Notou-se que o termo de similaridade da função de perda teve um impacto moderado na taxa de reconhecimento. Já o termo relativo ao decaimento dos pesos demonstrou ter pouco impacto no processo de treinamento.

Em seguida, foram testadas três funções de decisão baseadas nas múltiplas amostras na prova (MSPPP), demonstrando-se que ambos algoritmos, LBP e SSAE, podem se beneficiar consideravelmente desta abordagem.

Em terceiro lugar, entre as funções de decisão testadas, a de melhor resultado foi claramente a do *melhor escore*, obtendo um desempenho consistente superior em todos os experimentos. Nos testes realizados utilizando o banco de dados Honda/UCSD, o qual contêm muitas variações de iluminação, o LBP foi melhor do que o SSAE. Entretanto, o SSAE superou o LBP nos experimentos com VIDTIMIT para comprimentos de sequências maiores do que 120 quadros.

Por fim, o SSAE apresentou melhores resultados quando treinado com CMU-PIE. Esta observação apoiada por outros resultados experimentais, sugerem

que o SSAE pode alcançar desempenho muito superiores aos observados neste estudo, desde que treinamento a partir de uma maior base de dados.

6.2. Trabalhos Futuros

O aprendizado de representações a partir de imagens faciais em cenários SSPP é um problema que tem atraído grande interesse nos últimos anos, razão pela qual diferentes abordagens foram propostas com este fim, como está detalhado no segundo capítulo desta dissertação. Contudo, apesar dos resultados obtidos serem encorajadores, ainda é possível alcançar melhorias caso se superem algumas deficiências observadas.

Como se constatou no capítulo sobre a Análise Experimental, o SSAE é muito sensível a variações nas condições de iluminação. Sendo assim, alternativas que tentem corrigir esta fraqueza podem ser estudas no futuro.

Além disso, dado que os experimentos foram executados em bases de dados com poucos sujeitos, experimentos em bancos de dados com maior quantidade de pessoas estão sendo planejados como continuação desta pesquisa.

Por outro lado, a partir da observação de que a normalização geométrica favorece certo grau de frontalidade nas faces detectadas, experimentos que não utilizem esta normalização serão realizados como continuação deste trabalho.

Finalmente, experimentos utilizando um comitê de redes também serão avaliados no futuro.

Referências bibliográficas

1 SU, Y. et al. Adaptive generic learning for face recognition from a single sample per person, 2010.

2 TAN, X. et al. Face recognition from a single image per person: A survey. Pattern recognition, v. 39, n. 9, p. 1725-1745, 2006.

3 OJALA, T.; PIETIK; MENP. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 24, n. 7, p. 971-987, 2002.

4 OJANSIVU, V.; HEIKKIL. Blur insensitive texture classification using local phase quantization. In: _____ Image and signal processing. [S.l.]: Springer, 2008. p. 236-243.

5 LE, Q. V. Building high-level features using large scale unsupervised learning. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. [S.l.]: [s.n.]. 2013. p. 8595-8598.

6 GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. International conference on artificial intelligence and statistics. [S.l.]: [s.n.]. 2010. p. 249-256.

7 KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. [S.l.]: [s.n.]. 2012. p. 1097-1105.

8 RIFAI, S. et al. Higher order contractive auto-encoder. In: _____ Machine Learning and Knowledge Discovery in Databases. [S.l.]: Springer, 2011. p. 645-660.

9 RIFAI, S. et al. Contractive auto-encoders: Explicit invariance during feature extraction. Proceedings of the 28th international conference on machine learning (ICML-11). [S.l.]: [s.n.]. 2011. p. 833-840.

10 VINCENT, P. et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research, v. 11, p. 3371-3408, 2010.

11 GAO, S. et al. Single sample face recognition via learning deep supervised autoencoders. Information Forensics and Security, IEEE Transactions on, v. 10, n. 10, p. 2108-2118, 2015.

12 VIOLA, P.; JONES, M. J. Robust real-time face detection. International journal of computer vision, v. 57, n. 2, p. 137-154, 2004.

13 TURK, M. A.; PENTLAND, A. P. Face recognition using eigenfaces. Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. [S.l.]: [s.n.]. 1991. p. 586-591.

14 YANG, J. et al. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 26, n. 1, p. 131-137, 2004.

15 JOLLIFFE, I. Principal component analysis. [S.l.]: Wiley Online Library, 2002.

16 WU, J.; ZHOU, Z.-H. Face recognition with one training image per person. Pattern Recognition Letters, v. 23, n. 14, p. 1711-1719, 2002.

17 CHEN, S.; ZHANG, D.; ZHOU, Z.-H. Enhanced (PC) 2 A for face recognition with one training image per person. Pattern Recognition Letters, v. 25, n. 10, p. 1173-1181, 2004.

18 BELHUMEUR, P. N.; HESPANHA, J. P.; KRIEGMAN, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 19, n. 7, p. 711-720, 1997.

19 GAO, Q.-X.; ZHANG, L.; ZHANG, D. Face recognition using FLDA with single training image per person. Applied Mathematics and Computation, v. 205, n. 2, p. 726-734, 2008.

20 MART. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 24, n. 6, p. 748-763, 2002.

21 SHAN, S. et al. Extended Fisherface for face recognition from a single example image per person. Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on. [S.l.]: [s.n.]. 2002. p. II--81.

22 CHEN, S.; LIU, J.; ZHOU, Z.-H. Making FLDA applicable to face recognition with one sample per person. Pattern recognition, v. 37, n. 7, p. 1553-1555, 2004.

