Hugo Kenji Rodrigues Okada

Detecção Automática de Fases Temporais de Emoção em Vídeos a partir de Características da Face

Manaus - AM

Março, 2018

Hugo Kenji Rodrigues Okada

Detecção Automática de Fases Temporais de Emoção em Vídeos a partir de Características da Face

Dissertação apresentada ao Instituto de Computação da Universidade Federal do Amazonas, para a obtenção do Grau de Mestre em Informática.

Instituto de Computação – IComp

Universidade Federal do Amazonas - UFAM

Programa de Pós-Graduação em Informática

Orientador: Prof.ª Dr.ª Eulanda Miranda dos Santos

Manaus - AM

Março, 2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Okada, Hugo Kenji Rodrigues
O41d Detecção automática de fases temporais de emoção em vídeos a partir de características da face / Hugo Kenji Rodrigues Okada. 2018
64 f.: il. color; 31 cm.

Orientadora: Eulanda Miranda dos Santos
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Segmentação temporal. 2. Dinâmica temporal. 3. Análise de emoção. 4. Reconhecimento de padrões. I. Santos, Eulanda Miranda dos II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Detecção Automática de Fases Temporais de Emoção em Vídeos a partir de Características da Face"

HUGO KENJI RODRIGUES OKADA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Profa. Eulanda Miranda dos Santos - PRESIDENTE

Prof. José Reginaldo Hughes Carvalho - MEMBRO INTERNO

Prof. José Luiz de Souza Pio - MEMBRO EXTERNO

Manaus, 03 de Abril de 2018

Agradecimentos

Quero agradecer primeiramente a Deus, por ter me concedido todas as oportunidades e permitir que meus objetivos fossem concluídos, ao lado de pessoas maravilhosas como meus amigos e familiares.

Agradeço ao meu pai Yosuke Okada e à minha mãe Maria Lusilene Bueno Rodrigues, por sempre terem me mostrado o melhor caminho e ensinado como enfrentar os desafios ao longo de minha vida, e por sempre terem apoiado todas as minhas escolhas, decisões e projetos. Agradeço também aos meus amigos, por toda ajuda, apoio e trocas de ideias, que contribuíram para o desenvolvimento deste trabalho.

E claro que não podia deixar de agradecer à Prof.ª Dr.ª Eulanda Miranda dos Santos, minha orientadora e exemplo profissional, por ter me guiado com maestria e profissionalismo durante todo o processo. Aproveitando também para agradecer todos os professores do IComp.

Agradeço ao Instituto de Computação - IComp e a Universidade Federal do Amazonas - UFAM pela oportunidade oferecida.

"Tanto ao lutar quanto na vida cotidiana,
você deve ser determinado, ainda que calmo.
Vá de encontro à situação sem tensão,
mas também sem desleixo,
com o espírito estável, mas sem prejulgamentos.
(Miyamoto Musashi)

Resumo

É cada vez mais frequente o uso de técnicas computacionais para resolver problemas existentes no mundo real, como por exemplo, o reconhecimento de emoções humanas. Nesse contexto, é possível a utilização de conceitos e técnicas computacionais para analisar e identificar a emoção humana através de características extraídas de diferentes modalidades de dados, tais como face, gesto e escrita. Quando os dados são obtidos a partir de vídeos, de acordo com a literatura, qualquer sentimento humano representado é composto por quatro fases temporais que ocorrem em cinco etapas (*neutral, onset, apex, offset* e *neutral*), essas fases representam todo o ciclo de "vida"de uma emoção. Portanto, o exercício de definição das fases temporais é um passo muito importante, pois beneficia o trabalho dos sistemas de reconhecimento de emoções em vídeo. Este trabalho apresenta uma arquitetura baseada em técnicas de aparência e em similaridade voltadas para identificar de forma automática as fases temporais de emoções em vídeo considerando dados da face. Os testes contidos neste trabalho mostram que o método proposto identifica o padrão de reconhecimento das fases temporais de emoções em vídeos, a partir de dados de face, de forma independente da base de dados utilizada.

Palavras-chave: Reconhecimento de Padrões, Análise de emoções em vídeo, Fases Temporais, Métodos baseados em Aparência e em similaridade, Características de face.

Abstract

Computational techniques employed to solve real-world problems in the real world, such as the recognition of human emotions,has become more frequent. In this context, it is possible to use computational concepts and techniques to analyze and identify human emotions by applying features extracted from different data modalities such as face, gesture and writing. When data are obtained from video, according to the literature, any represented human emotion is composed of four temporal phases involving five steps (neutral, onset, apex, offset and neutral), these phases represent the entire "life" cycle of an emotion. Therefore, defining temporal phases is a very important step, since it supports video emotion recognition systems. This work presents an approach based on appearance and similarity techniques to automatically identify the temporal phases of emotions in videos taking into account face data. The experimental results provided in this work show that the proposed method is able to identify a pattern of emotions temporal phases recognition in videos based on face data features. The learned pattern is independent of the database used.

Keywords: Pattern Recognition, Video Emotion Analysis, Time Phases, Appearance-Based and Similarity-Based Methods, Face Characteristics.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AU	Action Unit
BOW	Bag Of Words
FABO	The Bimodal Face and Body Gesture Database
HMM	Hidden Markov model
HOG	Histogram of Oriented Gradients
LBP	Local Binary Patterns
MHI	Motion History Image
MMI	MMI Facial Expression
MSE	Mean Squared Error
MSM	Subspace Method (Método do Subespaço Mútuo)
MUG	Multimedia Understanding Group
PSNR	Peak Signal-to-Noise Ratio
SDK	Software Development Kit
SSIM	Structural Similarity Index (Índice de Similaridade Estrutural)
SVM	Support Vector Machine
SVMSMO	Sequential Minimal Optimization for Support Vector Machine

Sumário

1 Introdução

O termo emoção é muito utilizado e identificado no nosso dia a dia, portanto, demonstra, de acordo com o contexto do entendimento, ser algo notório e simples de ser compreendido, mas de acordo com a psicologia definir emoção não é algo tão fácil. Segundo Atkinson (2002), a expressão emoção é definida como um estado complexo e momentâneo que aparece em experiências de caráter afetivo, gerando mudanças em várias áreas do funcionamento psicológico e fisiológico, levando o indivíduo a gerar movimentos físicos.

Com o objetivo de compreender os sentimentos humanos, a visão computacional utiliza diversas estratégias para interpretar, identificar e responder às emoções humanas, podendo avaliar cada estado afetivo (GUNES; PICCARDI, 2009). Nesse contexto, na última década, a análise automática das expressões humanas tornou-se uma pesquisa importante para diversos setores, como recuperação de imagem, reconhecimento de padrões e análise de sentimentos (FASEL; LUETTIN, 2003).

Segundo Liu (2010), análise de sentimentos é o estudo de opiniões, sentimentos e emoções. Existem diversas formas de extração desses elementos, sendo que as informações podem ser obtidas a partir de diversas fontes tais como: textos, imagens ou vídeos. O grande desafio da análise de sentimentos é poder criar métodos computacionais para identificar opiniões e sentimentos. Embora muitos estudos tenham sido feitos nessa linha de pesquisa, a maioria dos trabalhos, especialmente os que investigam análise de sentimentos a partir de vídeos, utilizam apenas uma fonte de informação, também chamada modalidade, geralmente a face. No entanto, alguns estudos relatam que o uso de detecção multimodal provê melhores resultados. Algumas dessas modalidades estudadas são: voz, eletrocefalograma, batimentos cardíacos, respiração, gestos faciais e corporais.

No caso de reconhecimento de gestos, Escalera et al. (2013) relata que essa é uma tarefa muito complexa e desafiadora, devido ao fato de existirem inúmeras características envolvidas, e também algumas limitações técnicas, tais como a resolução espacial e temporal.

A literatura retrata que cada emoção exibida em vídeo é composta por quatro fases temporais: *Neutral* (onde a emoção encontra-se no estado de repouso), *Onset* (a emoção começa a ser construída), *Apex* (estado ápice da emoção) e *Offset* (fase de finalização da emoção). Existem alguns estudos que utilizam dinâmicas temporais como uma característica muito importante para a interpretação de emoções, pois estas informações possuem um significado psicológico para o estado das expressões (RUSSELL; FERNÁNDEZ-DOLS, 1997) (SCHMIDT; COHN, 2001). No entanto, apesar de sua funcionalidade, a definição das propriedades espaciais e dinâmicas de gestos faciais e corporais representam um grande

desafio na tarefa de reconhecimento de sentimentos (GUNES et al., 2015).

O reconhecimento de expressões através da face é a fonte mais usada para o reconhecimento de emoção. De acordo com Gelder (2009), 95% das pesquisas sobre emoções humanas utilizam apenas a face como modo para extrair características, e uma pequena porcentagem emprega o uso do corpo. Porém, somente nos últimos anos o uso de características multimodais passou a ser investigado com mais frequência, com o objetivo de obter melhores resultados. As expressões representam uma manifestação do estado afetivo, atividade cognitiva, intenção e personalidade de uma pessoa (DONATO et al., 1999). Mehrabian (1968), relata que um sentimento é representado 7% por texto, 38% pela voz e 55% pela expressão facial, por este ser um estudo antigo (1968), ainda não se tinha o uso de gestos corporais.

De fato, poucos estudos, tanto no contexto de dados monomodais quanto multimodais, voltados para o reconhecimento de emoções em vídeo, têm se preocupado em modelar, de forma automática, as fases temporais do comportamento das emoções. A seção seguinte apresenta o problema a ser resolvido e a justificativa para o desenvolvimento deste trabalho.

1.1 Definição do Problema e Justificativa

Schmidt e Cohn (2001), relatam que as dinâmicas temporais possuem um papel importante para a interpretação das emoções em vídeos. Entende-se que as informações sobre o tempo de uma ação, tanto facial quanto corporal, podem representar um significado psicológico, sendo este relevante para identificar os aspectos de estado de uma expressão. Com o intuito de melhorar o reconhecimento de expressões, o tratamento das fases temporais é considerado um ponto crucial. Cada expressão cresce a partir das fases temporais na seguinte ordem: *Neutral*, *Onset*, *Apex*, *Offset* e *Neutral*, dentre essas quatro, o *Apex* é o que melhor discrimina o sentimento, pois é nesta fase temporal que o sentimento atinge a sua representação máxima (CHEN et al., 2013).

No intuito de determinar as fases temporais, alguns trabalhos existentes na literatura rotularam as fases temporais de forma manual, utilizando votações humanas em busca das diferenças entre as fases (ZHANG; ZHANG, 2015), (ADSUL et al., 2010) e (GUNES; PICCARDI, 2009). Gunes e Piccardi (2009), utilizaram o rótulo manual, feito por pessoas de forma majoritária, para determinar as fases temporais de emoções dos vídeos da base de dados FABO (*The Bimodal Face and Body Gesture Database*), esta técnica precisou ser feita pois cada vídeo da base possui tempos de fases temporais diferente dos demais, essa diferença é existente pois a base de dados FABO possui vídeos com pessoas diferentes, várias emoções e iluminação variada, sendo esta uma tarefa bem demorada para ser registrada. Assim, uma forma de detecção automática das fases temporais deve também

definir as fases de modo mais rápido e preciso.

Existem na literatura métodos que realizam este procedimento de forma automática, alguns deles utilizam algoritmos que extraem as diferenças em uma área de movimento, exemplo HMM (*Hidden Markov model*) usado por Cohen et al. (2003), Otsuka e Ohya (1998) e Gunes et al. (2015), AU (*Action Unit*) utilizado por Pantic e Patras (2006) e Pantic e Patras (2004), MHI (*Motion History Image*) e divergência neutra empregado por Chen et al. (2013), essas técnicas, apesar de solucionarem o problema, possuem muitas etapas, apresentando, portanto, elevado custo computacional. Outra desvantagem dos métodos existentes na literatura é o de identificar as fases temporais apenas em uma base de dados, não encontrando o padrão entre as bases. A grande culpa desse problema é a diferença de iluminação entre as bases de dados, modificando os resultados da segmentação temporal.

Este trabalho tem como foco identificar as fases temporais das emoções, reduzindo o custo computacional. Esse processo, ao utilizar métodos baseados na aparência e em similaridade, reconhecem padrões em vídeo ou imagem e utilizam modelos matemáticos, baseados em aparência, como o valor de intensidade dos *pixels* ou níveis de histogramas.

1.2 Objetivos

Esta seção apresenta o objetivo geral e os objetivos específicos deste trabalho.

1.2.1 Objetivo Geral

Desenvolver um método que seja capaz de identificar de forma automática, as fases temporais de emoções humanas em vídeos a partir de dados adquiridos da face, mais precisamente, dados de expressões faciais.

1.2.2 Objetivos Específicos

- Adaptar métodos baseados em aparência e em similaridade para que possam ser usados para identificar mudanças entre os *frames*.

- Mensurar o grau de mudança de estado das características faciais que registram as fases temporais.

- Implementar um limiar adaptativo para identificar as mudanças das fases temporais.

