Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Classificação de expressões faciais, gênero e engajamento em videochamadas usando redes neurais convolucionais

Ana Beatriz Kapps dos Reis

JUIZ DE FORA
DEZEMBRO, 2023

Classificação de expressões faciais, gênero e engajamento em videochamadas usando redes neurais convolucionais

Ana Beatriz Kapps dos Reis

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Edelberto Franco Silva

Coorientador: Luiz Maurílio da Silva Maciel

JUIZ DE FORA

DEZEMBRO, 2023

Classificação de expressões faciais, gênero e engajamento em videochamadas usando redes neurais convolucionais

Ana Beatriz Kapps dos Reis

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Edelberto Franco Silva
Doutor Computação

Luiz Maurílio da Silva Maciel
Doutor em Engenharia de Sistemas e Computação

Saulo Moraes Villela
Doutor em Engenharia de Sistemas e Computação

Ronney Moreira de Castro
Doutor em Informática

JUIZ DE FORA
15 DE DEZEMBRO, 2023

Resumo

A rápida transição para o ensino remoto, impulsionada pela pandemia de COVID-19, trouxe consigo desafios significativos, notadamente a escassez de interação entre alunos e professores. Neste trabalho busca-se avaliar modelos baseados em redes neurais convolucionais para classificação de expressões faciais, gênero e nível de engajamento dos participantes de videochamadas. Ao oferecer aos professores uma visão aprofundada do estado emocional e participação dos estudantes, a proposta visa melhorar a compreensão do envolvimento e interesse dos alunos nas matérias, contribuindo para uma experiência educacional mais efetiva. Para avaliar os modelos de classificação, foram criados cenários realistas, envolvendo indivíduos reais. Devido à dificuldade de obter vídeos no cenário educacional, os vídeos foram criados no cenário de trabalho remoto. Acredita-se que nesse cenário é igualmente possível analisar o engajamento dos participantes. A aplicação prática desses cenários proporcionou *insights* valiosos. Na classificação emocional, as emoções se mostraram compatíveis com as expectativas. Quanto à classificação de gênero, o modelo obteve precisão na maioria das instâncias. Além disso, observou-se que o nível de engajamento apresentou resultados mais efetivos quando a câmera estava bem posicionada, isto é, diretamente na frente do rosto, permitindo uma captura adequada da região dos olhos.

Palavras-chave: redes neurais convolucionais, videochamadas, expressões faciais, cenários realistas, nível de engajamento, classificação emocional, classificação de gênero.

Abstract

The rapid transition to remote learning, driven by the COVID-19 pandemic, brought with it significant challenges, notably the lack of interaction between students and teachers. This work seeks to evaluate models based on convolutional neural networks for classifying facial expressions, gender and level of engagement of video call participants. By offering teachers an in-depth view of students' emotional state and participation, the proposal aims to improve understanding of students' involvement and interest in subjects, contributing to a more effective educational experience. To evaluate the classification models, realistic scenarios were created, involving real individuals. Due to the difficulty of obtaining videos in the educational setting, videos were created in the remote work setting. We believe that in this scenario it is also possible to analyze the participants' engagement. The practical application of these scenarios provided valuable insights. In the emotional classification, emotions were compatible with expectations. As for gender classification, the model was accurate in most instances. Furthermore, it was observed that the level of engagement presented more effective results when the camera was well positioned, that is, directly in front of the face, allowing adequate capture of the eye region.

Keywords: convolutional neural networks, video calls, facial expressions, realistic scenarios, engagement level, emotional classification, gender classification.

Agradecimentos

A todos os meus parentes, pelo encorajamento e apoio.

Aos professores Edelberto Franco Silva e Luiz Maurílio da Silva Maciel pela orientação e paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

"A persistência é o caminho do êxito".

Charles Chaplin

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BGR	Azul, Verde, Vermelho
CK+	Conjunto de dados estendido de Cohn-Kanade
CNNs	Redes neurais convolucionais
COVID-19	Coronavirus Disease 2019
FER	Reconhecimento de expressão facial
FERG	Banco de dados do grupo de pesquisa de expressão facial
JAFFE	Expressão facial feminina japonesa
OBS Studio	Estúdio de Software de Transmissão Aberta
OpenCV	Visão Computacional de Código Aberto
RAF-DB	Base de Dados Audiovisual Ryerson de Fala e Canção com Expressões Emocionais
ResNet-10	Rede Residual 10
ResNet-50	Rede Residual 50
RGB	Vermelho, Verde, Azul
R-CNN	Redes neurais convolucionais baseadas em região
SSD	Detector de Múltiplas Caixas em um Único Disparo
VGG13	Grupo de Geometria Visual versão 13
VGG16	Grupo de Geometria Visual versão 16
VGG19	Grupo de Geometria Visual versão 19

1 Introdução

As emoções no campo da educação são cruciais, pois elas possuem um papel essencial nos processos cognitivos responsáveis pela aprendizagem e assimilação de novas informações. Após o surto da COVID-19, o cenário de ensino-aprendizagem mudou drasticamente para plataformas digitais (MISHRA; GUPTA; SHREE, 2020). Com essa mudança repentina da sala de aula física para virtual, muitos desafios surgiram e um deles é a falta de motivação, uma vez que a ausência de interação face a face com o professor pode afetar o desempenho de aprendizagem do aluno.

O nível de engajamento é frequentemente manifestado através da análise emocional, tornando possível avaliar o envolvimento dos alunos durante as aulas. Um exemplo claro desse fenômeno é observado nas emoções positivas, como alegria, que exerce um impacto positivo, facilitando a autorregulação e promovendo um foco mais intenso nas tarefas de resolução de problemas. Isso, por sua vez, contribui para criar um ambiente propício ao engajamento dos alunos (KIURU et al., 2020). Em contrapartida, emoções negativas, como tristeza e raiva, desviam a atenção dos alunos e consomem recursos cognitivos durante as atividades de aprendizagem. O resultado desse impacto é uma desmotivação evidente entre os alunos (TORRES, 2020).

Em vista disso, faz-se necessário o uso de funcionalidades que possam auxiliar os professores durante a aula ministrada. Apesar de alguns estudos serem propostos dos últimos tempos para cá, o reconhecimento de expressões faciais ainda é um desafio (GUPTA; KUMAR; TEKCHANDANI, 2023) e há muito espaço para melhorar a robustez e desempenho dessas técnicas e aplicá-las em plataformas *online*. O presente trabalho analisa um modelo de classificação capaz de capturar as expressões faciais dos alunos para ambientes de aprendizagem *online*. Adicionalmente, o modelo também classifica o gênero e o nível de engajamento. Com isso, os tutores terão maior controle sobre envolvimento dos alunos em tempo real, trazendo informações e características importantes sobre as emoções dos alunos naquele momento.

1.1 Contextualização

À medida que a pandemia da COVID-19 se espalhou, houve um movimento crescente em direção ao ensino *online* devido ao fechamento das escolas, faculdades e universidades (MARTINEZ, 2020). Após esse período de pandemia, as plataformas *online* ganharam enorme atenção em todo o mundo (NOURAEY; BAVALI; BEHJAT, 2023) e trouxeram consigo muitos desafios na aprendizagem.

Existe uma ligação muito forte entre o estado emocional dos alunos, o desempenho do professor e os resultados da aprendizagem. De fato, o estado emocional, as decisões em tempo real tomadas pelo professor e os diversos parâmetros que cercam a aprendizagem nas interações educacionais realmente têm uma influência automática sobre a qualidade da tutoria como um todo. A necessidade de integrar o reconhecimento emocional nos sistemas de aprendizagem para obter uma educação de qualidade, tanto do ponto de vista do aluno quanto do professor (BOUHLAL et al., 2020).

Expressões faciais são movimentos musculares físicos traduzidos de impulsos emocionais, como levantar as sobrancelhas, franzir a testa ou curvar os lábios. Ao observar a mudança de expressões faciais automaticamente, muitas informações sobre os estados emocionais dos alunos podem ser determinadas (ZHENG et al., 2021). As expressões faciais são divididas em sete emoções básicas segundo o trabalho de Ekman e Oster (1979). Desde então, muitos pesquisadores aceitaram universalmente essas emoções básicas para pesquisa sendo elas: surpresa, felicidade, tristeza, medo, nojo, raiva e neutra. Portanto, analisar automaticamente as expressões faciais dos alunos ajuda a decidir o estado de envolvimento em cenários de tempo real. Ademais, a inclusão do gênero e o nível de engajamento pode fornecer *insights* adicionais para os professores.

1.2 Descrição do Problema

Considerando que as plataformas de ensino *online* carecem de ferramentas que proporcionem um suporte e supervisão eficaz aos professores durante suas aulas, torna-se evidente a necessidade de um modelo de classificação. Esse modelo pode destacar as características emocionais dos alunos além do gênero, proporcionando uma visão mais

completa do ambiente virtual de aprendizagem. Além disso, uma melhoria significativa pode ser alcançada ao incorporar o monitoramento do engajamento da turma. Permitindo que o professor observe em tempo real o nível de envolvimento dos alunos, essa adição é particularmente valiosa, uma vez que é desafiador determinar se o conteúdo está sendo compreendido de maneira clara do outro lado da tela.

1.3 Justificativa

Atualmente, as plataformas digitais possuem uma carência quando se trata de avaliar a motivação e a interação dos alunos, além de trazer a desconexão entre aluno e professor tornando a aula massiva e sem interação de ambas as partes. Em vista disso, a interação é a parte fundamental no controle do ambiente de aprendizagem, e esse assunto possui extrema relevância e preocupação por parte das instituições de ensino.

Considerando o cenário atual da transformação digital, e que após a pandemia o cenário de ensino-aprendizagem mudou drasticamente nos últimos tempos para plataformas digitais, há uma necessidade de evolução ainda maior para que o impacto dessas tecnologias seja cada vez mais eficaz, de forma que não prejudique a educação dos alunos.

Diante desse panorama, a aplicação de um modelo de classificação em tempo real pode oferecer suporte valioso aos professores, permitindo a coleta imediata de informações sobre o desempenho dos alunos durante a aula. Essa abordagem não apenas minimiza a desconexão na interação humana, mas também contribui para tornar o ensino remoto mais próximo da experiência presencial. Essa perspectiva ganha ainda mais relevância, especialmente considerando que muitas empresas adotaram o *home office* após a pandemia, estendendo a análise de engajamento para videochamadas de trabalho. Vale ressaltar que, neste trabalho, a análise foi conduzida no contexto de *home office* devido às limitações na obtenção de vídeos de aulas.

1.4 Objetivos

1.4.1 Objetivo geral

O objetivo principal deste trabalho é conduzir uma análise crítica de dois modelos classificadores propostos na literatura, com foco em aplicá-los e avaliá-los em um cenário realista. Para atingir esse propósito, optou-se por utilizar dois modelos reconhecidos como estado da arte, visando uma compreensão aprofundada de suas capacidades. Esses modelos apresentam técnicas que permitem a classificação de emoções, gênero e nível de engajamento.

A escolha desses modelos é fundamentada em sua relevância e desempenho reconhecidos na literatura. Ao aplicar esses classificadores em um contexto realista, espera-se não apenas aprimorar a experiência de ensino online, mas também proporcionar aos educadores valiosos *insights* sobre o estado emocional dos alunos durante as aulas, enriquecendo assim o processo educacional.

Além disso, os resultados apresentados não apenas validam os modelos utilizados, mas também fornecem um caminho para futuras investigações e desenvolvimentos que possam aprimorar o ensino remoto.

1.4.2 Objetivo específico

Os objetivos específicos para desenvolvimento deste trabalho são:

- Pesquisar e estudar a fundamentação teórica para embasamento deste trabalho como: reconhecimento de expressões faciais, modelos de redes neurais convolucionais (*convolutional neural networks* – CNN), conjunto de dados FER2013, detecção facial, classificação de emoções e gênero;
- Pesquisar as necessidades específicas para as plataformas online de ensino educacional;
- Revisão bibliográfica sobre os trabalhos existentes na literatura relacionados à avaliação de engajamento;
- Construção de vídeos para avaliação de um método de classificação de engajamento;

- Avaliação do modelo nos vídeos elaborados e análise dos resultados.

1.5 Organização

Esta monografia está organizada em cinco capítulos, além desta introdução. O Capítulo 2 apresenta a fundamentação teórica de conceitos importantes como: reconhecimento de expressões faciais, expressões básicas universais, modelos CNNs, conjunto de dados FER2013 e *Single Shot MultiBox Detector*. O Capítulo 3 apresenta os trabalhos relacionados. O Capítulo 4 detalha o funcionamento do classificador analisado. No Capítulo 5 são apresentados os cenários para experimentação, além dos resultados obtidos. Por último, o Capítulo 6 apresenta as conclusões.

2 Fundamentação Teórica

Neste capítulo são abordados conceitos importantes para a compreensão do trabalho desenvolvido. A Seção 2.1 descreve sobre o conceito de reconhecimento de expressões faciais usando alguns exemplos de características que podem ser extraídas de uma pessoa e explica a importância de compreendê-las para a interação social. A Seção 2.2 aborda quais são as expressões faciais básicas conhecidas universalmente. A Seção 2.3 apresenta os conceitos de CNNs, notadamente relevantes para classificação de expressões faciais. Na Seção 2.4 apresenta-se o conjunto de dados FER2013 que é utilizado para treinar modelos para classificação de expressões faciais. Por fim, a Seção 2.5 apresenta o modelo de detecção*Single Shot MultiBox Detector* utilizado para detecção facial.