23 BRUNELLI, R.; POGGIO, T. Face recognition: Features versus templates. IEEE Transactions on Pattern Analysis \& Machine Intelligence, n. 10, p. 1042-1052, 1993.

24 ROWLEY, H. A.; BALUJA, S.; KANADE, T. Neural network-based face detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 20, n. 1, p. 23-38, 1998.

25 MANJUNATH, B. S.; CHELLAPPA, R.; VON DER MALSBURG, C. A feature based approach to face recognition. Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on. [S.l.]: [s.n.]. 1992. p. 373-378.

26 LEE, T. S. Image representation using 2D Gabor wavelets. IEEE Transactions on pattern analysis and machine intelligence, v. 18, n. 10, p. 959-971, 1996.

27 LYONS, M. et al. Coding facial expressions with gabor wavelets. Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. [S.l.]: [s.n.]. 1998. p. 200-205.

28 LADES, M. et al. Distortion invariant object recognition in the dynamic link architecture. Computers, IEEE Transactions on, v. 42, n. 3, p. 300-311, 1993.

29 WISKOTT, L. et al. Face recognition by elastic bunch graph matching. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 19, n. 7, p. 775-779, 1997.

30 IZENMAN, A. J. Linear discriminant analysis. In: _____ Modern Multivariate Statistical Techniques. [S.l.]: Springer, 2013. p. 237-280.

31 HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. Neural computation, v. 18, n. 7, p. 1527-1554, 2006.

32 SALAKHUTDINOV, R.; MNIH, A.; HINTON, G. Restricted Boltzmann machines for collaborative filtering. Proceedings of the 24th international conference on Machine learning. [S.l.]: [s.n.]. 2007. p. 791-798.

33 BENGIO, Y. Learning deep architectures for AI. Foundations and trends

34 BENGIO, Y. Practical recommendations for gradient-based training of deep architectures. In: _____ Neural Networks: Tricks of the Trade. [S.l.]: Springer, 2012. p. 437-478.

35 ERHAN, D. et al. Why does unsupervised pre-training help deep learning? The Journal of Machine Learning Research, v. 11, p. 625-660, 2010.

36 TANG, Y.; SALAKHUTDINOV, R.; HINTON, G. Deep lambertian networks. arXiv preprint arXiv:1206.6445, 2012.

37 KAN, M. et al. Stacked progressive auto-encoders (spae) for face recognition across poses. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: [s.n.]. 2014. p. 1883-1890.

38 DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. [S.l.]: John Wiley \& Sons, 2012.

39 HINTON, G. E.; ZEMEL, R. S. Autoencoders, minimum description length, and Helmholtz free energy. Advances in neural information processing systems, p. 3-3, 1994.

40 OLSHAUSEN, B. A.; OTHERS. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, v. 381, n. 6583, p. 607-609, 1996.

41 HAYAT, M.; BENNAMOUN, M.; AN, S. Deep reconstruction models for image set classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on, v. 37, n. 4, p. 713-727, 2015.

42 POULTNEY, C. et al. Efficient learning of sparse representations with an energy-based model. Advances in neural information processing systems. [S.l.]: [s.n.]. 2006. p. 1137-1144.

43 CHAN, C.-H.; KITTLER, J.; TAHIR, M. A. Kernel fusion of multiple histogram descriptors for robust face recognition. In: _____ Structural, Syntactic, and Statistical Pattern Recognition. [S.l.]: Springer, 2010. p. 718-727.

44 BARTLETT, M. S.; MOVELLAN, J. R.; SEJNOWSKI, T. J. Face recognition by independent component analysis. Neural Networks, IEEE Transactions on, v. 13, n. 6, p. 1450-1464, 2002.

45 WRIGHT, J. et al. Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, v. 31, n. 2, p. 210-227, 2009.

46 LIU, X.; CHEN, T. Video-based face recognition using adaptive hidden markov models. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. [S.l.]: [s.n.]. 2003. p. I--340.

47 LEE, K.-C. et al. Video-based face recognition using probabilistic appearance manifolds. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. [S.l.]: [s.n.]. 2003. p. I--313.

48 SIM, T.; BAKER, S.; BSAT, M. The CMU pose, illumination, and expression (PIE) database. Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. [S.l.]: [s.n.]. 2002. p. 46-51.

49 GEORGHIADES, A. S.; BELHUMEUR, P. N.; KRIEGMAN, D. J. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE transactions on pattern analysis and machine intelligence, v. 23, n. 6, p. 643-660, 2001.

50 SANDERSON, C.; LOVELL, B. C. Multi-region probabilistic histograms for robust and scalable identity inference. In: _____ Advances in Biometrics. [S.l.]: Springer, 2009. p. 199-208.

51 KRAMER, A. H.; SANGIOVANNI-VINCENTELLI, A. Efficient parallel learning algorithms for neural networks. Advances in neural information processing systems. [S.l.]: [s.n.]. 1989. p. 40-48.

52 ZOU, W. et al. Deep learning of invariant features via simulated fixations in video. Advances in neural information processing systems. [S.l.]: [s.n.]. 2012. p. 3212-3220.

53 ZHANG, Z. et al. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. [S.l.]: [s.n.]. 1998. p. 454-459.

54 LE, Q. V. et al. ICA with reconstruction cost for efficient overcomplete feature learning. Advances in Neural Information Processing Systems. [S.l.]: [s.n.]. 2011. p. 1017-1025.

55 HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. Statistical learning with sparsity: the lasso and generalizations. [S.l.]: CRC Press, 2015.

56 GAO, Y.; QI, Y. Robust visual similarity retrieval in single model face databases. Pattern Recognition, v. 38, n. 7, p. 1009-1020, 2005.