1.3 Estrutura da Dissertação

Este trabalho está organizado da seguinte forma: O Capítulo 2 apresenta o referencial teórico usado na pesquisa para um melhor entendimento do trabalho. O Capítulo 3 apresenta os trabalhos relacionados. O Capítulo 4 expõe o método proposto, descrevendo cada etapa. O Capítulo 5 exibe os resultados obtidos nos experimentos. O capítulo 6 apresenta a conclusão deste trabalho e relata as propostas para trabalhos futuros.

2 Referencial Teórico

Este capítulo apresenta conceitos presentes na literatura que são importantes para o entendimento dos problemas relacionados à identificação das fases temporais em vídeos de expressões faciais para o reconhecimento de emoções. A seção 2.1 trata sobre reconhecimento de padrões, a seção 2.2 relata um sistema para o reconhecimento de pontos faciais, a seção 2.3 relata os conceitos sobre o reconhecimento de emoções em vídeo, a seção 2.4 descreve conceitos relacionados a fases temporais e a seção 2.5 apresenta os algoritmos baseados em aparência e em similaridade.

2.1 Reconhecimento de Padrões

De forma geral, existem na literatura duas maneiras de reconhecer determinado padrão (CONNELL; JAIN, 2001):

- Classificação supervisionada: onde o dado de entrada é rotulado como pertencente a uma das classes pré-definidas no problema.
- Classificação não supervisionada: onde o algoritmo reconhece e julga as características para determinar a classe do dado a ser identificado.

Portanto, o problema de reconhecer um padrão representa uma tarefa de classificação ou categorização, na qual as classes são pré-definidas ou são aprendidas com o decorrer do processo. A quantidade de estudos de reconhecimento de padrões tem crescido e suas técnicas têm sido utilizadas por diversas áreas. A tabela 1 apresenta exemplos de várias aplicações e suas áreas.

Tabela 1 – Exemplos de aplicações e áreas que utilizam o estudo de reconhecimento de padrões (SA, 2012).

Aplicação Científica	**Aplicação Industrial**	**Aplicação Médica**	**Aplicação Governamental**	**Aplicação Militar**
Astronomia	Reconhecimento de caracteres	Análise de eletrocardiogramas	Previsão meteorológica	Análise de fotografia aérea
Entomologia	Detecção de derrotas	Análise de radiografias e tomografias	Análise sísmica	Detecção e classificação de sonar
Arqueologia	Análise de assinaturas	Sistemas de diagnóstico clínico	Análise de poluição	Reconhecimento automático de alvos
Cibernética	Reconhecimento de fotografias	Análise de electroencefalogramas	Previsões econômicas	Detecção remota
Geologia	Análise e reconhecimento da fala	Exames microscópicos	Identificação de impressões digitais	Classificação e análise do radar
Educação	Visão por computador	Estudos genéticos	Determinação de crescimento urbano	

Um processo de reconhecimento de padrões possui necessariamente três fases (DUDA; HART; STORK, 2012):

- Aquisição de dados: Fase na qual os dados de entrada são recebidos como coordenadas, imagens, vídeos ou características (nome, tamanho, cor e etc). As características de um dado podem variar de acordo com os sensores de aquisição utilizados. Nesta etapa, nas abordagens clássicas também é feita a extração de características, onde os pontos de interesse presentes nos dados de entrada são selecionados;

- Representação de dados: Nesta etapa, é feita a classificação dos objetos, isto é, os dados são rotulados à um categoria descrita a partir de suas propriedades;

- Tomada de decisão: Fase responsável por tomar decisões com base nos resultados obtidos. Caso o resultado tenha pouca eficácia, pode-se voltar às etapas a fim de procurar novas características ou novos classificadores.

No trabalho de identificar as fases temporais da emoção em vídeos de face, a literatura mostra que uma etapa importante a ser realizada é a localização dos pontos principais da face, o quais são posteriormente usados como dados de entrada para extração de características. A próxima seção relata uma plataforma que realiza muito bem este objetivo.

2.2 Reconhecimento de pontos faciais

Face plus plus (PLUS, 2017), é uma plataforma composta por vários algoritmos voltados a visão computacional, permitindo utilizar tecnologias de reconhecimento e análise

da face através de imagens, baseadas em aprendizado profundo. Para identificar os pontos-chaves principais de uma face, o *Face plus plus* localiza e retorna todos os componentes do rosto, como contorno facial, contorno dos olhos, das sobrancelhas, lábios e nariz. O maior número de pontos que pode ser retornado pelo *Face plus plus* é de 106 pontos. A Figura 1 apresenta um exemplo do uso do *Face plus plus*.

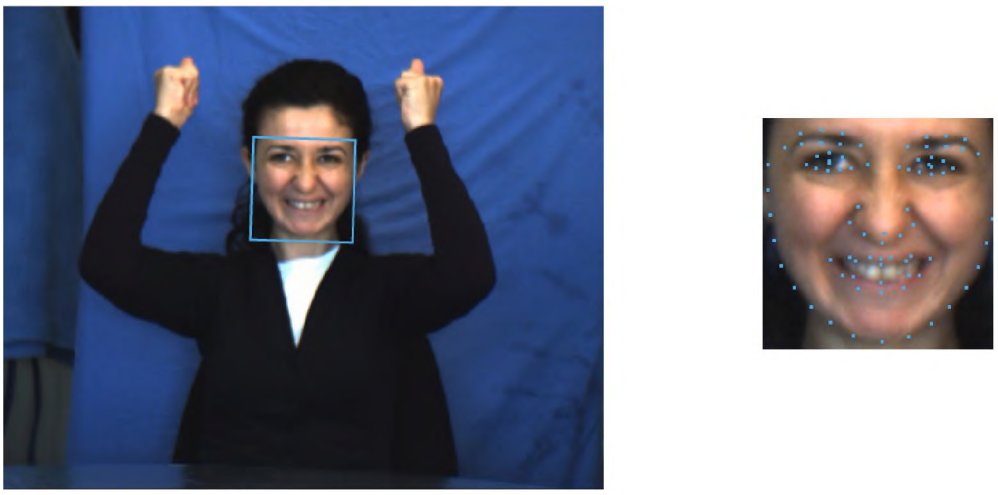

Figura 1 – Exemplo do uso *Face plus plus*, a imagem da direita detecta o rosto e a da esquerda apresenta os pontos faciais. Fonte: Adaptado de Gunes et al. (2015).

A plataforma *Face plus plus* está disponível *on-line* e de uso gratuito através do link *https://www.faceplusplus.com/*.

Reconhecimento de emoções em vídeo é o problema de identificação de padrões investigado neste trabalho. A próxima seção discute esse tipo de aplicação, apresentando suas características e seu uso.

2.3 Reconhecimento de emoções em vídeo

O termo "emoção"é o conjunto genérico de estados afetivos, os quais representam sentimentos (JAQUES; VICARI, 2005). Uma expressão (de emoção) é definida pela demonstração a outras pessoas, podendo ser de forma voluntária ou involuntária (PICARD; PICARD, 1997). A duração de uma emoção é normalmente de poucos segundos.

Existem muitas formas de uma emoção ser manifestada, e diversos são os métodos existentes para reconhecer a emoção transmitida. A voz, as ações do usuário em um sistema, as expressões faciais, os gestos corporais e os sinais fisiológicos (respiração, batimento cardíaco e etc) são exemplos de características utilizadas na literatura para o reconhecimento de emoções (OLIVEIRA; JAQUES, 2013). Essas propriedades podem ser usadas de três modos:

- Monomodal ou unimodal: Segundo Caridakis et al. (2007), neste modo ocorre a análise de apenas uma única forma de expressão de emoção;
- Bimodal: neste caso acontece a análise e união de duas formas de expressão de emoção;
- Multimodal: quando são usadas mais de duas formas de expressão de emoção.

O reconhecimento de emoções em vídeo a partir de expressões faciais, tem sido pesquisado com o uso de três tipos de base de dados: emoções representadas, onde, conforme solicitado, a expressão facial é emitida; emoções induzidas, por exemplo, ao assistir um vídeo; e emoções espontâneas, onde não existe uma prévia de qual sentimento será representado. Conforme esperado, as bases de dados com emoções representadas permitem que os algoritmos de classificação de emoções alcancem os melhores resultados, pois o seu processo apresenta as características existentes no estado emocional investigado e os padrões representados apresentam pouca variância.

Um vídeo que representa uma emoção por meio de expressão facial é composto pelo seguinte processo de evolução das fases temporais: *neutral, offset, apex, onset* e *neutral*. Então, para um melhor apoio para reconhecer os estados afetivos, se torna útil a detecção das fases temporais.

A seção a seguir trata sobre as fases temporais existentes em vídeos que contêm representação de emoções.

2.4 Fases Temporais

Um sentimento em vídeo é composto por quatro fases e cinco etapas. Essas etapas representam o início e o fim da ação, são chamadas de fases temporais e acontecem na seguinte ordem, *neutral, onset, apex, offset e neutral*.

A Figura 2 demonstra na emoção "felicidade"alguns *frames* de cada fase temporal.

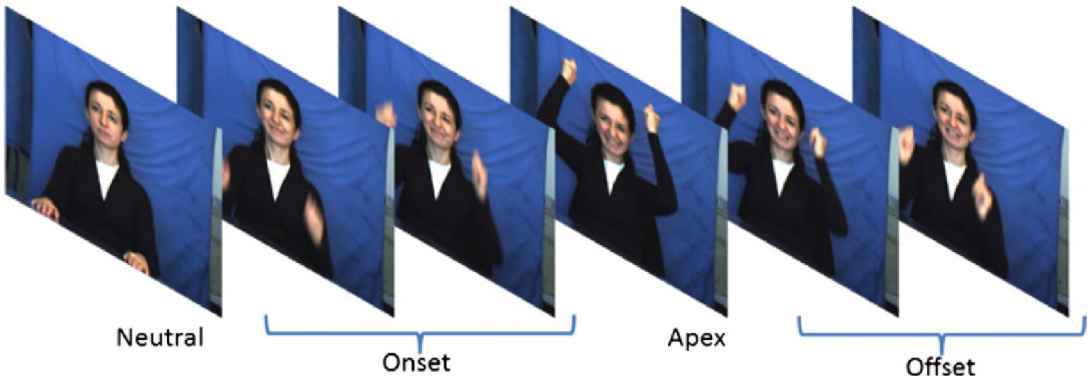

Figura 2 – Apresentação de alguns *frames* de cada fase temporal de acordo com o crescimento da emoção "felicidade". Este vídeo faz parte da base de dados FABO. Fonte: Adaptado de Gunes et al. (2015).

Com base em Koelstra e Pantic (2008), cada fase temporal ocorre a seguinte ação:

- *Neutral*: Esta fase não possui sinais de mudança da ação, ocorre no início e no fim do vídeo, pois nesse estado a expressão ainda não foi iniciada ou já foi concluída.;
- *Onset*: Onde os músculos se contraem e o aparecimento das alterações crescem. Ocorre quando a expressão está no início;
- *Apex*: Nesta etapa ocorre o ápice da ação, considerado o auge, pois é nessa etapa que ocorre o limite máximo de uma expressão;
- *Offset*: Onde os músculos estão em fase de relaxamento e retornando à fase onde a expressão encontra-se em parte de finalização *neutral*.

É importante destacar que, segundo a literatura ((GUNES et al., 2015), (PANTIC; PATRAS, 2004)), essas fases temporais não existem apenas quando emoções são representadas na forma de expressões faciais. Se, por exemplo, gestos corporais forem utilizados como fonte da dados para classificação de emoções em vídeos, a mesma sequência de fases temporais também pode ser observada. A Figura 3 apresenta o tempo de crescimento das fases temporais da mesma cena a partir de características obtidas de expressão facial e de gestos do corpo.

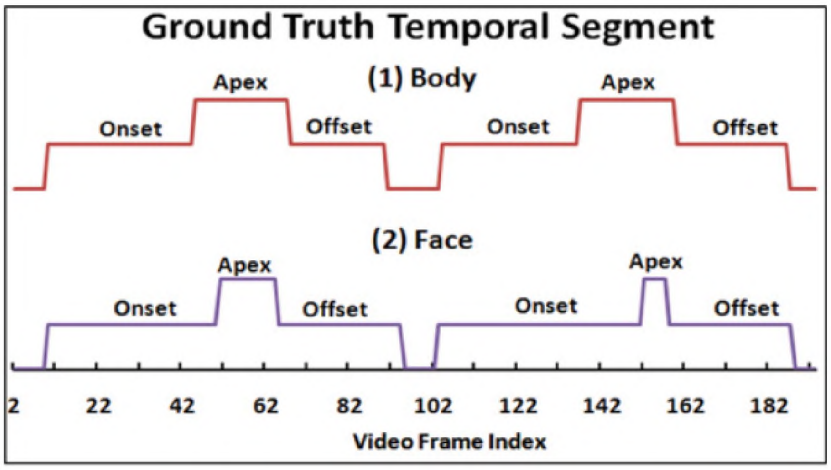

Figura 3 – Gráfico que mostra o crescimento das fases temporais no corpo (*body*) e face de acordo com o *ground truth* da base de dados FABO. Fonte: Gunes et al. (2015).