2.1 Reconhecimento de expressões faciais

Ao observar a mudança de expressões faciais, muitas informações sobre os estados emocionais do indivíduo são reveladas (ZHENG et al., 2021). Um exemplo dessas expressões pode ser visualizado na Figura 2.1, que mostra um aluno que está absorto no estudo, um aluno dormindo, em um estado feliz enquanto presta atenção na aula, estudando pacificamente, em estado de aborrecimento por não estar entendendo o conceito e, por fim, o aluno se apresenta em estado frustrado já que não consegue se concentrar.

Portanto, compreender as características extraídas do rosto é uma tarefa fundamental, e tal aptidão é vital em nossas comunicações diárias e interações sociais. Em comunidades de pesquisa, como interação humano-computador (IHC), neurociência e visão computacional, os cientistas conduziram uma extensa pesquisa para entender as emoções humanas. Tais estudos permitiriam a criação de computadores que podem compreender as emoções humanas, assim como nós mesmos, e obter interações contínuas entre humanos e computadores (BARSOUM et al., 2016).

Figura 2.1: Diferentes estados mentais dos seres humanos. Os estados são: antecipação, sono, felicidade, paz, irritação e frustração (GUPTA; KUMAR; TEKCHANDANI, 2023).

2.2 Expressões faciais básicas universais

Entre muitas entradas que podem ser usadas para derivar emoções, a expressão facial é de longe a mais popular. Um dos trabalhos pioneiros de Paul Ekman (EKMAN; OSTER, 1979) identificou 7 emoções que são universais em diferentes culturas. Essas sete emoções básicas são surpresa, tristeza, felicidade, medo, nojo e raiva, que são mostradas na Figura 2.2. A análise da expressão facial pode, portanto, ser realizada analisando em unidades de ação facial para cada uma das partes faciais (olhos, nariz, cantos da boca, etc.) (TIAN; KANADE; COHN, 2001).

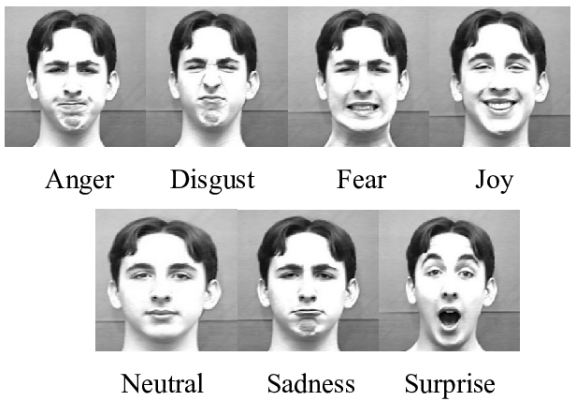


Figura 2.2: Expressões faciais de diferentes emoções: raiva, desprezo, medo, alegria, neutro, tristeza e surpresa (KWONG et al., 2018).

2.3 Modelos CNNs

O reconhecimento de emoções faciais (*facial expression recognition* – FER) é significativo para a análise de interações em contextos *online*. O FER preciso e robusto por modelos de computador continua sendo um desafio devido à heterogeneidade dos rostos humanos e variações nas imagens, como diferentes poses faciais e iluminação. Dentre todas as técnicas para FER, os modelos de aprendizado profundo, especialmente as CNNs, têm mostrado grande potencial devido à sua poderosa extração automática de recursos e eficiência computacional. As CNNs são distintas das redes neurais artificiais tradicionais porque têm a capacidade de codificar recursos de imagem relevantes diretamente das imagens de entrada brutas, tornando mais eficientes para implementar e reduzir o número de parâmetros na rede (SHEN et al., 2022).

Com o avanço da visão computacional, o desempenho do reconhecimento de emoções faciais melhorou e foi alcançado em imagens capturadas em ambientes controlados. Porém, ainda persistem os desafios no reconhecimento de emoções sob condições naturais devido às alterações em pose facial e as diferenças sutis entre as expressões. Na classificação de imagens, as CNNs mostraram grande potencial devido à sua eficiência computacional e capacidade de extração de características (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). Eles são os modelos profundos mais amplamente utilizados para FER (MEHENDALE, 2020).

A rede CNN emprega a operação matemática chamada convolução. A convolução é um tipo especializado de operação linear. Redes convolucionais são simplesmente redes neurais que usam convolução no lugar da multiplicação geral de matrizes em pelo menos uma de suas camadas. Na sua forma mais geral, a convolução é uma operação em duas funções de um argumento com valor real (KIM, 2016). Uma CNN normalmente possui três tipos de camadas: camadas convolucionais, camada de agrupamento (*pooling*) e camadas totalmente conectadas. Um exemplo da arquitetura CNN pode ser vista na Figura 2.3.

As convoluções funcionam como filtros que enxergam pequenos quadrados e vão "deslizando" por toda a imagem captando os traços mais marcantes. Explicando melhor, com uma imagem 32×32 e um filtro que cobre uma área de 5×5 da imagem com movimento de 1 salto (chamado de *stride*), o filtro passará pela imagem inteira, formando

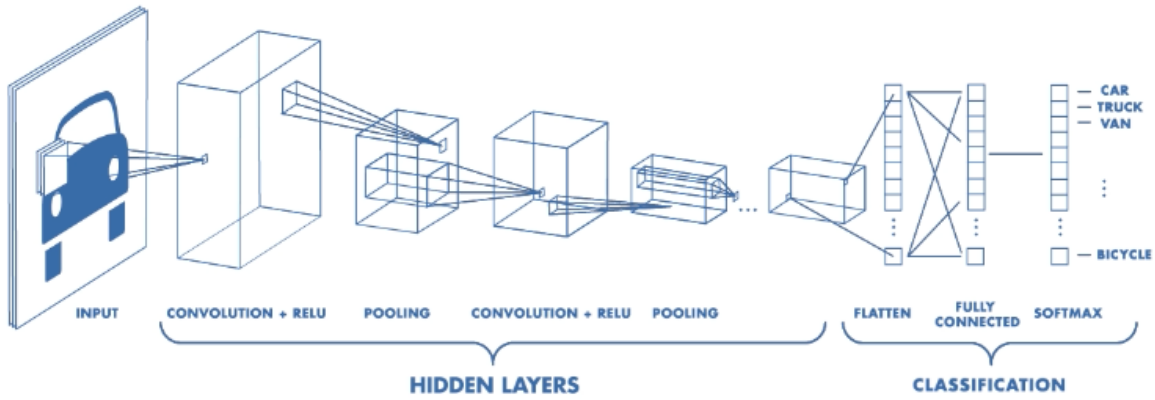


Figura 2.3: Arquitetura de uma CNN (MISHRA, 2020).

no final um *feature map* ou *activation map* de 28 × 28 como ser vista na Figura 2.4 (ALVES, 2018). A redução no tamanho da saída ocorre devido à impossibilidade de aplicação do filtro nas bordas. Isso pode ser evitado com o uso de *padding*, que será discutido mais adiante.

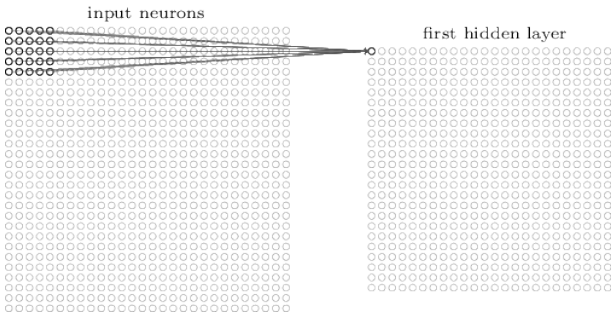


Figura 2.4: Entrada de 28 × 28 dimensões com campo receptivo de área 5 × 5 (ALVES, 2018).

.

O filtro, também conhecido como *kernel*, contém pesos inicializados aleatoriamente e é atualizado durante o processo de retropopagação (*backpropagation*). A região da entrada onde o filtro é aplicado é chamada de campo receptivo (*receptive field*). A compatibilidade entre o filtro e o campo receptivo resulta em um número alto, indicando presença de padrões. A ausência de compatibilidade se reflete em um resultado próximo a zero (ALVES, 2018).

Além de determinar o tamanho do filtro e o passo da convolução como hiperparâmetros, ao modelar uma CNN, é essencial tomar decisões sobre o uso de *padding*.

A não inclusão de padding resulta em uma imagem de saída menor do que a entrada, devido a impossibilidade de aplicar o filtro nas bordas da imagem. Uma forma comum de *padding* é adicionar uma borda preenchida com zeros, que permite que a imagem de saída tenha o mesmo tamanho da imagem original. A inclusão de *padding* tem o propósito de evitar que as camadas diminuam excessivamente rápido, proporcionando um aprendizado mais eficaz (ALVES, 2018).

É prática comum intercalar periodicamente uma camada de *pooling* entre camadas convolucionais consecutivas. Essa camada desempenha a função de reduzir gradualmente o tamanho espacial da representação, resultando na diminuição do número de parâmetros e da carga computacional na rede. Dessa forma, ela desempenha um papel crucial no controle do sobreajuste (*overfitting*) (BYUN, 2023). A camada de *pooling* opera de maneira independente em cada fatia de profundidade da entrada, realizando uma redução espacial por meio da operação de agrupamento, comumente uma operação de máximo. Um arranjo comum consiste em uma camada de *pooling* com filtros de tamanho 2×2, aplicados com um passo de 2. Esse arranjo reduz a amostragem de cada fatia de profundidade em 2 ao longo da largura e da altura, descartando 75% das ativações. Nesse cenário, cada operação de máximo envolveria, no máximo, 4 valores, correspondendo a uma pequena região 2×2 em alguma fatia de profundidade. Importante ressaltar que a dimensão de profundidade permanece inalterada durante esse processo (BYUN, 2023).

Finalmente, a tarefa de classificação é feita pela camada totalmente conectada. Sua entrada é a saída da camada anterior e sua saída são N neurônios, com N sendo a quantidade de classes do seu modelo para finalizar a classificação (PERES, 2021).

2.4 Conjunto de dados FER2013

O conjunto de dados FER2013 é usado para treinar modelos de classificação de expressões faciais. Esse conjunto é bastante conhecido na literatura, sendo composto por imagens de apenas 48×48 *pixels* e pode ser usado para comparar o desempenho dos métodos em situações de pouca luz (GOODFELLOW et al., 2013). O conjunto consiste em 35887 imagens rotuladas, divididas em 28709 imagens para treinamento, 3589 imagens públicas dedicadas para validação e teste e outras 3589 imagens de teste privadas (TALE-

Figura 2.5: Imagens de amostra do conjunto de dados FER2013 (TALEGAONKAR et al., 2019).

Tabela 2.1: Descrição do conjunto de dados FER2013 (TALEGAONKAR et al., 2019).

Rótulo	Número de amostras	Emoção
0	4593	Irritado
1	547	Nojo
2	5121	Medo
3	8989	Feliz
4	6077	Triste
5	4002	Surpreso
6	6198	Neutro

GAONKAR et al., 2019). O conjunto de dados FER2013 contém imagens que variam do ponto de vista da iluminação e escala. A Figura 2.5 mostra algumas imagens de amostra do conjunto de dados FER2013 e a Tabela 2.1 apresenta a distribuição do conjunto de dados de acordo com as classes.

2.5 *Single Shot MultiBox Detector*

A SSD, introduzida por Liu et al. (2016), destaca-se como a segunda geração de redes para detecção de objetos na literatura. A estrutura central da SSD consiste em um modelo de classificação, conhecido como *backbone*, e o próprio detector SSD.

A arquitetura original da SSD, ilustrada na Figura 2.6, baseia-se na rede pré-treinada de classificação VGG16, mas admite a substituição por outras redes como *MobileNet* e *ResNet*. O *backbone* da SSD desempenha um papel crucial na extração de características em múltiplas escalas, produzindo caixas envolventes (*bounding boxes*) e suas respectivas classes.

As principais contribuições da SSD incluem o uso de filtros convolucionais de pequenas dimensões para prever classes de objetos, o alinhamento das posições de caixa

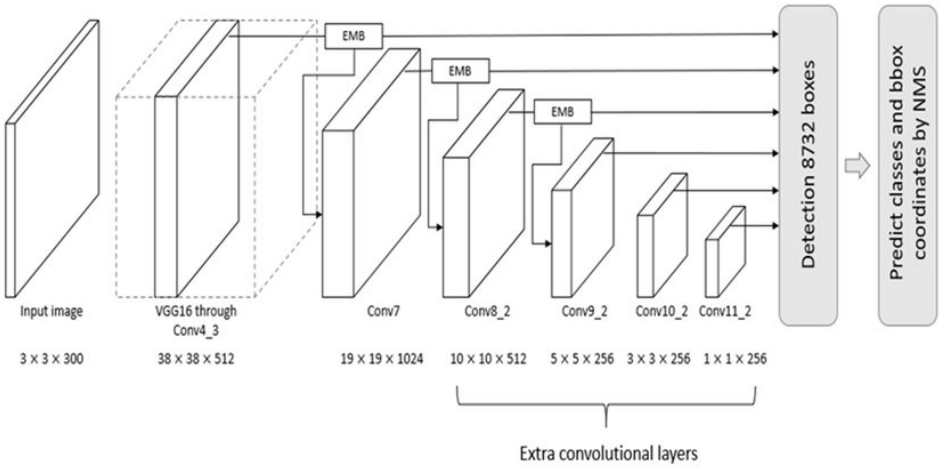


Figura 2.6: Arquitetura da SSD (LIU et al., 2016).

envolvente, a segmentação de caixas com base em sua razão de aspecto e a criação de representações de objetos para aprendizado em diversas escalas (ZOU et al., 2023).

A SSD aprimora a rede VGG16 ao adicionar camadas convolucionais que criam representações piramidais das imagens em diferentes escalas (LIU et al., 2016). Esse método visa garantir a detecção eficiente de objetos de diferentes tamanhos, tornando a rede invariante espacialmente. Ela não divide a imagem em grades fixas, mas prediz caixas de ancoragem para todas as localizações no mapa de características da imagem. Durante o treinamento, a SSD utiliza perdas de confiança e localização para otimizar a precisão da classificação e a localização das caixas preditas em relação aos objetos reais.

A rede passa por processos como normalização de caixas de ancoragem, mineração de negativos difíceis (*hard negative mining*) e aumento de dados antes da detecção (WENG, 2018). O equilíbrio entre positivos e negativos é mantido usando a técnica de mineração de negativos difíceis (LIU et al., 2016). Além disso, a aplicação de técnicas de aumento de dados, como o uso de *patches* randômicos, contribui para treinar um modelo mais robusto e sensível a diferentes tamanhos de objetos. Embora a SSD seja eficaz na detecção de objetos precisa e em tempo real, sua complexidade e exigência de grande conjunto de dados podem apresentar desafios de implementação em certas situações (LIU et al., 2016). Embora apresente desafios de implementação devido à sua complexidade e requisitos de grandes conjuntos de dados, a SSD destaca-se como uma ferramenta eficiente para a tarefa de detecção facial, o que a torna uma escolha relevante para o presente

trabalho.

3 Trabalhos Relacionados

Neste capítulo apresenta-se o método utilizado na busca de trabalhos relacionados para encontrar os artigos similares de maior relevância para o desenvolvimento do presente trabalho. Também são apresentados 4 (quatro) artigos de forma detalhada, da Seção 3.1 a Seção 3.4, com o devido impacto, tanto positivo quanto negativo na presente pesquisa. Por fim, a Seção 3.5 apresenta as considerações finais.

Para a recuperação de trabalhos relacionados à presente pesquisa foi utilizada uma *String* de busca com as seguintes palavras: Detecção de emoção facial, classificação de gênero, nível de engajamento, *classifying emotions*, *emotion detection using face*, *emotion detector system*, *Emotion detection using deep learning*, nas bases de dados científicas do IEEE e Google Scholar. Esta *String* consiste na junção de termos chaves associados ao tema proposto nesta pesquisa. Dentre os trabalhos encontrados, foram selecionados os mais relevantes para o presente trabalho, os quais são descritos neste capítulo.

Sendo assim, os artigos apresentados serviram de base para o correto desenvolvimento desta monografia. A seguir são apresentados os trabalhos relacionados mais relevantes.

3.1 Rede neural convolucional em tempo real para classificação de emoção e gênero

O estudo de Arriaga, Valdenegro-Toro e Plöger (2017) propõe a implementação de uma estrutura geral para a construção de redes neurais para aplicações em tempo real. Esta abordagem foi validada por meio da implementação de um sistema de visão em tempo real, capaz de executar simultaneamente tarefas como detecção de rosto, classificação de gênero e avaliação de emoções em uma única etapa, utilizando a arquitetura CNN desenvolvida. Adicionalmente, o artigo fornece detalhes sobre o processo de treinamento dos modelos e apresenta os resultados obtidos em conjuntos de *benchmark* padrão.

A abordagem proposta neste artigo destaca-se pela eficiência em tempo real na implementação de CNNs. Essa capacidade é especialmente valiosa em aplicações que requerem processamento rápido de informações visuais, como sistemas de visão computacional em robótica. A habilidade de realizar tarefas de detecção de rosto, classificação de gênero e classificação de emoções em tempo real é crucial para melhorar a interatividade e a tomada de decisões em ambientes onde a latência é um fator crítico.

Embora o artigo apresente resultados promissores, a acurácia de 66% na classificação de emoções no conjunto de dados FER2013 é um desafio evidente. A classificação de emoções é uma tarefa complexa devido à variabilidade das expressões faciais e ao contexto em que são feitas. A melhoria nessa área é essencial para tornar a tecnologia mais confiável em aplicações que dependem de uma análise precisa das emoções humanas. O trabalho, por sua vez, não se propõe a realizar a detecção do nível de engajamento, diferenciando-se do escopo do presente estudo.

Vale destacar que esse trabalho desempenhou um papel importante como referência na abordagem adotada para a classificação e emoção no presente trabalho.

3.2 Reconhecimento de expressões faciais usando rede convolucional atencional

No trabalho de Minaee, Minaei e Abdolrashidi (2021) foi proposto uma abordagem de aprendizado profundo utilizando rede convolucional atencional. Esse modelo é capaz de focar em partes importantes da face e alcançar uma melhoria significativa em relação aos modelos anteriores em vários conjuntos de dados, incluindo FER2013, CK+, FERG e JAFFE. Também é usada uma técnica de visualização que é capaz de encontrar regiões faciais importantes para detectar diferentes emoções com base na saída do classificador. Além disso, mostra-se através de resultados experimentais que diferentes emoções são sensíveis a diferentes partes da face.

O trabalho propõe um modelo, que primeiramente realiza a extração de recursos. Esse modelo consiste em quatro camadas convolucionais, com cada uma seguida por uma camada de *pooling* e uma função de ativação de unidade linear retificada (ReLU), além de

duas camadas totalmente conectadas. Além disso, é utilizado um módulo transformador espacial que tenta focar nas partes mais relevantes da imagem, estimando uma amostra sobre a região de interesse. Por fim, a seção que mostra a arquitetura do modelo proposto acrescenta que foi necessário usar uma rede com menos camadas, que tem a velocidade de inferência muito mais rápida e é mais adequada para aplicações em tempo real.

Após a seção que descreve o modelo proposto, há uma seção de resultados experimentais que fornece a análise detalhada do modelo implementado em vários conjuntos de dados de classificação de expressões faciais. A análise realizada em diversos conjuntos de dados pode ser considerada um ponto positivo deste trabalho, uma vez que representa uma defesa para o modelo proposto, mostrando assim seu desempenho. Essa análise experimental foi realizada em cima de quatro conjuntos de dados populares, sendo eles: FER2013, Cohn-Kanade, JAFFE e FERG. O artigo destaca que foi um desafio utilizar o conjunto de dados FER2013, pois há uma natureza de dados desequilibrada, em vista das diferentes classes emocionais, e algumas classes como felicidade e neutro possuem mais exemplos do que as outras.

Embora a abordagem de aprendizagem profunda proposta baseada em uma rede convolucional atencional demonstre melhor desempenho em vários conjuntos de dados, incluindo FER2013, CK+, FERG e JAFFE, não há menção explícita ou avaliação da eficiência do modelo em aplicações em tempo real. O reconhecimento de expressões faciais em tempo real é crucial para aplicações como interação humano-computador, realidade virtual e sistemas sensíveis às emoções.

3.3 Detecção de envolvimento do aluno usando métodos multimodais

O trabalho de Sharma et al. (2022) apresenta um sistema para detectar o nível de engajamento dos alunos usando apenas informações fornecidas pela típica *webcam* embutida em um computador *laptop* e foi projetado para funcionar em tempo real. O sistema combina informações sobre os movimentos dos olhos e da cabeça e emoções faciais para produzir o índice de concentração com três classes de engajamento: muito engajado,

nominalmente engajado e nada engajado. O sistema foi testado em um cenário típico de aprendizado *online* e os resultados mostraram que ele identifica corretamente cada período de tempo em que os alunos estavam engajados, nominalmente ou nada engajados. Adicionalmente, os resultados também mostraram que os alunos com melhores notas apresentaram maior índice de concentração.

O sistema proposto conta com a participação do instrutor e do aluno, quando o aluno está interagindo com o material didático, os dados da imagem do aluno (capturados pela *webcam*) são automaticamente analisados pelo sistema para avaliar o nível de concentração do aluno. Se o índice de concentração resultante cair abaixo de um valor limite pré-definido, um alerta é emitido para o professor e para o aluno.

Durante o artigo, são apresentados os algoritmos que foram usados no sistema para a verificação do engajamento: Cascata Haar e redes neurais convolucionais. O sistema detector primeiro realiza a detecção do rosto e região dos olhos do aluno utilizando a Cascata Haar. Após isso, o trabalho se baseia no modelo de (ARRIAGA; VALDENEGRO-TORO; PLÖGER, 2017) para realizar a classificação de emoções. Para calcular o nível de engajamento, o algoritmo se baseia no movimento dos olhos e o peso atribuído para as emoções dominantes. Como ponto negativo para este trabalho, especificamente faltam informações sobre o algoritmo implementado para classificação do engajamento. Essa ausência de detalhes impacta negativamente na compreensão do método empregado.

Vale ressaltar que esse trabalho desempenhou um papel fundamental como referência na abordagem adotada para o cálculo do nível de engajamento no presente estudo. A escolha desses métodos e a compreensão das categorias de engajamento foram influenciadas diretamente pelos princípios delineados por (SHARMA et al., 2022). Além disso, ao reconhecer a importância de uma explicação detalhada do algoritmo, este estudo busca preencher a lacuna identificada no trabalho de referência, proporcionando uma compreensão mais abrangente do processo de cálculo do nível de engajamento.

3.4 Sistema de detecção *online* de engajamento

O estudo de Gupta, Kumar e Tekchandani (2023) se refere à criação de um sistema de detecção de engajamento, capaz de identificar automaticamente o nível de envolvimento

do aluno em situações de tempo real. O artigo propõe uma abordagem baseada em métodos de aprendizado profundo. O trabalho realizou a detecção facial com a ajuda do modelo pré-treinado *Faster R-CNN*. Além disso, modelos de aprendizado profundo como *Inception-V3*, *VGG19* e *ResNet-50* foram implementados para cenários de aprendizado em tempo real para classificar as emoções dos alunos. Um algoritmo de avaliação de engajamento é proposto para calcular o índice de engajamento a partir de dados de saída de classificação de emoção facial. Com isso o sistema decide se o aluno está engajado ou não durante a aula *online*.

Durante a leitura deste trabalho, é possível notar alguns pontos positivos, como o detalhamento dos conjuntos de dados que foram usados para o experimento. Em vista disso, são descrito brevemente os conjuntos Wider face, FER2013, CK+, RAF-DB e Own dataset, além de pontuar os pontos positivos e negativos de cada base de dados. Além disso, foram descritas as técnicas utilizadas para a implementação do sistema e os métodos e modelos usados para fazer a detecção facial, pontuando os erros encontrados e porque tais métodos apresentaram melhor desempenho. Outro ponto que vale ressaltar, é que ao final do artigo, são apresentados os resultados experimentais e as análises comparando o desempenho de cada conjunto de dados, ressaltando que o modelo *ResNet-50* obteve um melhor desempenho.

Ao final da leitura, é possível marcar um ponto negativo referente à descrição do algoritmo usado para calcular o nível de engajamento. Os autores apresentam um trecho do algoritmo, mas fazem uma abordagem bastante sucinta, não detalhando o motivo para escolha das abordagens utilizadas nos cálculos. Somado a isso, o trabalho não apresenta as telas do sistema para exemplificar melhor o funcionamento do mesmo, limitando-se a citar como são as telas.

3.5 Considerações finais

Neste capítulo, foram descritos os trabalhos relacionados de maior relevância, com a apresentação dos resultados alcançados, pontos positivos e negativos de cada um deles. Além disso, os trabalhos citados nas Seções 3.1 e 3.4 foram usados como base para o classificador deste trabalho.

Desta forma, todos os 4 (quatro) trabalhos são importantes, tanto em métodos e modelos abordados. Eles podem servir como arcabouço para o desenvolvimento de um sistema para análise de videochamadas, com tecnologias e metodologias atuais e focadas na classificação de emoções.

Os problemas encontrados durante a realização destes trabalhos também possuem um grande valor quando aplicados em melhorias nos trabalhos posteriores, evitando principalmente a propagação de erros já conhecidos.

No presente trabalho, visando compreender melhor o problema, serão realizadas análises aplicando uma implementação do método proposto por Sharma et al. (2022) para o cálculo do nível de engajamento, além disso, usa como base para classificação emocional e de gênero o trabalho proposto por e Arriaga, Valdenegro-Toro e Plöger (2017). A partir dessa análise, espera-se compreender os principais desafios para construção de um sistema para análise de engajamento em videochamadas em contextos como educacional e empresarial.