Conforme mencionado na introdução, muitos trabalhos existentes na literatura identificam as fases temporais de forma manual. Porém, existem algumas soluções que tentam realizar essa tarefa de forma automática.

Os métodos baseados em aparência e em similaridade (*appearence based methods*), são conhecidos desta forma porque utilizam os dados de entrada (imagens, vídeos) sem a extração de nenhum conhecimento ou características. Deste modo, os algoritmos não possuem conceitos de aprendizado e treinamento, pois as informações necessárias são retiradas do próprio conjunto de objetos, sem qualquer intermédio externo (LOPES; FILHO; NO, 2005). A seção a seguir descreve algumas das técnicas que utilizam esta abordagem e que são utilizadas neste trabalho.

Os algoritmos baseados em aparência e em similaridade possuem como principal função buscar um limiar de diferença entre dois *frames*. Em relação aos vídeos, buscam uma diferença de comportamento. Os métodos como MSE (*Mean Square Error*), PSNR (*Peak Signal-to-Noise Ratio*), Distância Euclidiana, Distância *City Block*, Distância de Minkowski, MSM (*Subspace Method*) e SSIM (*Structural Similarity Index*), utilizam esses quadros de comparação.

2.5 Algoritmos Baseados em aparência e em similaridade

Esta seção apresenta as técnicas que são utilizadas como parte do método proposto para detecção das fases temporais, descrevendo suas definições e funcionamento.

2.5.1 LBP (*Local Binary Pattern*)

O método LBP (OJALA; PIETIKÄINEN; HARWOOD, 1996), é um descritor com base em textura com ótimo desempenho e muito utilizado em várias aplicações. Sendo altamente sensível a pequenas mudanças, o LBP detecta variâncias de níveis de cinza.

Tem como característica atribuir um tamanho de rótulo para cada pixel a ser extraído em uma janela de tamanho 3x3 ou superior, definindo de forma uniforme a vizinhança local como um conjunto de pontos de amostragem (OJALA; PIETIKAINEN; MAENPAA, 2002). A Figura 4 apresenta um exemplo de tamanhos de janela do LBP.

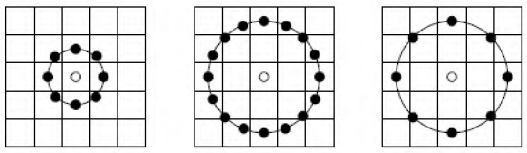

Figura 4 – Os circulos (8,1), (16,2) e (8,2) representam o número de pontos vizinhos e a distância dos pontos vizinhos ao ponto central. Fonte: Ojala, Pietikainen e Maenpaa (2002).

2.5.2 Erro Médio Quadrado

O MSE ou Erro Médio Quadrado é um método bastante conhecido e utilizado para analisar a fidelidade de uma imagem. Na maioria dos casos em que essa métrica foi utilizada, a mesma possui grande vantagem em analisar imagens que utilizam diferentes filtros (LOPES; FILHO; NO, 2005). O MSE possui a característica de reconhecer os detalhes das imagens através de alterações de energia, não perdendo a informação real da imagem. O MSE é uma medida de quão perto os pontos de dados estão da linha da curva (erro), quanto menor for o seu valor, maior será o seu ajuste para os dados.

A fórmula de cálculo do MSE esta descrita na Equação 2.1.

$$MSE = \frac{1}{m * n} \sum_{y=1}^{m} \sum_{x=1}^{n} \left[I\left(x, y\right) - I'\left(x, y\right) \right]. \tag{2.1}$$

Sendo que m e n representam as dimensões da matriz gerada das imagens, I é a imagem alvo e I' é a imagem de referência. Resultando na diferença entre as imagens.

2.5.3 PSNR (Peak Signal-to-Noise Ratio)

O PSNR é usado em muitos sistemas analógicos como uma métrica para julgar a qualidade das imagens. Devido sua baixa complexidade, o PSNR também é utilizado para medir a qualidade e avaliar algoritmos de processamento de imagem, e é considerado

uma refêrencia para o desenvolvimento de avaliações (HUYNH-THU; GHANBARI, 2008). Seu cálculo corresponde a um ajuste posterior ao MSE. A sua fórmula é representada pela Equação 2.2.

$$PSNR = 10log_{10}\left(\frac{S^2}{MSE}\right). \tag{2.2}$$

Sendo que S representa o valor máximo dos elementos. Como exemplo, em imagens com 256 tons de uma determinada cor, o S teria o valor de 255, pois este é o maior valor que pode ser alcançado.

2.5.4 Distância Minkowski

A distância de Minkowski é uma generalização das distâncias Euclidiana e City Block, seu cálculo é representado pela Equação 2.3, onde m é a sua ordem e X e Y representam as imagens a serem comparadas.

$$D_{Minkowski} = \sqrt[m]{\sum_{i=1}^{p} |X_i - Y_i|^m}. \tag{2.3}$$

2.5.4.1 Distância Euclidiana

Segundo Vicini e SOUZA (2005) a distância Euclidiana é a medida de distância mais utilizada em relação à análise de agrupamentos. As imagens que possuem a menor distância Euclidiana entre si são consideradas mais parecidas em relação às com maiores distâncias. No caso mais simples, onde existem n indivíduos, cada um com valores de p variáveis, é possível descobrir a distância Euclidiana entre eles. Seu cálculo é representado na Equação 2.4, onde i e i' são os indivíduos a serem comparados e o X corresponde a matriz dos indivíduos.

$$D_{ii'} = \left[\sum_{j=1}^{p} (X_{ij} - X_{i'j})^2\right]^{\frac{1}{2}}. \tag{2.4}$$

2.5.4.2 Distância *City Block*

A distância City Block, também é chamada de distância Manhattan, e é uma norma da distância de Minkowski, pois tem como objetivo encontrar o comprimento de menor diferença entre dois pontos. A distância City Block realiza uma soma das distâncias ao longo de cada dimensão.

Possui esse nome pois geralmente é usada para calcular a distância entre pontos de uma cidade. Seus parâmetros são subtraídos uns dos outros (LOPES, 2004). A Equação

2.5 representa o seu cálculo, onde as variáveis X e Y representam as imagens a serem comparadas.

$$D_{cityblock} = \sum_{i=1}^{n} |X_i - Y_i| . \tag{2.5}$$

2.5.5 MSM (Método do Subespaço Mútuo)

O MSM tem como base o método do subespaço, onde ocorre uma redução da dimensão dos dados de entrada. Segundo Yamaguchi, Fukui e Maeda (1998), o MSM tem bons resultados quando utilizado em tarefas de reconhecimento de padrões em imagens, pois consegue armazenar as estruturas dos objetos.

A entrada do MSM não é único vetor e sim, um subespaço. Desta forma é possível realizar a análise de similaridade entre conjuntos de padrões. A Equação 2.6 apresenta sua fórmula.

$$M = \frac{1}{n} \sum_{i=1}^{k} (Xi) (Xi)^T . \tag{2.6}$$

Seu método de reduzir a dimensão funciona da seguinte forma, ao usar uma imagem Id com dimensão d, resulta em um vetor de características Xk com tamanho k. O MSM gera um subespaço para cada conjunto de imagens, onde n é a constante. Esses subespaços são gerados a partir de k vetores da matriz M, sendo que para obter a matriz M não ocorre um cálculo da média dos seus valores.

2.5.6 SSIM (Índice de Similaridade Estrutural)

O SSIM é um método que calcula a diferença de similaridade entre duas imagens. Pode ser visto como um medidor que utiliza uma das imagens como a correta. A diferença em relação a outras técnicas como PSNR e MSE é que estas abordagens estimam erros absolutos, enquanto o SSIM tem como função perceber que a degradação da imagem é uma mudança que representa a informação estrutural, incluindo tanto termo de iluminação quanto de contraste (WANG et al., 2004).

Seu cálculo é representado pela Equação 2.7.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_2^x + \mu_2^y + C_1)(\sigma_2^x\sigma_2^y + C_2)}. \tag{2.7}$$

Sendo que σ_{xy} a covariância entre as imagens, $C1$ e $C2$ são usadas para normalizar o cálculo da similaridade, e μ_i é a média e σ_i representa a variância da imagem i.

O próximo capítulo apresenta os trabalhos existentes na área de análise de sentimentos em vídeo, relatando as técnicas utilizadas para o reconhecimento das fases temporais.

3 Trabalhos Relacionados

Este capítulo apresenta algumas das pesquisas existentes na área abordada neste trabalho, descrevendo suas técnicas e métodos. A seção 3.1 descreve os trabalhos que realizaram a análise das emoções com o tratamento das fases temporais. A seção 3.2 discute e compara os trabalhos apresentados neste capítulo.

Vários trabalhos realizam a classificação de emoções sem detectar as fases temporais. Um desses trabalhos é de Zhang e Zhang (2015), os quais propuseram uma estrutura de fusão a nível de semi-característica, que são as informações iniciais afetivas faciais, para ter uma interpretação mais confiável dos estados emocionais. Outro exemplo é o trabalho de Adsul et al. (2010), os quais apresentaram um trabalho com abordagem bimodal (face e corpo) para o reconhecimento de duas emoções consideradas extremas (felicidade e tristeza) e a expressão neutra. As características extraídas da face foram obtidas a partir de informações de cor, intensidade, textura e contorno. Para a extração de características do corpo, também foi utilizado o detector de bordas, sendo este o algoritmo *Canny*.

3.1 Reconhecimento de Fases Temporais

Gunes e Piccardi (2009), realizaram uma detecção automática dos segmentos temporais do rosto. Eles processaram 10 vídeos do banco de dados Fabo e extraíram um total de 152 características para face, sendo que duas técnicas foram usadas:

- Detecção Baseada em sequência - para detecção baseada em sequência de rosto foi utilizado o HMM para modelar o segmento temporal.
- Detecção Baseada em *Frame* - para detecção baseada em *frame* foram investigados os seguintes algoritmos: C4.5, Random Forest, BayesNet, SVM, SVMSMO (*Sequential Minimal Optimization for Support Vector Machine*), Multilayer Perceptron, e AdaBoost.

Apesar da detecção baseada em sequência ter apresentado um bom desempenho em termos de tempo de resposta e em custo computacional, em termos de percentual de detecção, essa técnica alcançou resultados pouco promissores. Tendo uma taxa geral de detecção de 28,7% para a face e 37,2% para o corpo.

Enquanto que a detecção baseada em *frame* apresentou melhores resultados, 57,27% na detecção através da face e 80,66% do corpo. Esses resultados indicam que, em geral, os movimentos do corpo podem produzir características mais relevantes para a detecção das fases temporais do que os movimentos da face, pois os movimentos faciais são mais

sutis nas transições entre uma fase e outra. Chen et al. (2013) relatam que o tempo das expressões pode ser diferente em diversas modalidades, ou até mesmo na mesma, pois, cada expressão pode possuir tempos diferentes. A maneira encontrada pelos autores para solucionar este problema é a normalização temporal. Entretanto, esse trabalho não realizou uma segmentação automática das fases temporais, e sim, utilizou os rótulos existentes na base de dados Fabo para definir, de forma manual, o início (*onset*), o ápice (*apex*) e o fim (*offset*) do sentimento. Em seguida, realizou uma interpolação linear com intuito de igualar a quantidade de *frames* para cada exemplo, cujo valor foi definido em 30. Para cada entrada de vídeo foi calculado o MHI, em seguida feita a extração de características faciais com os pontos de referência de rastreamento e HOG (Histograma de gradientes). Para extrair as características do corpo foi feito o rastreio da mão e cabeça em cada *frame*, verificando sua posição, movimento e aparência. Em seguida à concatenação das características da face e do corpo, o classificador SVM foi utilizado para a definição do sentimento. Apesar da simplicidade dos recursos utilizados, o método de abordagem superou significativamente o estado da arte relatado por Gunes e Piccardi (2009). Com os rótulos manuais das fases temporais divididos foram utilizados apenas os *frames* da fase temporal *apex*, definindo esta como a fase mais discriminativa da emoção.

O estudo feito por Valstar e Pantic (2012) efetuou uma abordagem em que foi aplicado um algoritmo híbrido SVM-HMM para resolver o problema da detecção das fases temporais em vídeos de emoções com dados a partir de face e corpo. O algoritmo HMM tem sido muito usado na literatura pois tem se mostrado eficaz na modelagem do tempo em problemas de classificação. Segundo os autores, o HMM apresenta dificuldade em definir o melhor ou o exato momento de identificar a diferença entre determinadas fases temporais vizinhas. Por outro lado, o SVM, mesmo não sendo adequado para solucionar modelagem de tempo, possui elevado poder de discriminação entre classes, sendo esta união o algoritmo híbrido. Esse método é aplicado nas probabilidades de emissão, que são geradas para cada *frame* do vídeo de entrada, sendo ajustadas por combinações gaussianas, e para ajudar corrigir é feita uma maximização das probabilidades. Apesar de ter alcançado elevada taxa de detecção correta na identificação da fase temporal *apex* (80%), esse método foi bem menos efetivo na tarefa de distinguir as outras fases (68% *offset*, 36% *onset*, 34% *neutral*).