Este trabalho se destaca por apresentar um classificador que não apenas realiza a classificação de emoções, mas também incorpora a classificação de gênero e o monitoramento do nível de engajamento. Além disso, este estudo vai além ao experimentar e validar o desempenho do classificador em cenários reais, proporcionando uma compreensão mais robusta de sua eficácia e aplicabilidade em situações práticas.

4 Modelo proposto

O modelo utilizado neste trabalho se baseia na implementação dos trabalhos de Arriaga, Valdenegro-Toro e Plöger (2017) e Sharma et al. (2022). Essas implementações estão disponíveis nos repositórios Arriaga (2020) e Yagoub (2023), respectivamente. A Figura 4.1 apresenta um fluxograma que ilustra as etapas e componentes envolvidos no classificador. Este fluxograma apresenta uma visão geral das decisões tomadas pelo modelo, desde a entrada de um vídeo até a saída das previsões de emoção, gênero e nível de engajamento.

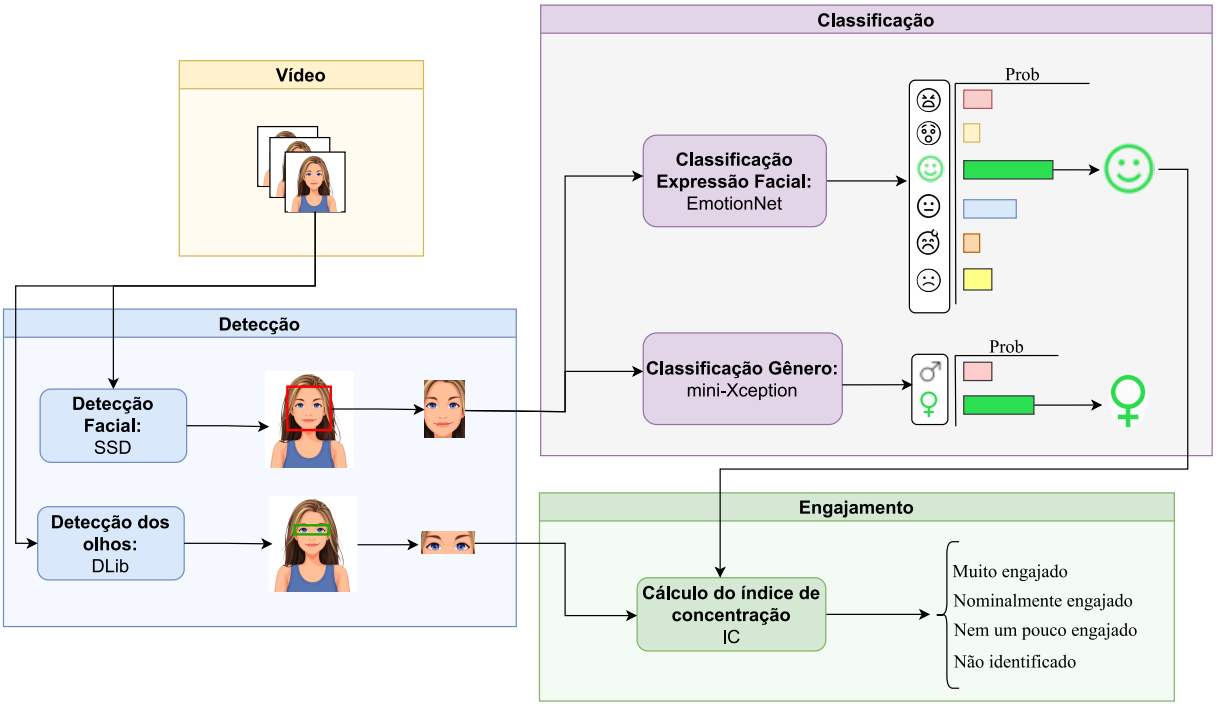


Figura 4.1: Etapas do modelo para obter o nível de engajamento, emoção e gênero.

4.1 Detecção facial

A detecção facial é realizada por modelo baseado em *Deep Learning*, que faz parte do conjunto de modelos da biblioteca OpenCV e é implementado usando o *framework* Caffe. Esse modelo é carregado a partir do arquivo **res10_300x300_ssd_iter_140000_fp16.caffemodel** utilizado para detecção facial. Ele é uma implementação da rede neural

SSD e foi treinado em imagens para reconhecer e localizar rostos. O res10 se refere ao *backbone* da rede (ResNet10). Além disso, o nome reflete informações sobre o tamanho da imagem de entrada (300 × 300 pixels), o tipo de dados para operações em ponto flutuante de 16 bits (fp16) e o número de iterações de treinamento (140.000).

Cada *frame* é processado na etapa de vídeo do fluxograma, que pode ser vista no diagrama da Figura 4.1. O *frame* é pré-processada para criar um *blob*, uma representação adequada para o modelo de rede neural utilizado. Este *blob* é então utilizado como entrada para o modelo de detecção facial pré-treinado, que retorna as detecções contendo caixas delimitadoras dos rostos identificados e suas respectivas probabilidades. Posteriormente, detecções consideradas fracas, com probabilidades abaixo de 0,3 são filtradas. Por fim, a face é recortada da imagem de acordo com as dimensões das caixas delimitadoras encontradas.

A detecção dos olhos é executada por meio da biblioteca Dlib, que oferece recursos avançados para identificação de pontos faciais, conhecidos como *landmarks*. A Figura 4.2 ilustra um exemplo de detecção utilizando *landmarks* faciais, onde pontos-chave em uma face foram identificados. A partir desses pontos, foram desenvolvidos dois métodos distintos para calcular a razão de piscar e a razão de olhar.

Figura 4.2: *Landmarks* são usados para rotular e identificar os principais atributos faciais em uma imagem (KING, 2014).

Para o cálculo da razão de piscar (RP), duas linhas são definidas em cada olho.

Uma linha horizontal conecta os cantos esquerdo e direito dos olhos, e uma linha vertical conecta os pontos médios superior e inferior. O valor da razão é determinado pela divisão entre os comprimentos da linha vertical e horizontal. A *RP* final é determinada pela média do olho esquerdo e direito.

A razão de olhar (*RO*) é calculada com base nos pontos faciais detectados. Ela isola a região dos olhos, e realiza uma limiarização (*thresholding*), ou seja, converte a imagem em escala de cinza para uma binária, para destacar a íris. A seguir, são contados os *pixels* brancos nas metades esquerda e direita da imagem limiarizada. Por fim, calcula-se a razão entre a quantidade de *pixels* brancos do lado esquerdo e direito. O valor de *RO* é dado pela média entre os olhos direito e esquerdo. Valores dessa razão indicam que os olhos estão focados na tela.

As fundamentações dessas razões derivam da adaptação proposta por Yagoub (2023) ao estudo conduzido por Sharma et al. (2022). Vale destacar que a abordagem de Yagoub (2023) concentrou-se exclusivamente no uso do olho esquerdo para esses cálculos, porém, no presente trabalho, optou-se por utilizar a média da razão dos dois olhos, por considerar que seria uma medida mais coerente. Os valores de *RP* e *RO* serão empregadas em uma etapa posterior para determinar o nível de engajamento.

4.2 Classificação de expressões faciais

Para fazer a classificação de expressões faciais, foi utilizado o modelo de CNN EmotionNet pré-treinado no conjunto de dados FER2013 baseado na implementação do autor Goodfellow et al. (2013). A arquitetura é composta por camadas convolucionais seguidas por camadas totalmente conectadas para realizar a classificação. A rede implementada foi inspirada na arquitetura VGG13 de Gupta, Kumar e Tekchandani (2023), porém com um pequeno ajuste na camada totalmente conectada e nas funções de ativação usadas em toda a rede.

A detecção obtida pelo modelo de detecção SSD (vista na Seção 4.1) passa por um pré-processamento envolvendo a conversão para escala de cinza (*grayscale*), redimensionamento para o tamanho desejado (48 x 48 *pixels*) e conversão para um tensor. A imagem pré-processada é então enviada para o modelo de rede neural (EmotionNet) para

inferência. O modelo produz previsões de probabilidade para cada classe de emoção. A função de ativação *softmax* é aplicada às previsões para normalizar as probabilidades, garantindo que somem 1. A emoção mais provável é considerada a emoção dominante para aquela imagem e é exibida. Vale ressaltar que para classificação de emoção, o modelo proposto por Goodfellow et al. (2013) não utiliza a emoção **Nojo**. Portanto, esse classificador não contém essa emoção.

4.3 Cálculo de engajamento

No trabalho de referência de Sharma et al. (2022), foi apresentada uma abordagem multimodal destinada a avaliar o engajamento em tempo real. O sistema apresentado por eles categoriza os níveis de engajamento dos alunos em três categorias: muito engajado, nominalmente engajado e nem um pouco engajado. No presente trabalho, seguiu-se a adaptação realizada por Yagoub (2023) para o cálculo do nível de engajamento. Além disso, a detecção da região dos olhos é utilizada para o cálculo de engajamento. Se não for possível fazer essa detecção, então o modelo que foi desenvolvido no presente trabalho adotou o retorno de mais um estado, denominado não identificado. Foi necessário acrescentar essa mudança no algoritmo para o estudo no momento da experimentação.

A partir das razões RP e RO descritas na Seção 4.1, define-se o peso dos olhos (PO) da seguintes forma:

- Se RP for menor que 0,2 então PO é definido como 0, indicando que os olhos estão fechados;
- Se RP estiver entre 0,2 e 0,3 então PO é definido como 1,5, indicando que os olhos estão semi-abertos;
- Se RP for maior que 0,3 indica que os olhos estão abertos, então PO é determinado com base no cálculo de RO. O cálculo de RO indica se a pessoa está olhando na direção da tela. Se o valor de RO estiver no intervalo entre 1 e 1,7, então PO é definido como 5; caso contrário, é definido como 2;

O índice de concentração (IC) resultante é determinado multiplicando-se o PO,

definido anteriormente, pelo peso emocional (PE) da emoção dominante (obtida na etapa de classificação de emoções na Seção 4.2). A Tabela 4.1 apresenta os pesos emocionais que serão usados no cálculo do IC. O resultado é dividido pelo valor máximo possível 4,5. Dessa forma, tem-se a equação:

$$IC = \frac{PE \times PO}{4,5}. \tag{4.1}$$

Tabela 4.1: Emoções dominantes e seus pesos correspondentes.

Emoção dominante	Peso emocional (PE)
Neutro	0,9
Feliz	0,6
Surpreso	0,6
Triste	0,3
Raiva	0,25
Surpreso	0,3

O valor máximo possível de 4,5 é utilizado como um fator de normalização para garantir que o IC resultante esteja na faixa de 0 a 1,0. O cálculo do IC é uma combinação de pesos associados às emoções PE e aos olhos PO, e o objetivo é normalizar esse valor para uma escala mais interpretável.

O nível de engajamento foi definido em quatro classes:

1. **Muito engajado:** valor do IC é maior que 65%.
2. **Nominalmente engajado:** valor do IC do aluno está entre 25% e 65%.
3. **Nem um pouco engajado:** valor do IC é menor que 25%.
4. **Não identificado:** O cálculo do IC retorna 0.

Cabe salientar que é possível ainda fazer uma análise levando em consideração somente a expressão facial ou apenas o movimento dos olhos. Para IC_E calculado levando em conta apenas a expressão facial, o peso emocional (PE) é dividido pelo valor máximo possível 0,9, podendo ser definido pela equação:

$$IC_E = \frac{PE}{0,9}. \tag{4.2}$$

Já o IC_O calculado levando em conta apenas a informação dos olhos divide o peso dos olhos (PO) pelo valor máximo possível **5**, conforme equação:

$$IC_O = \frac{PO}{5,0}. \tag{4.3}$$

Essas duas abordagens distintas buscam avaliar a concentração a partir de perspectivas complementares.

4.4 Detecção de gênero

A abordagem de Arriaga, Valdenegro-Toro e Plöger (2017) propõe treinar o modelo **mini-Xception** para fazer a classificação de gênero. O modelo é uma rede neural treinada em um conjunto de dados que inclui exemplos de rostos rotulados por gênero. Essa abordagem foi utilizada como base para este trabalho. Utilizou-se o modelo pré-treinado no conjunto de dados do IMDB (ROTHE; TIMOFTE; GOOL, 2018), que contém 460.723 imagens RGB, onde cada imagem pertence à classe mulher ou homem. O modelo alcançou a acurácia de 96% neste conjunto de dados.

A imagem da face detectada na etapa descrita na Seção 4.1 é pré-processada por meio de normalização e escala para ser submetida à **mini-Xception** para inferência. A rede gera uma saída que representa a probabilidade da imagem pertencer a cada uma das classes de gênero: mulher ou homem. A classe com a probabilidade mais alta é escolhida como a previsão de gênero para o rosto.

4.5 Saída do classificador

A Figura 4.3 exibe a saída do classificador onde é possível verificar um *frame* que exibe o vídeo de entrada, com retângulos delimitadores indicando a localização do rosto detectado. É possível visualizar informações do nível de engajamento, gênero e a emoção predominante. Além disso, um gráfico de barras representa a distribuição das probabilidades das emoções.

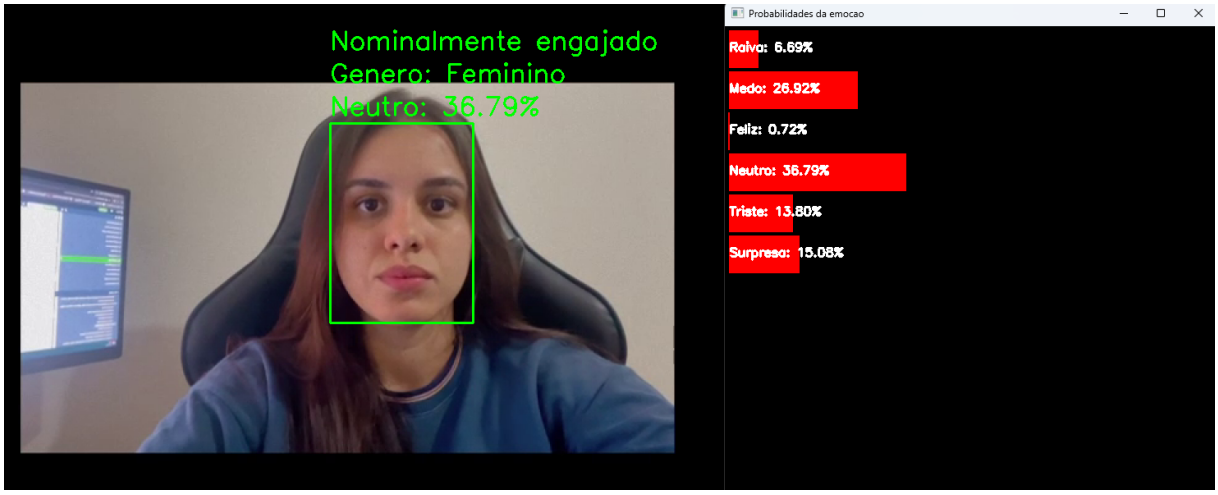


Figura 4.3: Saída do classificador exibindo o nível de engajamento, gênero e a emoção predominante, além do gráfico de probabilidades das emoções.

5 Experimentos e Resultados

Neste capítulo, são apresentados os experimentos conduzidos para avaliar o classificador. Será apresentado o conjunto de dados que foi construído para realização dos experimentos. Além disso, os cenários escolhidos serão discutidos em detalhes, fornecendo uma descrição abrangente e, em seguida, são analisados os resultados obtidos durante os experimentos.

5.1 Conjunto de dados

Para a realização dos experimentos, optou-se por utilizar um cenário real no ambiente corporativo, envolvendo a gravação das câmeras de três indivíduos durante uma reunião de trabalho. Esses vídeos foram posteriormente empregados como dados de entrada para o classificador, permitindo a análise das expressões faciais, gênero e nível de engajamento em um contexto dinâmico.

Cada indivíduo registrou sua câmera ao longo de toda a reunião e, posteriormente, os vídeos foram segmentados em períodos de 50 segundos para cada cenário específico em momentos predefinidos. Os vídeos foram segmentados devido à sua extensão, com duração total de duas horas. Essa medida foi adotada para otimizar a análise, visto que vídeos extensos podem dificultar a identificação de padrões específicos e o aprofundamento na compreensão de diferentes cenários. Foram estabelecidos quatro cenários distintos, cada um representando diferentes contextos e dinâmicas de interação. Apesar de não se enquadrar em um ambiente educacional, esse cenário empresarial demanda atenção, foco, discussões em grupo e outros aspectos relevantes para o entendimento do engajamento e das emoções dos participantes.

Para fins de apresentação dos resultados, os indivíduos serão designados da seguinte forma:

- Participante 1: Mulher

- Participante 2: Homem
- Participante 3: Mulher

Dessa forma, o classificador foi aplicado aos vídeos individuais de cada participante, analisando segmentos de 50 segundos em cada cenário. Esse procedimento resultou na obtenção de 1500 *frames* para cada vídeo, que foram agrupados em 50 conjuntos de 30 *frames*, ou seja, com 1 segundo de duração. É importante mencionar que o Participante 3 foi incluído somente no cenário 1 pois o mesmo apresentou problemas na gravação dos demais cenários.

Os vídeos foram registrados por meio da ferramenta **OBS Studio**. O Participante 1 escolheu utilizar a câmera do celular, posicionando-a ligeiramente acima de seu rosto. Já o Participante 2 optou por transformar a câmera do celular em uma *webcam*, colocando-a em seu lado. Por fim, o Participante 3 utilizou a câmera embutida no notebook, capturando seu rosto de frente. Todos os vídeos estão em formatos **mp4**. Além disso, em todos os vídeos foram configurados com uma borda extra onde foi possível fixar uma resolução de 640×480.

No contexto desses experimentos, é relevante observar que todos os participantes forneceram consentimento e assinaram um termo de uso de imagem para participar do estudo. Esse procedimento foi adotado para garantir a conformidade ética e a privacidade dos participantes durante a análise das gravações.

5.2 Cenário 1: comunicação diante da câmera

No primeiro cenário, foi possível utilizar o vídeo gravado pelos três participantes. O contexto abordado foi uma discussão pós-*Sprint*, que ocorre a cada 15 dias. Durante essa reunião, cada indivíduo compartilhou suas percepções sobre os aprendizados e desafios enfrentados durante a *Sprint*, detalhando suas experiências e destacando as principais lições aprendidas e as demandas mais relevantes para o time.

O objetivo deste cenário é simular e avaliar o comportamento do classificador em relação à comunicação diante da câmera. O experimento busca entender como o classificador reage à interação do indivíduo com a câmera, identificando as principais emoções

presentes nesse contexto específico. Além disso, visa medir o nível de engajamento do participante ao discutir sobre os aprendizados e desafios, assim como avaliar a capacidade do classificador em categorizar adequadamente o gênero dos indivíduos.

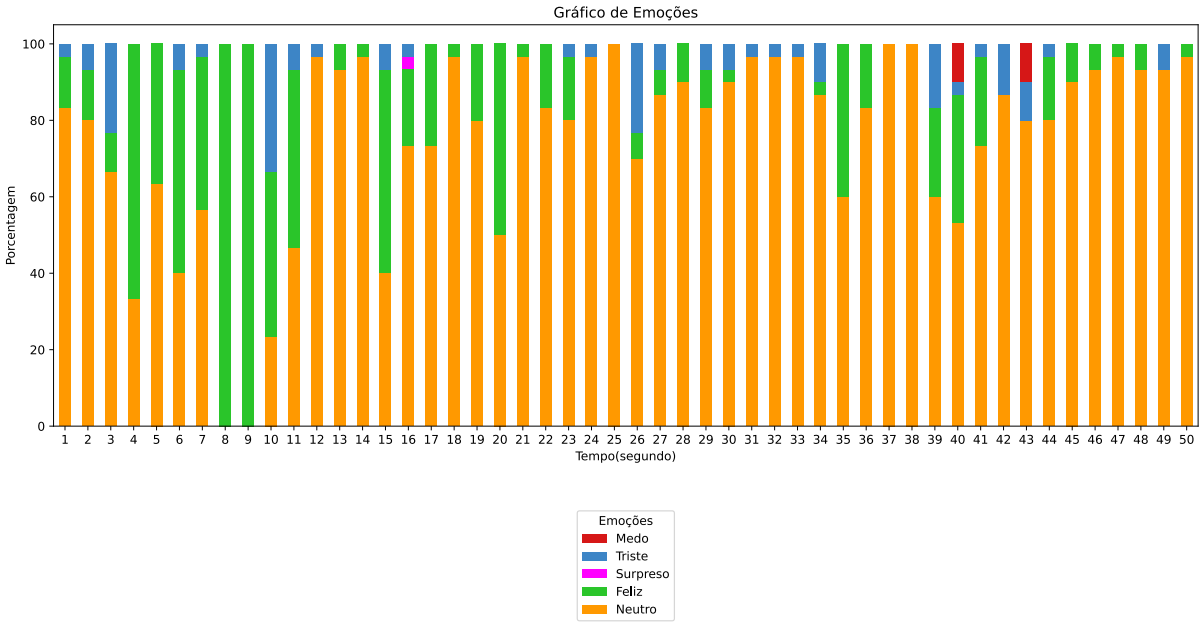


Figura 5.1: Emoções por segundo para o Participante 1 no cenário 1.

Na Figura 5.1 é possível analisar as emoções predominantes para o Participante 1 ao longo de cada segundo do vídeo. Inicialmente, a emoção **Neutro** se destaca, especialmente nos primeiros segundos, enquanto o participante compartilha informações sobre sua *Sprint*. Nota-se uma transição a partir do segundo 6, em que é alterada para emoção **Feliz**. Uma análise qualitativa revela uma alteração nas expressões faciais do Participante 1 nesse momento. Posteriormente, a emoção **Neutro** retorna à predominância. Além disso, em alguns momentos, observam-se *frames* que também exibem as emoções **Triste**, **Medo** e **Surpreso**.

Na Figura 5.2, em relação ao Participante 2, pode-se observar que a emoção mais proeminente ao longo do tempo foi a **Triste**, predominando na maior parte da análise. Entretanto, nos segundos 36 e 37, a emoção **Feliz** ganhou destaque, indicando uma alteração perceptível na expressão facial do Participante 2. Além dessas emoções proeminentes, é notável a presença intermitente das emoções **Medo** e **Neutro** em determinados momentos.

Ao analisar as emoções por segundo do Participante 3, destaca-se que a emoção

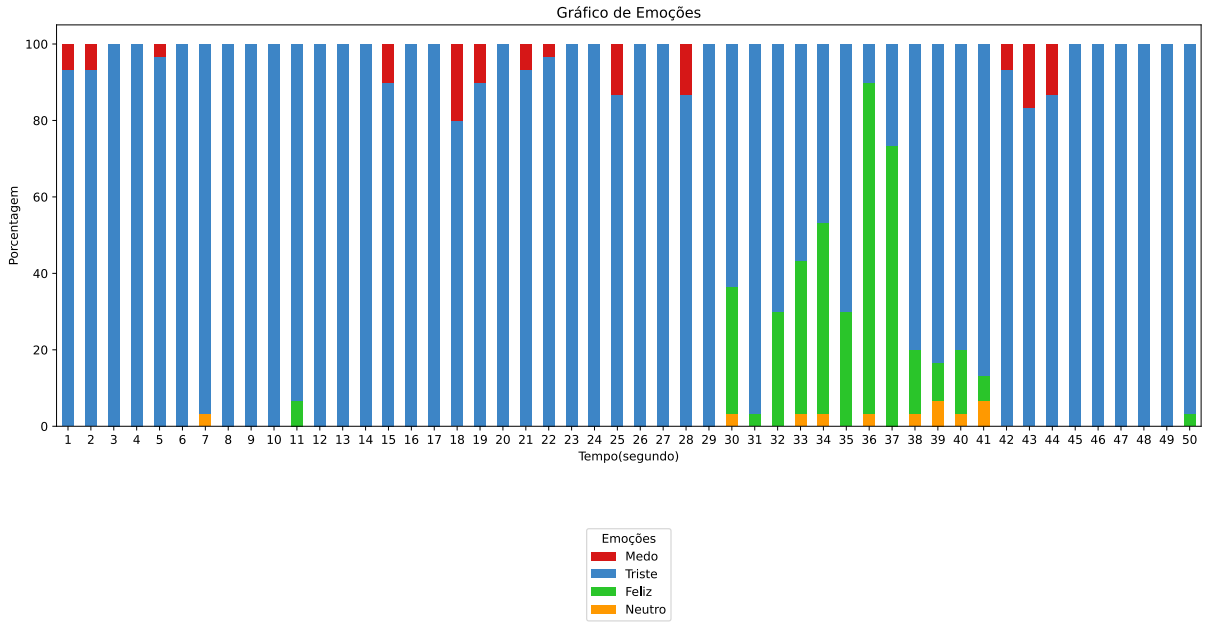


Figura 5.2: Emoções por segundo do Participante 2 no cenário 1.

Neutro prevaleceu na maior parte do tempo. Contudo, no intervalo entre 27 e 30, ou até mesmo nos segundos: 18, 20 e 31 a emoção **Feliz** se destaca. Além dessas emoções, outras duas expressões faciais também estão presentes: **Raiva** e **Medo**, mesmo que sem predominância.

Tabela 5.1: Porcentagem do gênero encontrados ao longo dos *frames* do vídeo para os participantes 1, 2 e 3 no cenário 1.

Participante	Mulher	Homem
1	**61,16%**	38,84%
2	5,78%	**94,22%**
3	**100%**	0%

Conforme mencionado anteriormente, o classificador também realizou a classificação de gênero para os três participantes. A Tabela 5.1 exibe as porcentagens calculadas no cenário 1 a partir dos 1500 *frames* de cada participante. Conforme esperado, o Participante 1 obteve a maior porcentagem para o gênero feminino, totalizando 61,16%. Apesar disso, o percentual de erro foi alto, totalizando 38,84%. O Participante 2 foi classificado como homem com uma porcentagem de 94,22%, enquanto o Participante 3 apresentou uma classificação de 100% para o gênero feminino. Assim, o classificador obteve uma acurácia significativa na classificação de gênero para os participantes.