Em Gunes et al. (2015), para definir as fases temporais, foram extraídas como características a área de movimento, com base na técnica MHI. Este algoritmo gera uma representação de uma sequência do movimento existente no vídeo. Também foi utilizada divergência neutra, que é um método que mede o grau de diferença entre o *frame* atual com o *frame* neutro. Esta combinação fornece informações complementares para a definição das fases temporais. Na definição do MHI, é possível separar as fases *onset/offset* de *apex/neutral*, pois as fases *onset* e *offset* geram um maior movimento. Com a divergência neutra, é presumível separar a fase *neutral* da fase *apex*, pois existe uma grande diferença de representação, separando também a fase *onset* da fase *offset*. Onde

ocorreu o crescimento da divergência neutra, a fase foi considerada *onset*, enquanto quando a divergência neutra diminuiu, foi considerada como *offset*. Esse processo tem um custo computacional elevado pois utiliza combinações de várias técnicas para definir as fases temporais. Outra desvantagem é que na prática, em uma base de dados não manipulada, é difícil obter uma detecção precisa dos pontos principais da face e do corpo que julguem as fases temporais, devido às variações de iluminação e oclusões. Essa pesquisa também relata que as correlações das fases temporais entre diferentes modos, como exemplo face e corpo, são áreas praticamente inexploradas nas pesquisas atuais. Nesse trabalho foram selecionados 288 vídeos da base, sendo a precisão média de acerto obtida foi 83,1%.

Embora ainda não exista um método padrão para a detecção de fases temporais na área de reconhecimento de emoções, a comparação entre métodos existentes é possível, mesmo que sejam utilizadas bases de dados diferentes. Além disso, mesmo se for utilizada a mesma base de dados, não é garantido que os sistemas serão treinados e testados com o mesmo número de vídeos e com as mesmas regras. Com base nesses pontos, a tabela exibida na próxima seção destaca os principais artigos discutidos neste capítulo, relatando suas técnicas e modo com que definiram as fases temporais, porém, sem fazer comparações em termos de desempenho.

3.2 Discussão

Ao avaliar os trabalhos apresentados neste capítulo, é possível notar que para resolver a modelagem das fases temporais de modo automático, normalmente são combinadas técnicas diferentes, as quais são aplicadas em etapas distintas. As abordagens automáticas utilizam vários passos para solucionar o problema, tendo fases distintas e independentes. Portanto, esses métodos podem acarretar um elevado custo computacional.

Outra solução comum, é a segmentação das fases temporais de modo manual, onde um determinado número de pessoas é escolhido para indicar o *frame* em que cada fase temporal começa e termina, utilizando como parte final o voto majoritário. Nessa coleta, as pessoas devem analisar os vídeos e criar rótulos declarando o *frame* ou o segundo em que acontece cada fase temporal. De modo geral, os trabalhos que utilizam esta técnica manual aproveitam os rótulos para utilizar apenas a fase temporal *apex*, devido conter o ápice da emoção.

A Tabela 2 apresenta os autores e as técnicas utilizadas na detecção das fases temporais nos trabalhos descritos acima.

Tabela 2 – Comparação de soluções para definição das Fases Temporais.

Artigo	**Definição de Fases temporais**	**Solução**
Gunnes e Piccardi 2009	Manual	Base em Sequência e *Frame*
Valstar e Pantic 2012	Automática	SVM-HMM
Adsul et al. 2013	Manual	Definiu apenas neutro
Chen et al. 2013	Automática	MHI e HOG
Gunnes et al. 2015	Automática	MHI e Divergência Neutra
Zhang 2015	Manual	Rótulos Manuais

A modelagem das fases temporais se mostra muito importante como forma de auxiliar o reconhecimento de emoções em vídeo. Este trabalho busca investigar se é possível definir as fases temporais, independentemente da base de dados, de um modo mais geral e automático, com baixo custo computacional, a partir de métodos baseados em aparência e em similaridade. A arquitetura proposta nesta pesquisa é detalhada no próximo capítulo.

4 Método Proposto

Este capítulo apresenta a estrutura e todo o funcionamento do método proposto neste trabalho. A Seção 4.1 demonstra o diagrama que representa as etapas do método proposto, em seguida, as próximas seções descrevem o andamento de cada módulo que o compõe.

4.1 Arquitetura do Método

Como visto na literatura, o trabalho de identificar as fases temporais é realizado através de várias etapas. O método proposto neste trabalho possui 3 etapas. Após a entrada dos dados, que são vídeos de pessoas representando emoções, é realizado o rastreio e a identificação de N pontos faciais, com o intuito de ter apenas as partes da face que caracterizam a emoção. Com isso, são gerados N janelas de corte de cada frame do vídeo. Em seguida ocorre a etapa de comparação e similaridade entre os *frames*, onde são comparados, com os algoritmos de aparência e similaridade, os N cortes gerados pela etapa anterior, relacionando sempre na ordem do *frame* atual com o próximo. Esta fase produz como resultado um valor de diferença entre os *frames*. Por fim, o limiar adaptativo é então utilizado para poder definir um valor que identifique uma mudança brusca, sendo esta mudança o ponto de transição entre o fim de uma fase temporal e o início de outra fase temporal. O resultado final são os *frames* de início e fim de cada fase temporal.

A figura 5 exibe a estrutura em etapas do método proposto. A seguir, nas próximas seções, cada passo está descrito de forma detalhada.

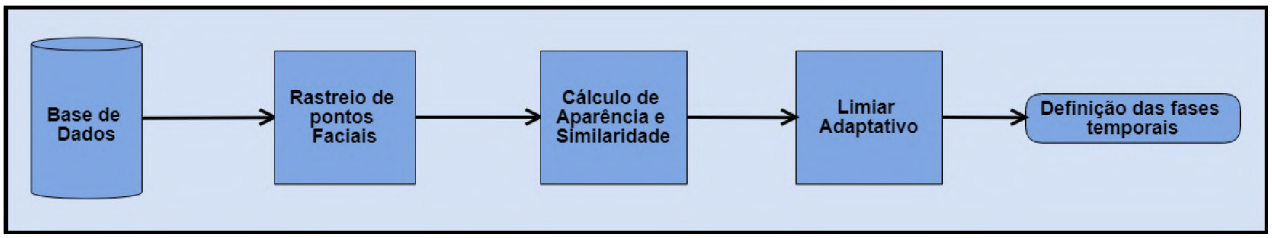

Figura 5 – Arquitetura do método dividida em 3 etapas: Rastreio de pontos Faciais, Cálculo de Aparência e similaridade, Limiar Adaptativo. Após todas essas etapas, as fases temporais são definidas. Fonte: O autor (2018).

4.2 Rastreio de pontos faciais

Nesta etapa os pontos faciais devem ser detectados por meio estratégias de rastreio de pontos faciais. Essa etapa é necessária para extrair apenas os pontos importantes da face,

com o objetivo de utilizar apenas os movimentos considerados úteis para a identificação das fases temporais do sentimento. O método proposto inicia com essa etapa porque a literatura mostra que a extração prévia de características obtém confiança e melhores resultados (GUNES; PICCARDI, 2009), (ADSUL et al., 2010) e (GUNES et al., 2015).

Após dividir cada vídeo em *frames*, para a etapa de rastreio de pontos faciais, escolhemos usar a plataforma *face plus plus*. Esse sistema funciona de forma *on-line* e identifica até 106 pontos da face de forma precisa, ou seja, o parâmetro N, que define a quantidade de pontos faciais a detectar, pode assumir até o valor 106, caso o *face plus plus* seja utilizado. Esse valor pode variar de acordo com o algoritmo de detecção de pontos faciais empregado e também deve ser ajustado para o problema investigado. Em seguida, é realizado um corte de janela de distância de Y *pixels* ao redor de cada ponto chave identificado. Essa janela de corte de distância é definida com o objetivo de identificar as pequenas mudanças existentes em cada ponto facial. O valor Y é, portanto, mais um parâmetro a ser ajustado para a execução adequada do método proposto. A figura 6 mostra um exemplo do uso da plataforma *face plus plus*.

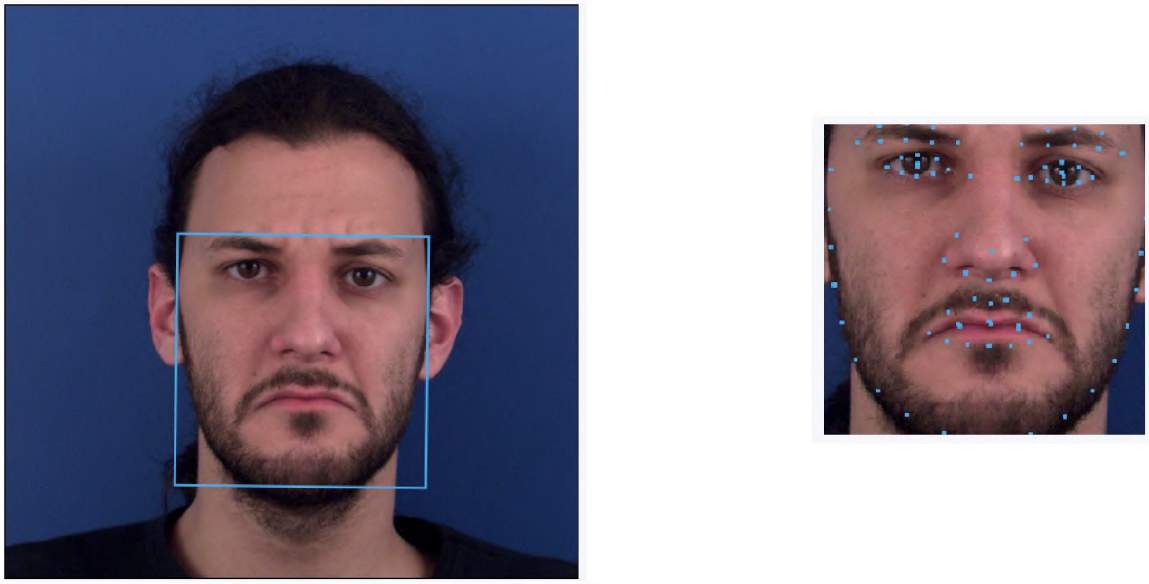

Figura 6 – Exemplo do uso do *Face plus plus* em um frame de um vídeo da base de dados MUG. O rosto é detectado na imagem da esquerda e enquanto o resultado do rastreio dos N pontos faciais é exibido na imagem da direita. Fonte: O autor (2018).

4.3 Cálculo de Aparência e Similaridade

Para comparar as janelas dos N pontos da face obtidos no módulo anterior, neste módulo nós utilizamos algoritmos baseados em aparência e em similaridade: MSE, LBP, PSNR, Distância Euclidiana, Distância City Block, Distância de Minkowski, MSM e SSIM.

Os algoritmos baseados em aparência e em similaridade foram escolhidos pois possuem como principal função encontrar uma diferença entre dois *frames* e realizam este trabalho sem precisar de conhecimento prévio ou treinamento. Nesta etapa, temos N imagens para cada *frame* do vídeo, geradas pela janela de corte. A Figura 7 apresenta um exemplo dos pontos encontrados na face, e a Figura 8 detalha, com zoom de aumento de imagem, o tamanho da janela de corte.

Figura 7 – Exemplo do uso do *Face plus plus* em um frame de um vídeo da base de dados MUG. Esse exemplo tem como valor de N igual a 25. Onde cada ponto marcado representa uma imagem. Fonte: O autor (2018).

Figura 8 – Exemplo de uma janela de corte feita pelo ponto marcado encontrado pelo *Face plus plus* com zoom de 500%, e o valor de Y igual a 20. Sendo este ponto o canto direito da boca. Fonte: O autor (2018).

Os algoritmos de aparência e em similaridade realizam o seu trabalho em dois *frames* que estão em sequencia no vídeo, sendo que cada *frame* possui suas N imagens geradas pelo corte de janela. Este módulo compara a janela de corte de um *frame* com a janela de corte do próximo *frame*. É importante mencionar que o ponto facial x do frame j é o mesmo ponto x no frame $j + 1$. Essa relação é garantida pelo sistema *face plus plus*, o qual cria rótulos com o mesmo nome para a identificação do mesmo ponto facial, mesmo que os pontos estejam em posições diferentes. Após gerar o valor de diferença de cada comparação, é realizada a soma das características de todos os $N - 1$ resultados. Em seguida, esse processo é repetido para todos os *frames* contidos no vídeo. Após a obtenção do resultado de cada comparação, é feita uma normalização dos resultados, onde cada resultado é dividido pelo maior dos resultados e multiplicado por 100. Assim temos um tratamento para diminuir a distância de diferença de valores entre os resultados. Normalização é apenas uma medida de distância relativa ao maior valor. A Figura abaixo apresenta um exemplo desta normalização.

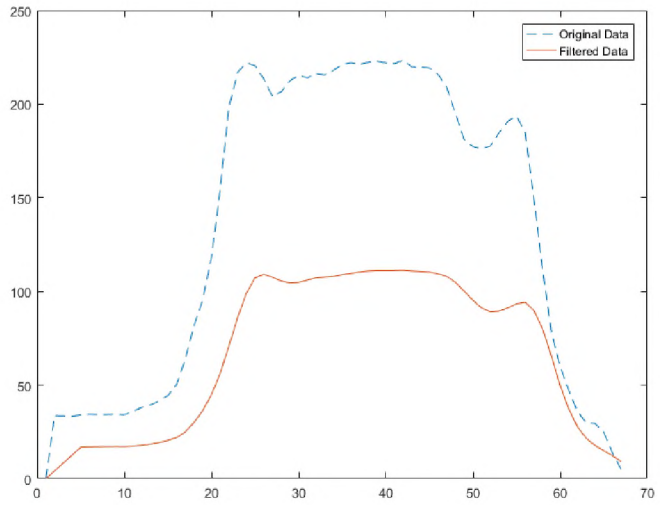

Figura 9 – O gráfico apresenta os valores gerados pelo algoritmo MSE, o *original data* são os valores antes da normalização e o *filtered data* após a normalização, os números do eixo X representam os números dos *frames*, e os números do eixo Y significam os valores gerados pelo uso do algoritmo. Fonte: O autor (2018).

Em seguida, é feito o cálculo da variância da seguinte forma: a cada entrada de um resultado é realizado o cálculo da variância e anotado, é produzido como resultado final um vetor com o número de variâncias determinado pelo número de resultados. Isto é, a cada inserção de um resultado é realizado um novo cálculo da variância. A Figura a seguir apresenta em três passos como funciona essa inserção de cada resultado para a construção do vetor de variâncias.

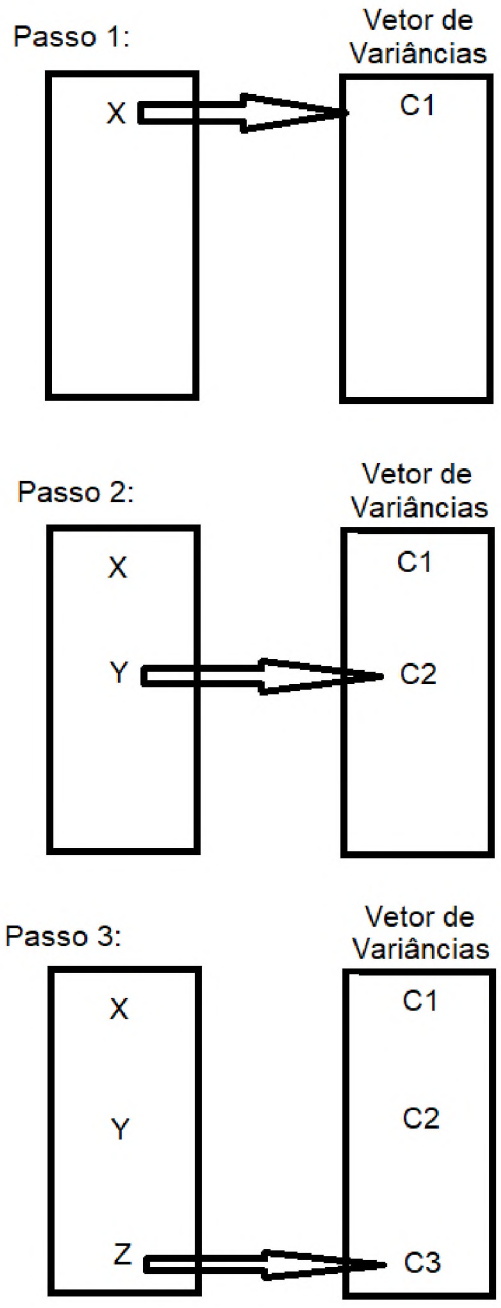

Figura 10 – A cada entrada de um resultado é feito o cálculo da variância e anotado no vetor de variâncias. Onde X,Y e Z são os resultados. E C1, C2 e C3 as variâncias. Fonte: O autor (2018).

4.4 Limiar Adaptativo

Nesta etapa foi utilizada a heurística de limiar adaptativo para avaliar os valores. De acordo com Gama et al. (2004) e Baena-García et al. (2006), esta heurística se baseia na distribuição de dados ocorridos na entrada, determinando um valor de advertência para ocorrer uma detecção. Essa estratégia foi escolhida para ser utilizada neste módulo do método proposto por tratar-se de uma estratégia não supervisionada, ou seja, não precisa de um treino para poder julgar os resultados. Dessa forma, o limiar adaptativo pode ser utilizado diretamente em situações reais, como no problema de detecção de fases temporais da emoção investigado neste trabalho. De acordo com Gregoratto et al. (2016), heurísticas de limiar adaptativo possuem uma vantagem se comparadas a técnicas de aprendizagem não supervisionada, a qual é a seguinte: esse processo não necessita de ajuste de diferentes parâmetros para cada *frame* de um vídeo. Além disso, Liu et al. (2013) reconhece que seu funcionamento é parecido com a de um cérebro humano quando trabalhado com memória curta, sendo utilizado por vários trabalhos e apresentado bons resultados.

A heurística existente no limiar adaptativo considera a ocorrência de um *buffer* que recebe os valores do vetor de variâncias. Este *buffer* recebe os valores um por um e a cada vez que recebe um novo valor é acumulada a média e o desvio. O limiar adaptativo é calculado de acordo com a fórmula:

$$L = M + (K * D)\,. \tag{4.1}$$

Onde L é o limiar, M a média gerada a partir dos valores recebidos, K é a constante que julga o nível do peso do desvio para a fase temporal e D o desvio.

O melhor valor da constante K que se encaixa para o método foi encontrado depois de vários testes experimentais. Sendo considerado mais um parâmetro do método proposto utilizado para julgar e amenizar o peso do desvio em relação aos resultados da variância.

Portanto, o valor da variável L é acumulado a cada passagem pelo vetor de variâncias, sendo este considerado o *buffer* do limiar adaptativo. Quando o limiar (L) for menor ao próximo valor do vetor de variâncias, em relação à posição atual no vetor, é declarado que neste ponto existe uma diferença maior do que o esperado e, então, é registrado um ponto de mudança de fase temporal. Em seguida o valor do limiar é zerado, assim como os valores da média e do desvio. O vetor de variâncias também é zerado e refeito a partir do ponto encontrado. Esse processo ocorre até que termine o vetor de variâncias.

Em seguida, os pontos identificados pelo método são verificados. Como a sequência das fases temporais é a mesma para todas as emoções (*neutral, onset, apex, offset e neutral*) , cada ponto de mudança indicado pelo limiar adaptativo é rotulado com o ponto final de uma fase temporal e o início da próxima fase temporal.

Portanto, o resultado final obtido são as quatro fases temporais da emoção. Desta

forma, é possível definir as fases temporais de forma independente da base de dados, sem o uso de um aprendizado profundo e com menos custo computacional. A Figura 11 apresenta um exemplo, através de gráfico, do resultado gerado pelo método comparando-o com o resultado obtido do *Ground Truth* da base.

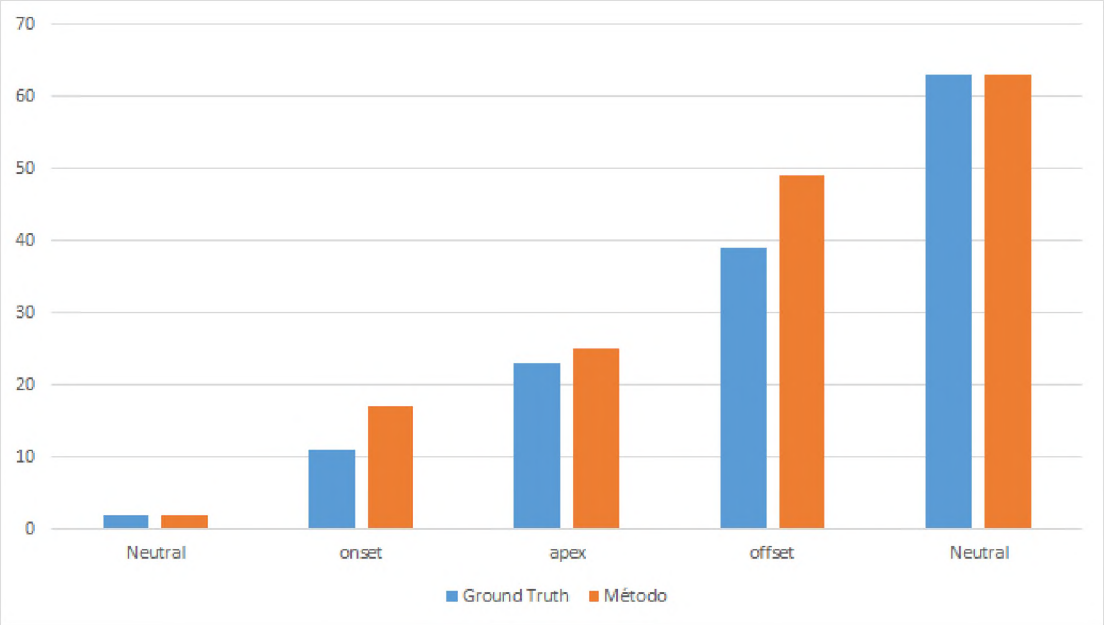

Figura 11 – Os rótulos representam o número do *frame* encontrado. Esse gráfico apresenta o resultado de detecção dos *frames* de cada fase temporal. Tendo os resultados gerados pelo método e pelo *Ground Truth*. Fonte: O autor (2018).

Como observamos no gráfico, as fases temporais que possuem maiores dificuldades em detectar são as fases com maior variação de movimento, que são elas: *onset e offset*. Em compensação, as fases com menor variação de movimento (*neutral e apex*), encontram um resultado bem próximo do existente no *ground truth*. Devemos também levar em consideração que um vídeo é composto por vários *frames*, sendo que a mudança de um *frame* para o outro ocorre de forma rápida, o que dificulta ter como resultado o *frame* exato de cada fase temporal.

Neste capítulo descrevemos o método proposto com detalhes, afim de explicar a implementação do sistema. O capítulo seguinte apresenta os experimentos realizados com o método proposto, as bases de dados investigadas são descritas, e o método proposto é comparado a um *baseline* presente na literatura.

5 Experimentos e Resultados

Neste capítulo são apresentados os experimentos realizados com o uso do método descrito neste trabalho. A seção 5.1 detalha as bases de dados utilizadas. Na seção 5.2 são discutidos os resultados obtidos através dos experimentos com o método proposto, também é definido o melhor algoritmo, dentre os algoritmos baseados em aparência e em similaridade investigados, para ser utilizado na segunda etapa do método proposto. Por fim, são apresentados resultados obtidos em experimentos que utilizam dados bimodais, mais precisamente, dados de face e de corpo.

Os experimentos realizados possuem o objetivo mostrar que o método proposto é capaz de descobrir um padrão para identificar as fases temporais nos *frames* dos vídeos e descartar os métodos baseados em aparência e em similaridade não úteis para o trabalho. Os testes foram realizados em três bases de dados públicas (Fabo, MMI e MUG), as quais possuem os rótulos das fases temporais. Dessa forma, é possível identificar e comparar os momentos em que ocorrem as mudanças de fases temporais indicados pelo método proposto e os momentos reais registrados nos rótulos dos dados.

5.1 Bases de dados

5.1.1 Fabo (*Bi-modal Face and Body Gesture Database*)

Essa base de dados é uma base Bimodal (face e corpo) encenada, extensamente utilizada na área da pesquisa que trata de análise de emoções em vídeos. Os vídeos foram capturados a partir de duas câmeras, uma direcionada para a face e outra para o corpo, desta forma se tem a mesma emoção gravada em dois modos. A base é composta por vídeos de 23 indivíduos, com idades entre 18 a 50 anos, sendo 12 do sexo feminino e 11 masculino. Os vídeos representam 10 tipos de emoções: neutro, incerteza, raiva, surpresa, medo, ansiedade, felicidade, desgosto, tédio e tristeza. A base tem 510 vídeos com rótulos das fases temporais.

Cada vídeo da base tem como início a fase temporal *neutral*, onde não ocorre nenhuma representação, em seguida, acontece o movimento da emoção e para finalizar, retorna-se ao *neutral*. Durante as gravações, foram utilizadas estratégias para estimular as emoções nos indivíduos, como por exemplo, qual seria a reação caso o indivíduo ganhasse na loteria. A base possui aproximadamente 9 GB de tamanho. A Figura 12 mostra dois exemplos de instância da base de dados FABO, a primeira coluna mostra a imagem obtida pela câmera do corpo e a segunda coluna mostra a imagem obtida a partir da câmera da face.

Figura 12 – Exemplos de sequência dos vídeos da Base de dados FABO. Fonte: Adaptado de Gunes et al. (2015).

Os rótulos contidos na base servem como *Ground truth*, representando o *frame* no qual cada fase temporal ocorre e qual emoção é expressada. O rótulo foi feito por seis pessoas e escolhido por meio de voto majoritário.