Na Figura 5.4, é possível analisar o Índice de Concentração (IC) do Participante

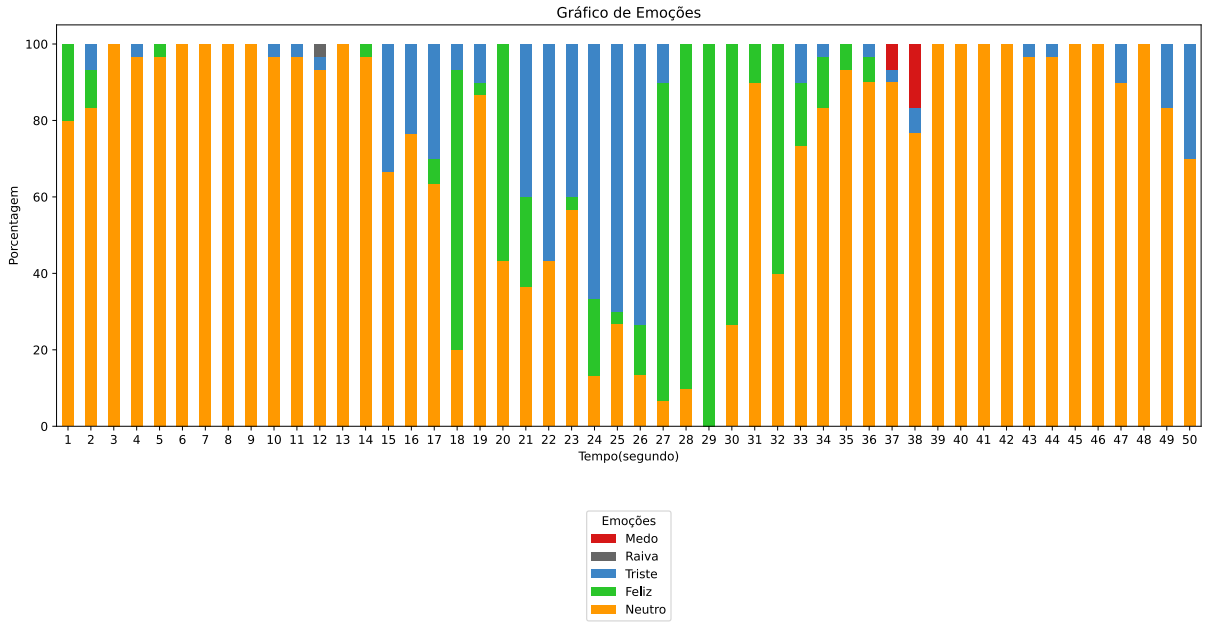
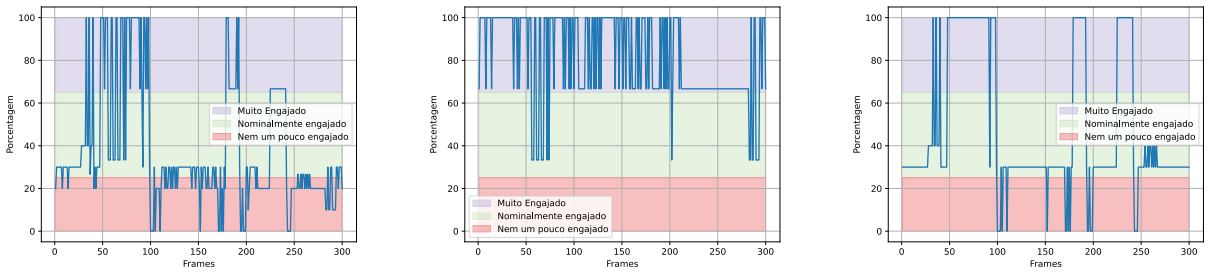


Figura 5.3: Emoções por segundo do Participante 3 no cenário 1.



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.

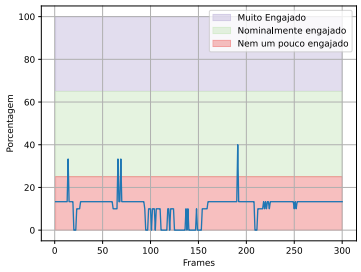
(b) Índice de concentração baseado apenas na predição da expressão facial.

(c) Índice de concentração baseado apenas nos movimentos dos olhos.

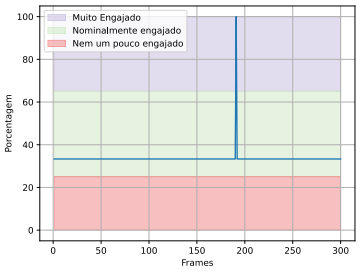
Figura 5.4: Índice de concentração do Participante 1 em diferentes abordagens no cenário 1.

1, considerando parâmetros como a expressão facial (PE) e o peso dos olhos (PO) citados anteriormente, além de uma análise individual de cada um desses elementos (IC_E e IC_O). Cabe destacar que o vídeo foi subdividido em segmentos de 10 segundos, especialmente nos momentos de mudanças significativas, pontos cruciais de fala ou foco, a fim de proporcionar uma análise mais detalhada nesses momentos específicos. Vale ressaltar que quando o IC atinge o valor 0, então não foi possível identificar a região dos olhos e a classe predominante foi **Não identificado**.

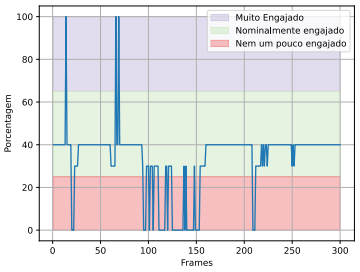
A Figura 5.4 (a) mostra a evolução do IC do Participante 1 baseado na expressão facial e no movimento dos olhos. Entre os *frames* 0 e 100 é possível perceber que houve variações entre as classes, principalmente uma grande concentração na classe **Muito en-**



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.5: Índice de concentração do Participante 2 em diferentes abordagens no cenário 1.

gajado. Entre 100 e 175 o participante obteve um baixo nível de engajamento. A partir de uma análise qualitativa, observou-se no vídeo que nesse intervalo o participante direciona o seu olhar para baixo não sendo possível detectar a região dos olhos. Após esse período ocorreu novamente variações entre as classes, agora com grande concentração na classe **Nem um pouco engajado**. Fazendo uma análise qualitativa do vídeo nesse intervalo é possível perceber que o participante está com um olhar posicionado para lateral de sua câmera, o que pode dificultar a captura da região dos olhos.

A Figura 5.4 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. É possível perceber uma grande concentração na classe **Muito engajado** entre os *frames*. Em certos intervalos é possível verificar que houve uma mudança para a classe **Nominalmente engajado**. Vale ressaltar que no gráfico de emoções o participante atingiu as emoções **Neutro** e **Feliz** com pesos 0,9 e 0,6 respectivamente. Esse pesos influenciam no cálculo do IC baseado na Equação (4.2).

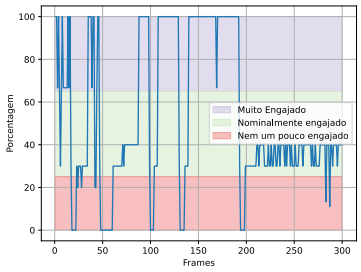
A Figura 5.4 (c) ilustra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Nota-se que houve intervalos onde a classe **Muito engajado** predominou, como por exemplo entre 49 e 90. Porém, em certos *frames* não foi possível detectar a região dos olhos, então o valor de PO foi igual a 0. Esses padrões sugerem uma dinâmica interessante na influência dos movimentos dos olhos na avaliação da concentração ao longo do tempo. Em dados momentos do vídeo, observou-se que o participante estava com olhar direcionado para baixo ou para o lado da câmera, isso pode ter influenciado nos momentos que a concentração dele estar zerada.

Na Figura 5.5 é possível analisar o IC do Participante 2. A Figura 5.5 (a) mostra a evolução da concentração baseada na expressão facial e no movimento dos olhos. É possível perceber que em nenhum momento o IC atingiu as classes **Muito engajado**. O participante durante os *frames* apresentou um baixo nível de concentração devido a grande concentração na classe **Nem um pouco engajado**. É possível perceber também que em alguns *frames* o PO foi igual a 0. Isso pode ser explicado uma vez que sua câmera estava direcionado ao lado do seu rosto, o que dificultou a captura dos olhos.

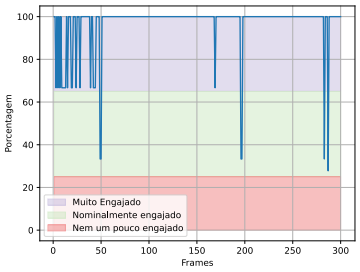
A Figura 5.5 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. É possível ver que houve uma grande concentração na classe **Nominalmente engajado** com o valor de PE igual a 0,3. Exceto no *frame* 191, onde a classe dominante foi **Muito engajado**. Vale lembrar que nesse intervalo o participante apresentou a emoção triste na maior parte do tempo e isso significa o PE foi definido como 0,3. E para o caso em que o resultado do IC_E foi igual a 1, então a emoção neutro foi a predominante para esse *frame*.

A Figura 5.5 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Diferente dos resultados do Participante 1, houve uma grande concentração entre as classes **Nominalmente engajado** e **Não identificado** (onde PO é igual a 0). Para os *frames* 14, 66 e 69 a classe predominante foi a **Muito engajado**. Quando a classe predominante foi **Nominalmente engajado**, os valores de IC_O variaram entre 0,3 e 0,4. Para o caso de IC_O igual a 0,3 significa que o PO foi definido como 1,5, indicando que o participante estava com os olhos semi-abertos. Para o caso de IC_O igual a 0,4, o valor de RP ficou maior que 0,3 e RO estava fora do intervalo 1 e 1,7, por isso PO foi definido como 2. Uma vez que esses dois valores estão entre 0,25 e 0,65 então a classe atingida foi **Nominalmente engajado**.

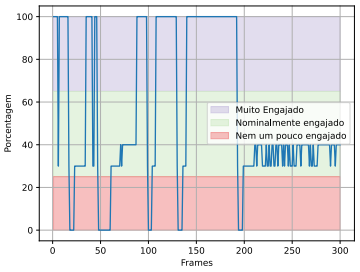
Na Figura 5.6 observa-se o IC do Participante 3. A Figura 5.6 (a) mostra a evolução da concentração baseada na expressão facial e no movimento dos olhos. É possível perceber que o participante teve grande concentração na classe **Muito engajado**. Porém, a partir do *frame* 200, o participante atingiu uma grande concentração na classe **Nominalmente engajado**. Nesse intervalo de tempo é possível perceber que apesar deste participante estar com a câmera posicionada no centro de seu rosto, ele desvia o



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.6: Índice de concentração do Participante 3 em diferentes abordagens no cenário 1.

olhar para o lado, então no momento de calcular o **RO** os *pixels* brancos na metade esquerda e direita não estão centralizados, isso significa que o olhar não está totalmente voltado para a câmera.

A Figura 5.6 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. É possível perceber que esse participante atingiu a classe **Muito engajado** na maior parte dos *frames*. Vale destacar que a emoção que teve predominância nesse intervalo foi a emoção neutro e isso significa que PE foi definido como 0,9.

Por fim, a Figura 5.6 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. É possível notar uma variação entre as classes **Muito engajado** e **Nem um pouco engajado**. Isso se deve ao fato de que em alguns *frames* o olho do participante se mostrou fechado, dificultando a captura da região dos olhos. Além disso, a partir do *frame* 200 a participante tem grande concentração na classe **Nominalmente engajado**. Fazendo uma análise qualitativa nesse intervalo de tempo é possível perceber que o participante direciona seu olhar para o lado. Por fim, Comparado ao Participante 2 ele obteve maior concentração e isso se deve ao fato de que esse participante utilizou a própria câmera embutida em seu computador, facilitando a captura da região dos olhos e contribuindo para um nível de concentração maior.

5.3 Cenário 2: concentração e atenção

Este momento específico captura a fase em que os participantes estão dedicados a ouvir as experiências e resultados da *Sprint* de outros membros da equipe. Este cenário destaca-se como um período de concentração e atenção, proporcionando um contexto valioso para avaliar as expressões faciais determinadas pelo classificador.