5.1.2 MMI (*The MMI Facial Expression Database*)

A sigla MMI vem da iniciativa MeM, onde os M são as iniciais dos dois principais autores (M. Pantic e M. Valstar). Embora outros autores tenham se juntado aos esforços de desenvolvimento desta base, a sigla permaneceu em uso. O MMI *Face Database* contém 1280 vídeos e mais de 2900 amostras em imagens estáticas. É uma base de dados Monomodal, ou seja, possui somente cenas de faces, com posições de duas câmeras, uma localizada para registrar a parte frontal do rosto e a outra captura o rosto de perfil. Os vídeos representam 19 indivíduos de ambos os sexo (44% do sexo feminino e 56% do sexo masculino), com idade entre 19-62 anos e de etnias variadas como europeia, asiática e sul-americana. Os vídeos representam 6 tipos de emoções: raiva, medo, desgosto, surpresa, alegria e tristeza. Todas as sequências de vídeo foram gravadas com uma taxa de 24 *frames* por segundo, sendo assim, cada amostra possui uma duração média de 520 *frames*.

Dois especialistas foram solicitados para descrever os segmentos das fases temporais, onde cada amostra tem como início a fase *neutral*, em seguida o movimento da emoção e para finalizar, retorna à fase *neutral*. A base possui aproximadamente 7 GB de tamanho. A Figura 13 apresenta o *frame Apex* de algumas amostras contidas na base de dados MMI, mostrando também as posições de câmeras contidas na base (frontal e perfil).

Figura 13 – Exemplo de amostras da base de dados MMI, mostrando um *frame Apex* e posições das câmeras. Fonte: Pantic et al. (2005).

5.1.3 MUG (*Multimedia Understanding Group*)

A base de dados MUG foi criada para superar algumas limitações existentes em outras bases, como alta resolução, iluminação uniforme, dentre outras. Possui como objetivo ajudar a pesquisa no campo do reconhecimento de emoções. O banco de dados possui mais de 1462 vídeos de face, sendo representados por 86 indivíduos (35 mulheres e 51 homens), com idade entre 20 e 35 anos. Todas as capturas foram gravadas com uma taxa de 19 *frames* por segundo, e gerando uma resolução de 896x896 *pixels* (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010).

Existem duas partes no banco de dados. Na primeira parte os indivíduos apresentam, de forma encenada, seis expressões básicas dos sentimentos (raiva, desgosto, medo, felicidade, tristeza e surpresa). A segunda parte contém emoções induzidas em laboratório. A base possui aproximadamente 38 GB de tamanho. A imagem 14 apresenta dois *frames* como exemplo de alguns sentimentos existentes na base de dados.

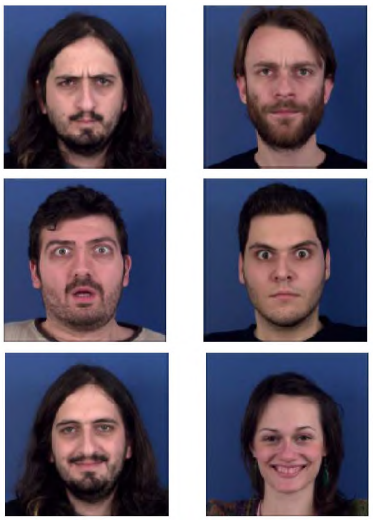

Figura 14 – Exemplo de amostras da base de dados MUG com dois *frames* de cada emoção existente, sendo eles na sequência de cima para baixo, desgosto, surpresa e feliz). Fonte: Aifanti, Papachristou e Delopoulos (2010).

5.2 Resultados Obtidos

Antes de realizar o uso do método proposto neste trabalho para todas as bases de dados, foi necessário efetuar experimentos para definir o melhor algoritmo de aparência ou de similaridade a ser utilizado na etapa de comparação entre *frames* do vídeo no processo de identificação das fases temporais das emoções em vídeos de face. Nesses experimentos foram utilizados 510 vídeos de face da base de dados FABO, os quais possuem os rótulos das fases temporais. Vale relembrar que os vídeos possuem diferenças entre si, como pessoas e movimentos diferentes.

O método proposto foi executado com cada um dos oito algoritmos investigados, ou seja, oito versões diferentes do métodos foram comparadas, seguindo todos os passos descritos no capítulo 4 deste trabalho. Nessa série de experimentos o valor da variável N, responsável pelo número de pontos faciais, foi definido em 25. Como explicado no método proposto, após a etapa de Cálculo de aparência e similaridade, é feita a soma das características. Entretanto, esse processo produz valores elevados e cujas escalas são diferentes para cada algoritmo. Para um melhor resultado do trabalho e para facilitar a comparação entre os algoritmos investigados, nós realizamos uma normalização dos valores gerados. A estratégia de normalização utilizada define um máximo valor. No caso destes experimentos, o máximo valor foi definido como 100, desta forma, temos um limite máximo de até 3 dígitos para cada resultado. Então, cada resultado é dividido pelo maior valor dos resultados e multiplicado por 100. Dessa forma, a escala final de valores será entre 0 a 100. A Figura 15 apresenta os valores gerados pelo algoritmo MSE, antes e depois da normalização dos resultados.

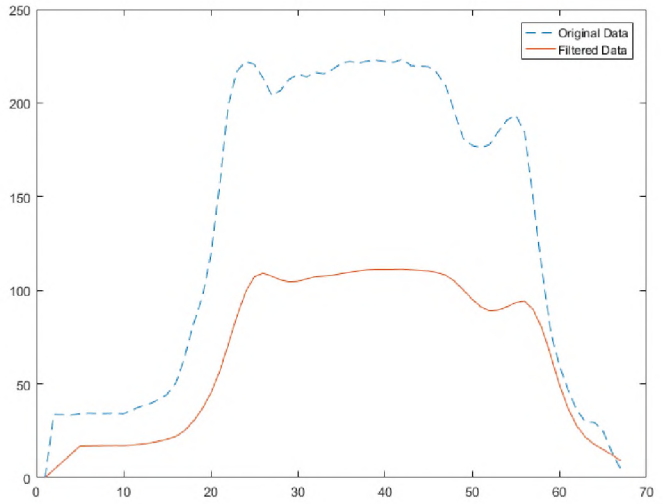

Figura 15 – O gráfico apresenta os valores gerados pelo algoritmo MSE, o *original data* representa os valores antes da normalização e o *filtered data* após a normalização, os números do eixo X representam os números dos *frames*, e os números do eixo Y significam os valores gerados pelo uso do algoritmo. Fonte: O Autor (2018).

Para medir o desempenho de cada algoritmo baseado em aparência ou em similaridade comparamos o resultado gerado pelas oito versões do método proposto com o valor dos rótulos contidos na base de dados FABO. Com o objetivo de tornar mais justa a comparação entre os oito algoritmos investigados, foi adicionada uma taxa de tolerância no valor de 10% do total de *frames* do vídeo em uso. Por exemplo, se a fase temporal *apex* termina no *frame* 30 e o vídeo possui o total de 60 *frames*, a faixa de acerto será aumentada em 6 *frames*, resultando no fato de que a fase temporal *apex* passa a estar na faixa do *frame* 24 ao 36, dita como o *frame* final. Se o rótulo desta fase temporal, contido na base, estiver dentro da faixa, considera-se a predição como acerto, caso contrario é considerado como erro. Essa heurística foi utilizada neste trabalho porque o rótulo da base de dados FABO foi feito por pessoas através de votações, com isso, essa heurística assume a existência de uma possível variância causada pela diferença entre os votos. A Figura a seguir apresenta o tamanho da faixa de acerto de acordo com a heurística aqui explicada.

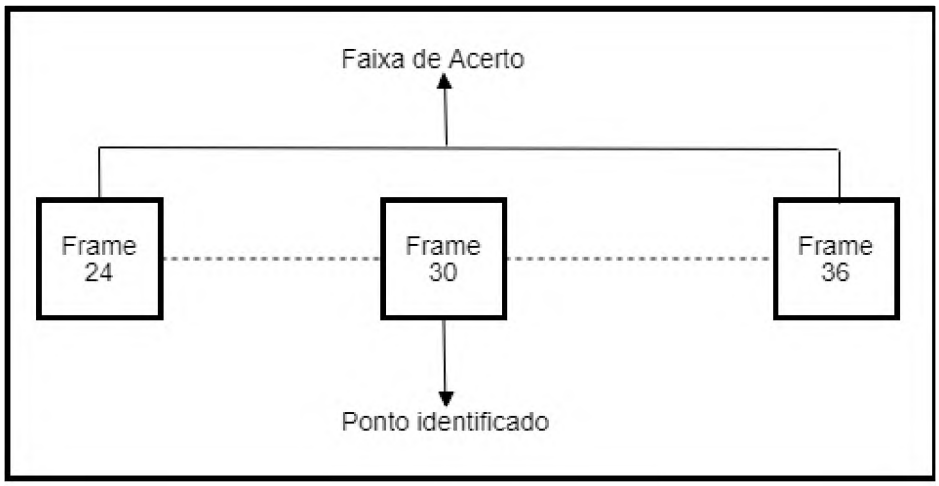

Figura 16 – Exemplo da faixa de acerto, considerada de acordo com a heurística. Fonte: O Autor (2018).

Após o uso do método proposto temos a seguir a Figura que apresenta o gráfico que relata as porcentagens de acerto em relação aos rótulos existentes na base de dados FABO obtidas pelos algoritmos baseados em aparência e em similaridade (MSE, LBP, PSNR, Distância Euclidiana, Distância City Block, Distância de Minkowski, MSM e SSIM).

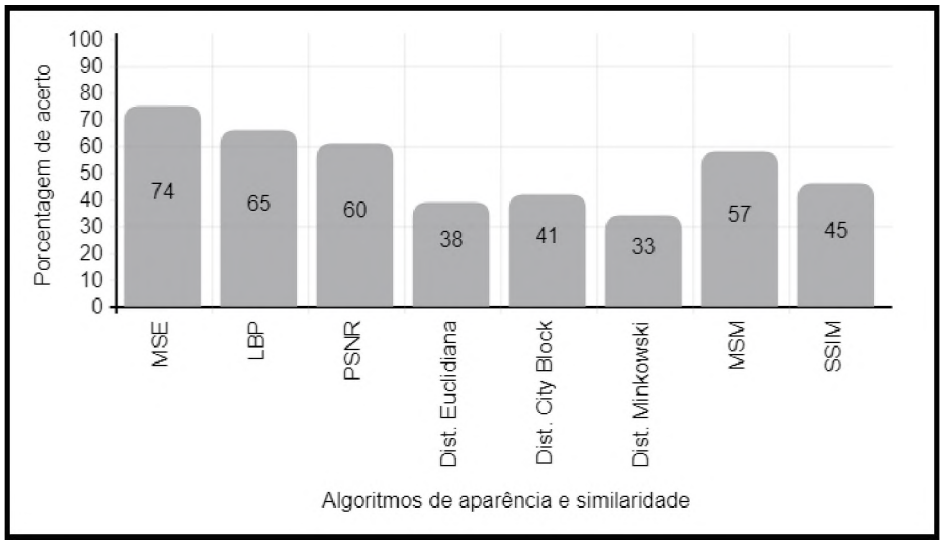

Figura 17 – Gráfico com a porcentagem de acerto obtida por cada algoritmo baseado em aparência e similaridade utilizado neste experimento. Fonte: O Autor (2018).

De acordo com este experimento foi possível identificar que a versão do método proposto que utiliza a métrica MSE obteve o maior percentual de taxa de acerto. Os algoritmos baseados em similaridade, como distância (Euclidiana, City Block, Minkowski) apresentaram os piores resultados. Nós devemos levar em consideração que esses algoritmos calculam apenas a distância de um determinado ponto de uma imagem a outra, não utilizando uma característica mais a fundo, como: textura, níveis de cinza, níveis de

energia, redução do subespaço, informação estrutural, iluminação e contraste. A litaratura mostra que o algoritmo SSIM pode produzir ótimos resultados em problemas de micro expressões, porém, essa característica o torna muito sensível a pequenos movimentos. Como os vídeos aqui estudados não são de micro expressões, o SSIM não apresentou bons resultados.

Após este experimento, definimos que o MSE é o melhor algoritmo baseado em aparência e em similaridade para o uso do método proposto. De acordo com isso, o MSE foi usado na versão do método proposto utilizado na próxima série de experimentos, na qual são investigadas as 3 bases de dados relacionadas neste capítulo (FABO, MMI e MUG). Seguindo todos os passos do método proposto explicados no capítulo 4 deste trabalho, foi definido o valor de 25. Este valor foi definido pois a aplicação *face plus plus*, em sua versão gratuita, tem como opção 25 ou 106 pontos faciais. Ao realizar experimentos, com essas duas opções, não obtivemos uma grande diferença nos resultados, apenas um maior custo computacional.

Nessa segunda série de experimentos, para medir a eficácia do método proposto foi feita uma comparação dos resultados obtidos com os rótulos contidos nas bases de dados. Foi utilizada a mesma heurística de faixa de acerto, descrita na primeira série de experimentos, para definir a quantidade de acertos do método proposto. A Figura 18 apresenta em forma de gráfico os resultados do método proposto para as 3 bases de dados.

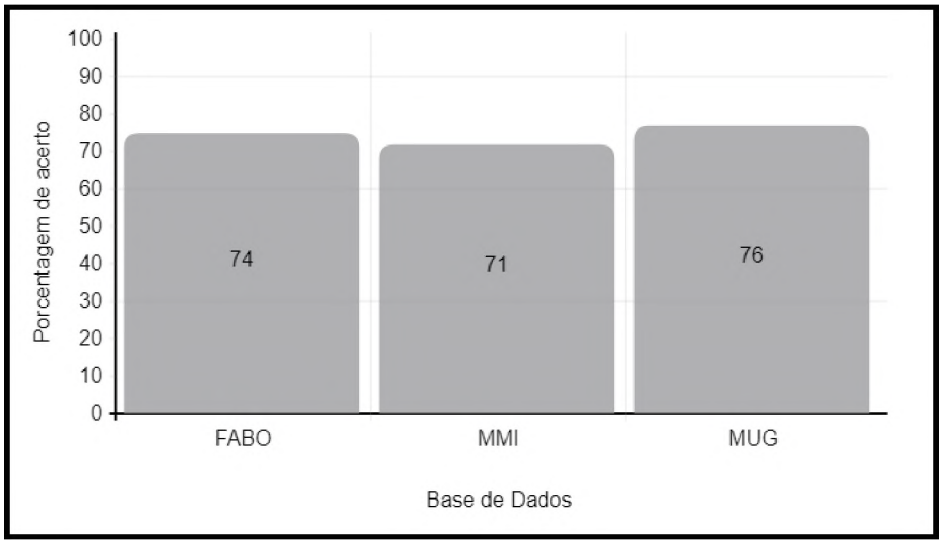

Figura 18 – Gráfico com a porcentagem de acerto do método proposto em cada base de dados. Fonte: O Autor (2018).

Podemos concluir, através dos resultados contidos na figura 18, que ao utilizar a base de dados MUG obtivemos o melhor resultado, isso se deve ao fato pois a base de dados MUG, além de possuir a melhor resolução de vídeo dentre as bases estudadas, dispõe de uma iluminação uniforme. Ocasionando em uma melhor precisão ao identificar os *pixels* dos *frames*. O pior resultado foi registrado ao utilizar a base de dados MMI, apesar da base de dados MMI ter uma taxa alta de *frames* por segundo, suas amostras possuem pouca iluminação, ocasionando sombras nos indivíduos e atrapalhando o uso do método.

O *baseline* Gunes et al. (2015), foi utilizado para julgar e comparar os resultados do método proposto. O *baseline* apresentou taxa de acerto de 83% em sua própria base de dados, mas ao ser testado em outras bases de dados não obteve a mesma eficácia. A Figura 19 apresenta, em forma de gráfico, a comparação do método com o *baseline*.

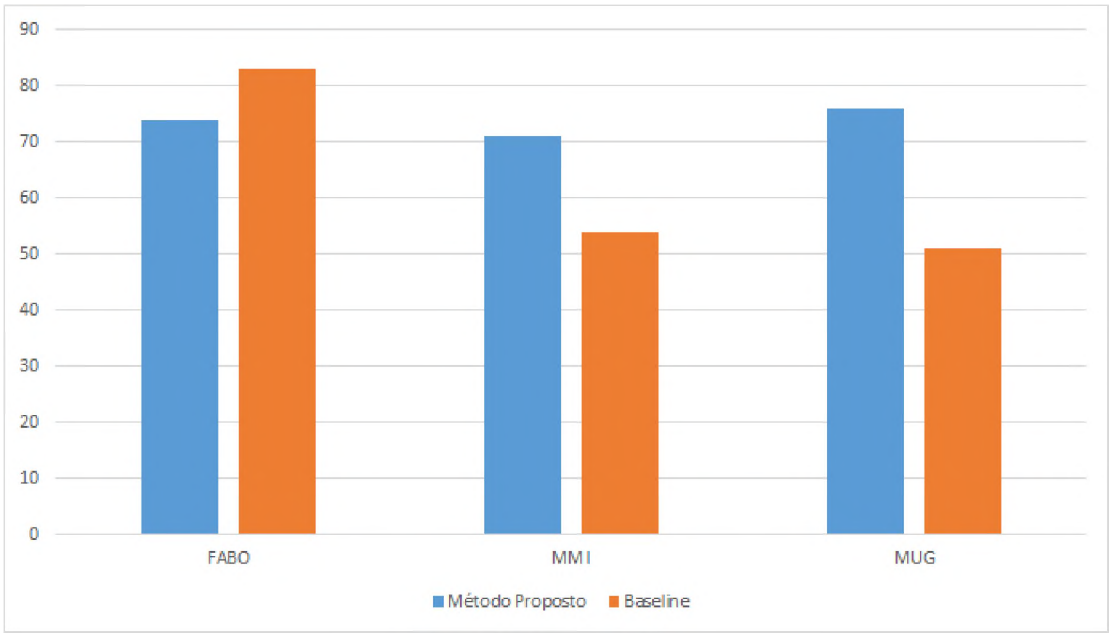

Figura 19 – Gráfico com a porcentagem de acerto do método proposto e do baseline nas bases de dados FABO, MMI e MUG. Fonte: O Autor (2018).

Como forma de comparar a contribuição do método proposto para a literatura, o *baseline* obteve o acerto de 83% ao identificar as fases temporais na base de dados FABO, e ao aplicarmos o mesmo método na base de dados MMI e MUG tivemos a eficácia abaixo dos 55%. Apesar do método proposto neste trabalho ter como acerto 74% na base de dados FABO, encontramos o padrão dentre as 3 bases de dados testadas, obtendo como resultado sempre a faixa de 70%.

Os resultados obtidos nessa segunda série de experimentos mostram que o método proposto pode alcançar taxas de detecção correta significativamente elevadas. Porém, conforme foi mencionado na introdução, o problema de classificação de emoções em vídeo é normalmente realizada por meio de fusão de dados multimodais. Com isso, para verificar se o método proposto é também efetivo ao ser utilizado no contexto de dados multimodais,

experimentos adicionais foram realizados, nos quais foram utilizados dados de duas fontes: face e gesto, ambas obtidas de vídeos. A próxima seção descreve esses experimentos e apresenta os resultados obtidos.

5.2.1 Experimentos com dados de face e de gestos de corpo.

Como relatado no início deste capítulo, a base de dados FABO é composta por dois modos de vídeos de emoção , precisamente face e corpo. Esses dados são obtidos a partir de vídeos cujas fases temporais estão devidamente rotuladas. Dada essa característica da base FABO, nós aproveitamos para realizar um experimento do método proposto aplicado a características obtidas das duas fontes de dados. A única mudança realizada no método proposto foi a utilização do extrator de característica MHI (*Motion History Image*) para obter informações do corpo.

O MHI representa todo o movimento existente em um vídeo, sendo uma representação compacta do movimento temporal. A técnica MHI gera uma imagem final, essa imagem é construída no decorrer que os *frames* do vídeo estão sendo comparados (DAVIS, 1999). A Figura 20 apresenta um exemplo da imagem final gerada no uso do MHI.

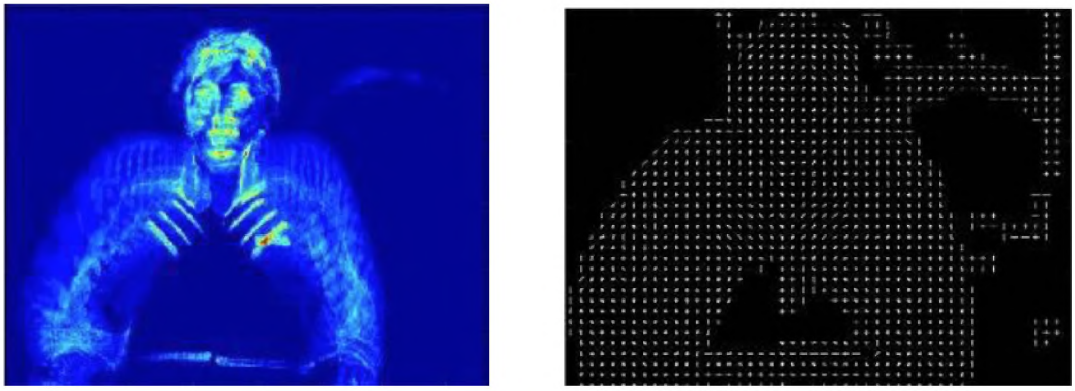

Figura 20 – Imagem final gerada no uso do MHI. Fonte: O Autor (2018).

Existem inúmeras aplicações que utilizam o MHI, como reconhecimento de movimentos humanos e rastreamento de objetos em movimento (XIANG; GONG, 2006), detecção de movimento em câmeras de vigilância (YIN; COLLINS, 2009), jogo interativo e narrativo para crianças (BOBICK et al., 1999), reconhecimento de voz virtual (YAU et al., 2006), entre outros.

O funcionamento do MHI ocorre em três etapas: na primeira é realizada a subtração de valores de intensidade das imagens, onde são subtraídos os valores entre dois *frames* sucessivos e armazenados em uma nova matriz. Isto acontece até que se tenha como resultado $n - 1$ (números de *frames*). Em seguida, ocorre a função de binarização. Esta função tem como objetivo eliminar características consideradas irrelevantes para o resultado.

Por fim, é feita a adição ponderada dos resultados das diferenças binarizadas, destacando e representando em uma imagem as diferenças dentre todos os *frames*.

Segundo Ahad et al. (2012), o MHI possui as seguintes vantagens:

- Método simples e robusto;
- Não é sensível aos ruídos;
- Expressa o fluxo de movimento temporal de cada pixel de intensidade;
- Pode ser utilizado por computadores de baixo custo.

A fórmula do cálculo do MHI é apresentada pela equação abaixo:

$$MHI\,(x,y) = \begin{cases} \tau \; se \; acontecer \; movimento(x,y), \\ 0 \; senao \; MHI(x,y) < (\tau - \delta), \end{cases}$$

onde x e y são os *frames* do vídeo, τ é o tempo atual e δ é o tempo máximo de duração da constante.

A seguir, as características do corpo, geradas pelo MHI, são adicionadas como a última característica do vetor final da face. Dessa forma, a fusão é feita em nível de características, ou seja, um único vetor de características é gerado e submetido ao limiar adaptativo. Portanto, após a fusão das características em um único vetor, o método proposto segue o fluxo normal, conforme visto na arquitetura apresentada no capítulo anterior. A Figura 21 apresenta uma comparação entre os resultados obtidos a partir do uso do método proposto na base de dados FABO com um modal (face) e dois modais (face e corpo).

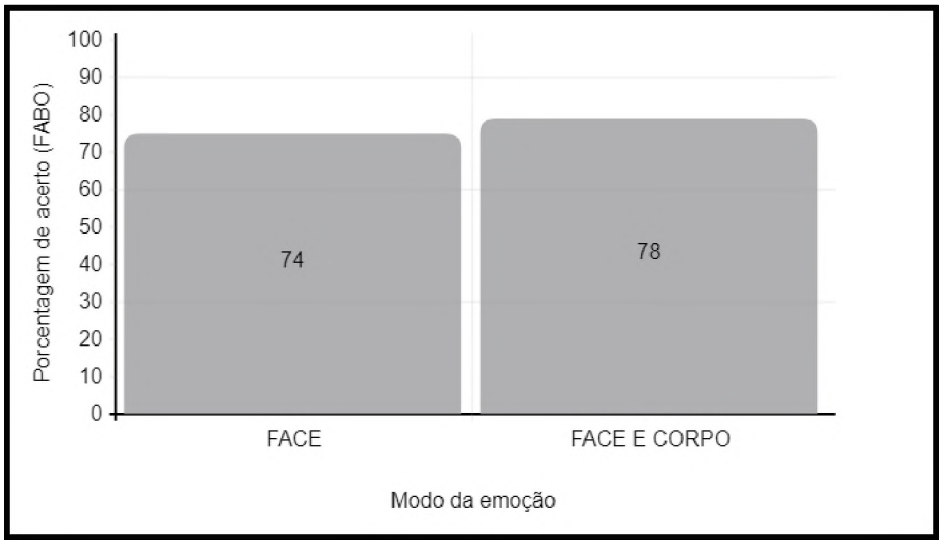

Figura 21 – Gráfico de porcentagem de acerto do uso do método proposto na base de dados FABO com um modal (face) e dois modais (face e corpo). Fonte: O Autor (2018).

Como visto no resultado da figura 21 ocorreu um aumento no acerto ao utilizarmos duas fontes de informações para a detecção das fases temporais. Esses resultados indicam que o método proposto é também eficiente para detectar as fases temporais caso a fonte de dados seja proveniente somente de gestos. Por fim, a fusão multimodal ajuda a melhorar o desempenho do método, dado que a taxa de acerto aumentou ao se trabalhar com as duas fontes de dados. Porém, infelizmente não existem muitas bases de dados públicas bimodais, com isso, não foi possível replicar esse estudo com outras bases para comprovar essas hipóteses levantadas a partir dos resultados obtidos na base FABO.