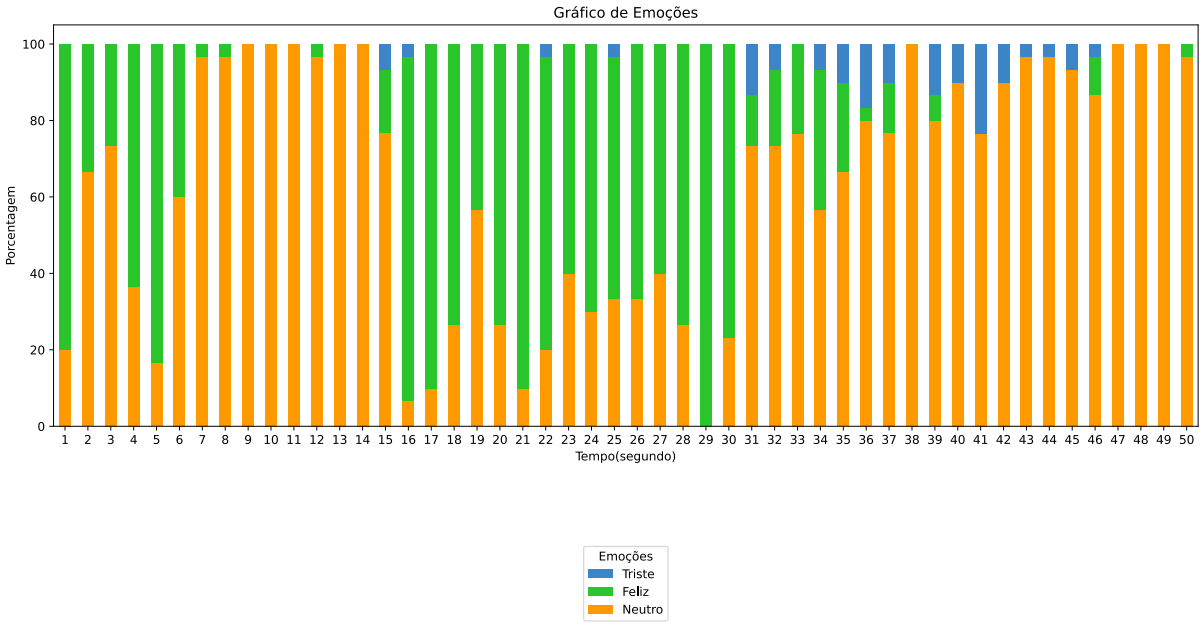


Figura 5.7: Emoções por segundo do Participante 1 no cenário 2.

Ao analisar a Figura 5.7, é possível observar as emoções predominantes em cada segundo do vídeo do Participante 1. As emoções **Neutro** e **Feliz** destacam-se durante a maior parte da gravação. Entre o segundo 7 e o 15, o Participante 1 mantém uma expressão totalmente neutra, alterando-se para uma emoção **Feliz**, a partir do segundo 16, como evidenciado no vídeo. A partir do segundo 31 até o final do vídeo, a emoção volta a ser **Neutro** e permanece constante. Vale ressaltar que, embora a emoção **Triste** seja capturada em alguns momentos, não é predominante na narrativa emocional.

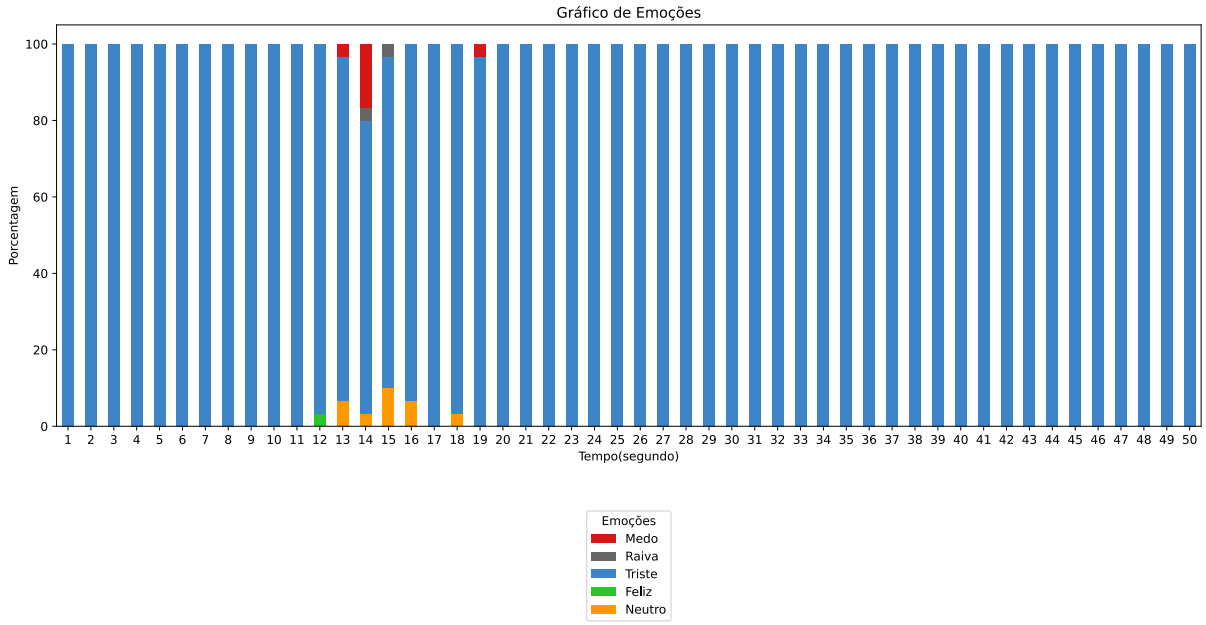


Figura 5.8: Emoções por segundo do Participante 2 no cenário 2.

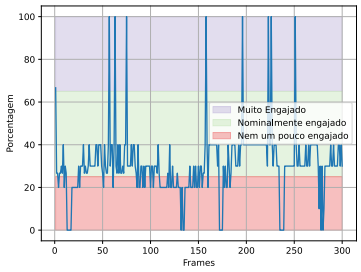
No caso do Participante 2 que pode ser visto na Figura 5.8, a análise revela que a emoção predominante ao longo de todo o período foi **Triste**, notadamente no cenário atual que requer foco e atenção ao ouvir os demais participantes. Cabe ressaltar que, embora não sejam as emoções predominantes, também são identificadas expressões de **Medo**, **Raiva**, **Feliz** e **Neutro** em alguns momentos.

Tabela 5.2: Porcentagem do gênero encontrado em cada *frame* dos participantes 1 e 2 no cenário 2.

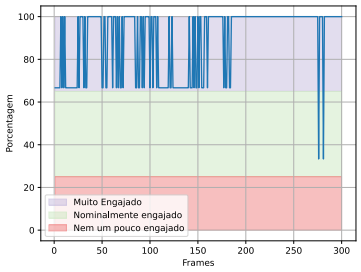
Participante	Mulher	Homem
1	**70,35%**	29,65%
2	35,96%	**64,04%**

A Tabela 5.2 apresenta os resultados da classificação de gênero. Conforme previsto, o Participante 1 obteve a maior porcentagem para o gênero feminino, totalizando 70,35%. Já o Participante 2 foi predominantemente classificado como homem, alcançando a porcentagem de 64,04%. Vale ressaltar que comparado ao cenário 1, o Participante 2 obteve um valor de erro maior.

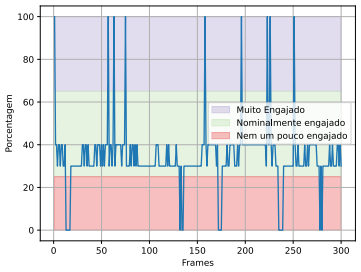
A Figura 5.9 (a) mostra a evolução da concentração do Participante 1 baseado na expressão facial e no movimento dos olhos. É possível perceber que o participante obteve grande predominância na classe **Nominalmente engajado**. Além disso, em alguns momentos o valor de *PO* foi igual a zero, nesses *frames* é possível ver o olhar



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

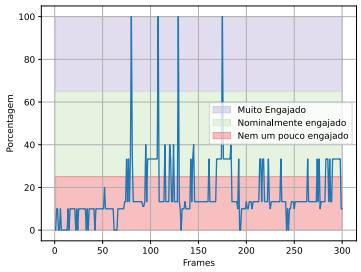
Figura 5.9: Índice de concentração do Participante 1 em diferentes abordagens no cenário 2.

voltado para baixo dificultando a captura da região dos olhos. Vale ressaltar também que a classe **Muito engajado** foi atingida em alguns *frames*.

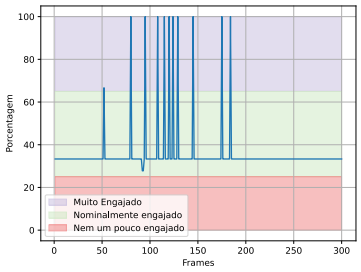
A Figura 5.9 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Observa-se que houve uma grande concentração na classe **Muito engajado**. Vale notar que no gráfico de emoções nesse cenário para o Participante 1, as emoções predominantes nesses segundos foram **Feliz** e **Neutro** com peso 0,6 e 0,9 respectivamente. Vale destacar também que esse cenário esperava dos participantes uma grande foco diante da câmera.

A Figura 5.9 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Observa-se que ocorreram variações entre as classes, porém a classe com maior concentração para este cenário foi **Nominalmente engajado**. Isso pode ocorrer devido ao fato de que a câmera do Participante 1 estava direcionada acima da região dos olhos e isso dificulta a captura da região dos olhos, apesar dele estar olhando diretamente para câmera.

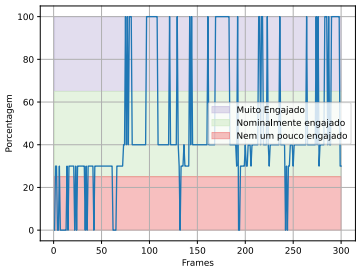
A Figura 5.10 (a) mostra a evolução da concentração do Participante 2 baseado na expressão facial e no movimento dos olhos. Percebe-se que a classe que teve maior concentração foi **Nem um pouco engajado**. O participante também apresenta em alguns *frames* variação entre as classes **Nominalmente engajado** ou **Muito engajado**. Em comparação com o Participante 1 neste cenário, o Participante 2 apresentou um baixo nível de engajamento onde ele pode ter sido influenciado pela emoção. Vale destacar que o Participante 1 obteve predominância na emoção neutro enquanto o Participante 2 foi



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.10: Índice de concentração do Participante 2 em diferentes abordagens no cenário 2.

classificado predominantemente na emoção triste.

A Figura 5.10 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Observa-se que a classe com maior concentração foi a classe **Nominalmente engajado**, onde o IC_E teve um valor igual a 0,3. Isso quer dizer que o valor de **PE** foi 0,3 que é referente ao peso da emoção triste. Vale lembrar que nesse cenário a emoção que predominou durante a maior parte do tempo foi triste. A classe **Muito engajado** foi atingida em alguns *frames*, isso se deve ao fato do participante ter atingido a classe neutro como predominante. Em comparação com o Participante 1 neste cenário, o Participante 2 obteve um baixo nível de engajamento.

A Figura 5.10 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. É possível analisar que houve algumas alterações em que o participante esteve por um longo momento **Muito engajado** ou **Nem um pouco engajado**. Em alguns momentos a classe **Nem um pouco engajado** foi atingida e também tiveram algumas ocorrências em que o valor do IC foi 0. Em comparação com o Participante 1, este participante obteve um nível de engajamento melhor, já que atingiu mais vezes a predominância na classe **Muito engajado**. Vale ressaltar que a concentração considerando apenas os olhos ou apenas as emoções apresentaram variações significativas.

5.4 Cenário 3: momento de reflexão sem interação com a câmera

Este cenário captura a fase em que os participantes estão engajados na reflexão e na escrita sobre o *feedback* da *Sprint*. Durante esse momento, eles expressam suas considerações sobre aspectos positivos e negativos, para que posteriormente, no cenário 4, compartilhem suas análises com a equipe. O objetivo deste cenário é proporcionar ao experimento uma análise em cima deste contexto, no qual os participantes não estão interagindo diretamente com a câmera.

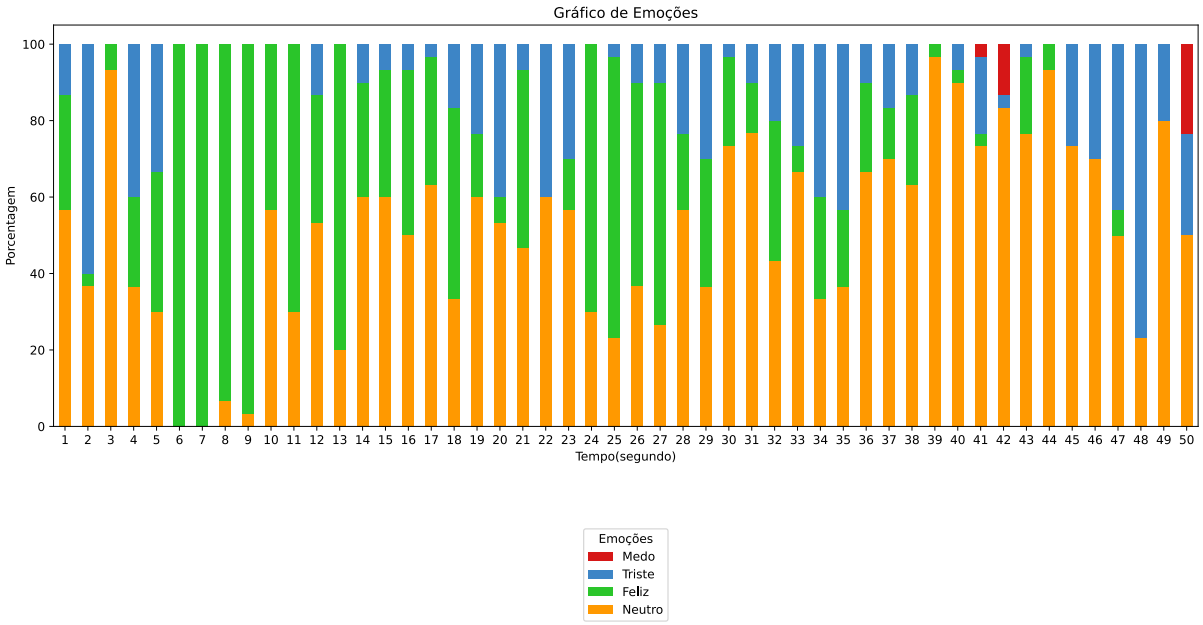


Figura 5.11: Emoções por segundo do Participante 1 no cenário 3.