6 Conclusão e Trabalhos Futuros

Este trabalho teve como objetivo desenvolver um método para identificar de forma automática as fases temporais de emoções expressas em vídeos. O método utiliza algoritmos baseados em aparência e em similaridade com características extraídas de expressões faciais. Embora, na literatura, existam trabalhos que investigam essa mesma problemática, a maioria das soluções existentes apresenta custo computacional elevado e uso de aprendizado profundo.

O método proposto foi investigado em três diferentes bases de dados e comparado ao baseline de (GUNES et al., 2015). Antes, porém, oito métodos baseados em aparência e em similaridade foram comparados para que fosse definido o algoritmo mais adequado para uso na etapa de cálculo de aparência e de similaridade do método proposto. Os testes realizados indicaram o MSE como algoritmo que alcançou melhor desempenho ao compor o método proposto.

Em termos de comparação com o *baseline*, o método de (GUNES et al., 2015) obteve 83% de precisão ao identificar as fases temporais na base de dados FABO, enquanto método proposto alcançou 74%. Porém, nas demais bases investigada, isto é, MMI e MUG, o desempenho dos dois métodos apresentou comportamento diferente. O *baseline* alcançou precisão abaixo dos 55% em ambas bases, enquanto o método proposto apresentou taxas de precisão acima de 70%. Portanto, encontramos um padrão de desempenho nas três bases de dados testadas, pois a precisão obtida está na faixa de 70%.

Dado que o *baseline* aqui testado apresentou taxa de acerto de 83% na base de dados para a qual foi proposto, mas, ao ser testado em outras bases de dados, o mesmo desempenho não foi verificado, nós podemos concluir que o método proposto é capaz de identificar de forma automática as fases temporais dos seis sentimentos básicos, representados nas bases de dados investigadas, independentemente da base de dados utilizada. Além disso, esse processo de detecção automática não utiliza aprendizado profundo e apresenta menor custo computacional.

Por outro lado, este trabalho apresenta algumas limitações, como os seguintes: os vídeos das bases de dados são encenados, contêm apenas uma pessoa em cada vídeo, a câmera está posicionada de frente para o indivíduo e é possível ver de forma clara a face completa da pessoa.

É importante também destacar que nós realizamos experimentos adicionais para testar o desempenho do método proposto quando mais de um modo de informação de emoção é utilizado. Nesse caso, foram usadas características extraídas de expressões faciais e de gestos do corpo. Os resultados mostram que a fusão dessas fontes de informação pode

melhorar a precisão da tarefa de reconhecimento das fases temporais de emoções humanas, pois, ao utilizarmos os dados de face e de gestos de corpo da base de dados FABO, os resultados foram superiores às taxas obtidas com o uso somente da expressão facial. Porém, infelizmente não foi possível realizar uma análise mais completa devido à quase ausência de bases bimodais com rótulos disponíveis publicamente. Para um melhor aprofundamento nesta questão, outras bases devem ser investigadas. Essa demanda pode ser considerada uma indicação para trabalhos futuros. Outra forma de aprofundar esse estudo seria obter as porcentagens de acerto obtidas ao utilizarmos no processo de classificação apenas a sequência de cada fase temporal individualmente. Dessa forma, nós poderemos identificar a fase temporal mais relevante para a classificação das emoções. Por fim, essa análise individual das fases temporais pode ser feita considerando cada emoção separadamente, pois, o impacto das fases pode variar de emoção para emoção.

Referências

ADSUL, S.; HULE, S.; ANTONY, J.; NAIK, A.; TALELE, K. Emotion detection and recognition using facial expressions and body gestures. Sardar Patel Institute of Technology, Andheri (West), Mumbai-58, 2010. Citado 3 vezes nas páginas 22, 35 e 40.

AHAD, M. A. R.; TAN, J. K.; KIM, H.; ISHIKAWA, S. Motion history image: its variants and applications. *Machine Vision and Applications*, Springer, v. 23, n. 2, p. 255–281, 2012. Citado na página 56.

AIFANTI, N.; PAPACHRISTOU, C.; DELOPOULOS, A. The mug facial expression database. In: IEEE. *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*. [S.l.], 2010. p. 1–4. Citado 2 vezes nas páginas 14 e 49.

ATKINSON, R. L. *Introdução à psicologia de Hilgard*. [S.l.]: Artmed, 2002. Citado na página 21.

BAENA-GARCÍA, M.; CAMPO-ÁVILA, J. del; FIDALGO, R.; BIFET, A.; GAVALDÀ, R.; MORALES-BUENO, R. Early drift detection method. *4th International Workshop on Knowledge Discovery from Data Streams (ECML/PKDD)*, Berlin, Germany: Springer, 2006. Citado na página 44.

BOBICK, A. F.; INTILLE, S. S.; DAVIS, J. W.; BAIRD, F.; PINHANEZ, C. S.; CAMPBELL, L. W.; IVANOV, Y. A.; SCHÜTTE, A.; WILSON, A. The kidsroom: A perceptually-based interactive and immersive story environment. *Presence: Teleoperators and Virtual Environments*, MIT Press, v. 8, n. 4, p. 369–393, 1999. Citado na página 55.

CARIDAKIS, G.; CASTELLANO, G.; KESSOUS, L.; RAOUZAIOU, A.; MALATESTA, L.; ASTERIADIS, S.; KARPOUZIS, K. Multimodal emotion recognition from expressive faces, body gestures and speech. In: *Artificial intelligence and innovations 2007: From theory to applications*. [S.l.]: Springer, 2007. p. 375–388. Citado na página 28.

CHEN, S.; TIAN, Y.; LIU, Q.; METAXAS, D. N. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, Elsevier, v. 31, n. 2, p. 175–185, 2013. Citado 3 vezes nas páginas 22, 23 e 36.

COHEN, I.; SEBE, N.; GARG, A.; CHEN, L. S.; HUANG, T. S. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, Elsevier, v. 91, n. 1, p. 160–187, 2003. Citado na página 23.

CONNELL, S. D.; JAIN, A. K. Template-based online character recognition. *Pattern Recognition*, Elsevier, v. 34, n. 1, p. 1–14, 2001. Citado na página 25.

DAVIS, J. Recognizing movement using motion histograms. *Technial Report 487, MIT Media Lab*, Citeseer, v. 1, n. 487, p. 1, 1999. Citado na página 55.

DONATO, G.; BARTLETT, M. S.; HAGER, J. C.; EKMAN, P.; SEJNOWSKI, T. J. Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 21, n. 10, p. 974–989, 1999. Citado na página 22.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012. 21–39 p. Citado na página 26.

ESCALERA, S.; GONZÀLEZ, J.; BARÓ, X.; REYES, M.; LOPES, O.; GUYON, I.; ATHITSOS, V.; ESCALANTE, H. Multi-modal gesture recognition challenge 2013: Dataset and results. In: ACM. *Proceedings of the 15th ACM on International conference on multimodal interaction*. [S.l.], 2013. p. 445–452. Citado na página 21.

FASEL, B.; LUETTIN, J. Automatic facial expression analysis: a survey. *Pattern recognition*, Elsevier, v. 36, n. 1, p. 259–275, 2003. Citado na página 21.

GAMA, J.; MEDAS, P.; CASTILLO, G.; RODRIGUES, P. Learning with drift detection. In: SPRINGER. *Brazilian symposium on artificial intelligence*. [S.l.], 2004. p. 286–295. Citado na página 44.

GELDER, B. D. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, The Royal Society, v. 364, n. 1535, p. 3475–3484, 2009. Citado na página 22.

GREGORATTO, C. d. J. et al. Detecção de comportamento anormal em vídeos de multidão. Universidade Federal do Amazonas, 2016. Citado na página 44.

GUNES, H.; PICCARDI, M. Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, IEEE, v. 39, n. 1, p. 64–84, 2009. Citado 5 vezes nas páginas 21, 22, 35, 36 e 40.

GUNES, H.; SHAN, C.; CHEN, S.; TIAN, Y. Bodily expression for automatic affect recognition. *Emotion Recognition: A Pattern Analysis Approach*, John Wiley & Sons, Inc., p. 343–377, 2015. Citado 12 vezes nas páginas 13, 14, 22, 23, 27, 29, 30, 36, 40, 48, 54 e 59.

HUYNH-THU, Q.; GHANBARI, M. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, IET, v. 44, n. 13, p. 800–801, 2008. Citado na página 32.

JAQUES, P. A.; VICARI, R. M. Estado da arte em ambientes inteligentes de aprendizagem que consideram a afetividade do aluno. *Revista Informática na Educação*, p. 15–38, 2005. Citado na página 27.

KOELSTRA, S.; PANTIC, M. Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics. In: IEEE. *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. [S.l.], 2008. p. 1–8. Citado na página 29.

LIU, B. Sentiment analysis: A multi-faceted problem. *IEEE Intelligent Systems*, v. 25, n. 3, p. 76–80, 2010. Citado na página 21.

LIU, J.; FENG, J.; TAN, L.; MA, D. An algorithm of auto-update threshold for singularity analysis of pipeline pressure. *Mathematical Problems in Engineering*, Hindawi, v. 2013, 2013. Citado na página 44.

LOPES, E. C.; FILHO, J. C. B.; NO, R. T. Detecção de faces e características faciais. *Porto Alegre: PUCRS*, 2005. Citado 2 vezes nas páginas 30 e 31.

LOPES, M. C. S. *Mineração de dados textuais utilizando técnicas de clustering para o idioma português*. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, 2004. Citado na página 32.

MEHRABIAN, A. Communication without words. *Psychological today*, v. 2, p. 53–55, 1968. Citado na página 22.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citado na página 31.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 7, p. 971–987, 2002. Citado 2 vezes nas páginas 13 e 31.

OLIVEIRA, E. d.; JAQUES, P. A. Classificação de emoções básicas através de imagens capturadas por webcam. *Revista Brasileira de Computação Aplicada*, v. 5, n. 2, p. 40–54, 2013. Citado na página 27.

OTSUKA, T.; OHYA, J. Spotting segments displaying facial expression from image sequences using hmm. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 442–447. Citado na página 23.

PANTIC, M.; PATRAS, I. Temporal modeling of facial actions from face profile image sequences. In: IEEE. *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. [S.l.], 2004. v. 1, p. 49–52. Citado 2 vezes nas páginas 23 e 29.

PANTIC, M.; PATRAS, I. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, IEEE, v. 36, n. 2, p. 433–449, 2006. Citado na página 23.

PANTIC, M.; VALSTAR, M.; RADEMAKER, R.; MAAT, L. Web-based database for facial expression analysis. In: IEEE. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. [S.l.], 2005. p. 5–pp. Citado 2 vezes nas páginas 14 e 49.

PICARD, R. W.; PICARD, R. *Affective computing*. [S.l.]: MIT press Cambridge, 1997. v. 252. Citado na página 27.

PLUS, F. plus. Face plus plus cognitive services. 2017. Disponível em: <https://www.faceplusplus.com>. Citado na página 26.

RUSSELL, J. A.; FERNÁNDEZ-DOLS, J. M. *The psychology of facial expression*. [S.l.]: Cambridge university press, 1997. Citado na página 21.

SA, J. M. D. *Pattern recognition: concepts, methods and applications*. [S.l.]: Springer Science & Business Media, 2012. Citado 2 vezes nas páginas 15 e 26.

SCHMIDT, K. L.; COHN, J. F. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American journal of physical anthropology*, Wiley Online Library, v. 116, n. S33, p. 3–24, 2001. Citado 2 vezes nas páginas 21 e 22.

VALSTAR, M. F.; PANTIC, M. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, IEEE, v. 42, n. 1, p. 28–43, 2012. Citado na página 36.

VICINI, L.; SOUZA, A. M. Análise multivariada da teoria à prática. *Santa Maria: UFSM, CCNE*, 2005. Citado na página 32.

WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, IEEE, v. 13, n. 4, p. 600–612, 2004. Citado na página 33.

XIANG, T.; GONG, S. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, Springer, v. 67, n. 1, p. 21–51, 2006. Citado na página 55.

YAMAGUCHI, O.; FUKUI, K.; MAEDA, K.-i. Face recognition using temporal image sequence. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 318–323. Citado na página 33.

YAU, W. C.; KUMAR, D. K.; ARJUNAN, S. P.; KUMAR, S. Visual speech recognition using image moments and multiresolution wavelet images. In: IEEE. *Computer Graphics, Imaging and Visualisation, 2006 International Conference on*. [S.l.], 2006. p. 194–199. Citado na página 55.

YIN, Z.; COLLINS, R. Moving object localization in thermal imagery by forward-backward motion history images. In: *Augmented Vision Perception in Infrared*. [S.l.]: Springer, 2009. p. 271–291. Citado na página 55.

ZHANG, Y.; ZHANG, L. Semi-feature level fusion for bimodal affect regression based on facial and bodily expressions. In: INTERNATIONAL FOUNDATION FOR AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS. *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. [S.l.], 2015. p. 1557–1565. Citado 2 vezes nas páginas 22 e 35.