Na Figura 5.11, em relação ao Participante 1, é possível analisar que a maior parte do tempo a emoção **Neutro** predominou, além disso houve uma pequena mudança para emoção **Feliz** entre os segundos 6 e 0 e entre 24 e 27. Além disso, a emoção **Triste** só teve predominância no segundo 48, mas foi possível captá-la em alguns outros momentos. Vale destacar que a emoção medo, mesmo que sem predominância, apareceu nos segundos 41, 42 e 50.

Para o caso do Participante 2, é possível analisar na Figura 5.12 que a emoção que teve maior predominância foi **Triste**, sendo capturada na maior parte do tempo, exceto no segundo 40 em que a emoção **Raiva** ganhou destaque. Além destas, as emoções

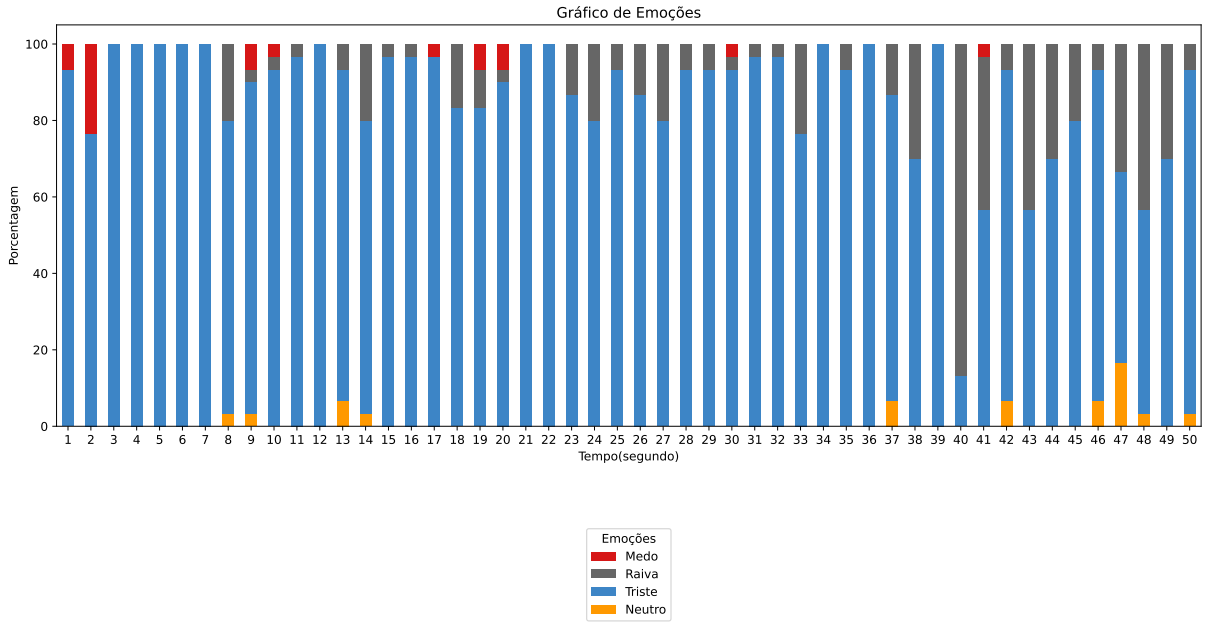


Figura 5.12: Emoções por segundo do Participante 2 no cenário 3.

Neutro e **Medo** também foram capturadas em alguns momentos.

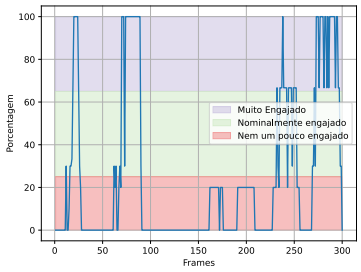
Tabela 5.3: Porcentagem do gênero encontrado em cada *frame* dos participantes 1 e 2 no cenário 3.

Participante	Mulher	Homem
1	**77,81%**	22,19%
2	3,47%	**96,53%**

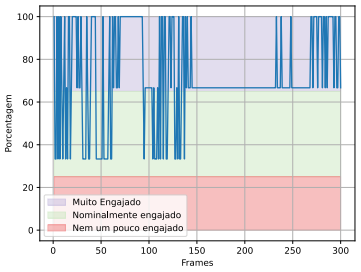
A Tabela 5.3 apresenta os resultados da classificação de gênero. Conforme previsto, o Participante 1 obteve a maior porcentagem para o gênero feminino, totalizando 77,81%. Já o Participante 2 foi predominantemente classificado como homem, alcançando a porcentagem de 96,53%.

A Figura 5.13 (a) mostra a evolução da concentração do Participante 1 baseado na expressão facial e no movimento dos olhos. O participante apresenta uma baixo nível de concentração uma vez que atingiu o IC igual a 0. Fazendo uma análise qualitativa do vídeo, é possível perceber que o participante estava com olhar baixo e envolvido em atividades como digitar no teclado, isso dificulta a captura dos olhos. Apesar disso, em alguns intervalos há uma concentração na classe **Muito engajado**, em momentos em que o olhar está voltado para câmera.

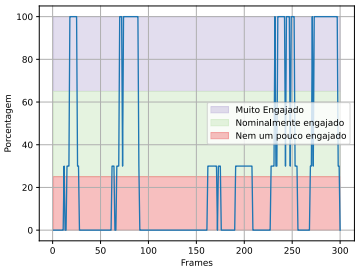
A Figura 5.13 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Observa-se que houve bastante concentração na classe **Muito**



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



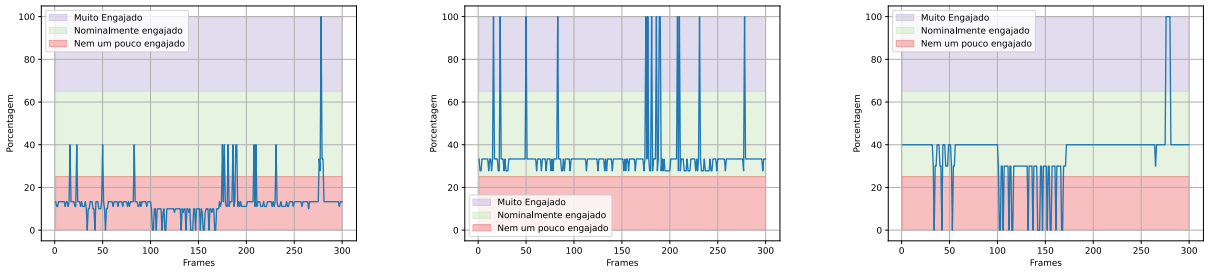
(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.13: Índice de concentração do Participante 1 em diferentes abordagens no cenário 3.

engajado. Entretanto, entre os *frames* 0 e 190 houve bastante variação com a classe **Nominalmente engajado**. Vale lembrar que nesse cenário o Participante 1 obteve predominância com as emoções **Neutro** e **Feliz**.

A Figura 5.13 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Nota-se que houve algumas variações entre as classes. Ao analisar o vídeo nesse cenário, é possível observar que o participante mantém o olhar baixo e, em alguns momentos, em outra direção, além de estar envolvido em atividades como digitar no teclado. Esses comportamentos podem estar relacionados ao fato de que o classificador teve dificuldade em capturar adequadamente a região dos olhos, o que resultou na identificação de que o participante não estava concentrado. Esse fato se deve aos *frames* com IC_O igual a zero. Contudo, também houve variações com as classes **Nominalmente engajado** e **Muito engajado**, no momento em que o olhar do participante estava voltado para câmera.

A Figura 5.14 (a) mostra a evolução da concentração do Participante 2 baseado na expressão facial e no movimento dos olhos. Observa-se que somente no *frame* 279 o valor de IC alcançou a classe **Muito Engajado**. O participante atingiu, em alguns momentos, a classe **Nominalmente Engajado**, mas, predominantemente, manteve-se na classe **Nem um pouco Engajado**. Vale ressaltar que em alguns momentos o **IC** foi igual a 0. Em comparação com o Participante 1, observa-se uma variação mais expressiva nos valores, situando-se entre 0 e 10%. Isso pode ser atribuído à orientação do posicionamento da câmera, o que dificulta a captura dos olhos, e também à prevalência das



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.

(b) Índice de concentração baseado apenas na predição da expressão facial.

(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.14: Índice de concentração do Participante 2 em diferentes abordagens no cenário 3. (a) expressão facial e movimentos dos olhos, (b) somente expressão facial e (c) somente movimento dos olhos.

emoções presentes nos *frames*. No caso, observou-se que o Participante 1 apresentou uma maior predominância na emoção feliz ou neutra, enquanto o Participante 2 destacou-se na expressão de tristeza.

A Figura 5.14 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Destaca-se a obtenção da classe **Muito Engajado**, juntamente com grande concentração na classe **Nominalmente engajado**. É importante destacar também que ao longo desse cenário, o Participante 2 expressou predominantemente a emoção de tristeza.

A Figura 5.14 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Com isso é possível notar uma grande concentração das classes **Nominalmente engajado** e **Nem um pouco engajado** atingindo o valor de IC_O igual a 0. Analisando o vídeo, foi possível perceber que o olhos do participante estava direcionado para cima ou para baixo, o que se reflete no momento de captura da região dos olhos para o cálculo do IC_O.

5.5 Cenário 4: discussão e interação ativa com a câmera

Como nos cenários 2 e 3, somente os Participantes 1 e 2 foram incluídos neste cenário. Neste cenário os participantes compartilham os pontos positivos e negativos da *Sprint*, os quais foram escritos no Cenário 3. Este momento marca a interação ativa com a câmera, uma vez que os participantes se envolvem em discussões mais dinâmicas com o restante

do time.

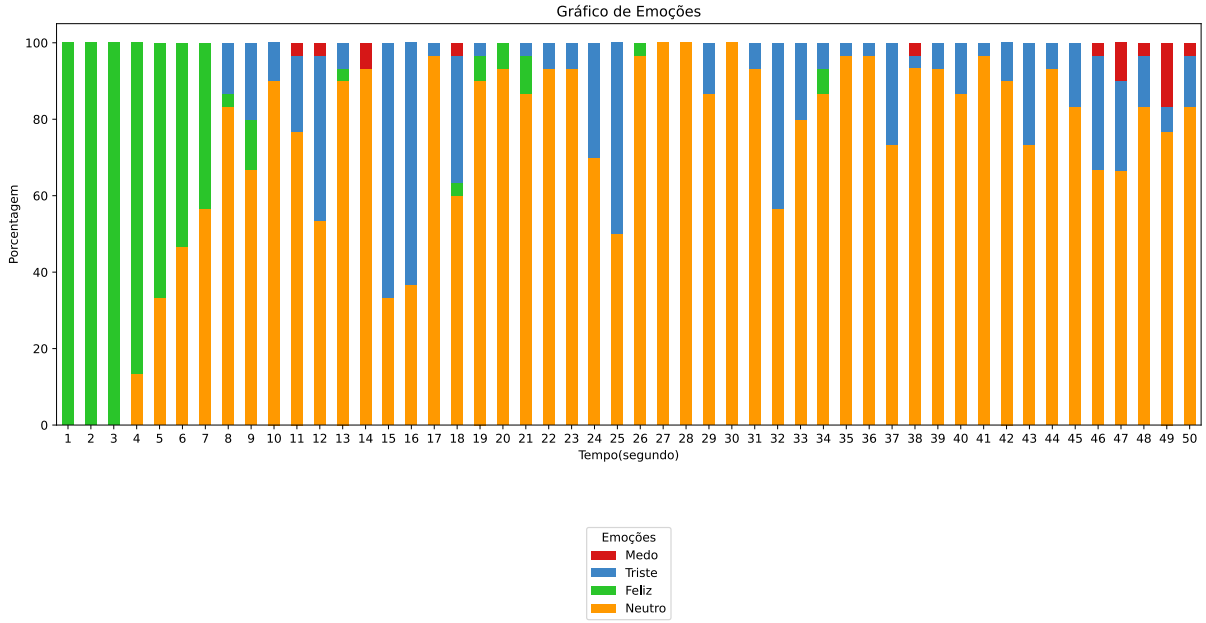


Figura 5.15: Emoções por segundo do Participante 1 no cenário 4.

Na Figura 5.15, em relação o Participante 1, é possível analisar que na primeira parte do tempo até o segundo 7, a emoção **Feliz** se destaca, e após isso, na maior parte do tempo a emoção **Neutro** ganha predominância, exceto no segundo 15 e 16 onde a emoção **Triste** ganha destaque. Além destas, é possível ver que a emoção **Medo** também foi capturada apesar de não ter predominância em nenhum momento.

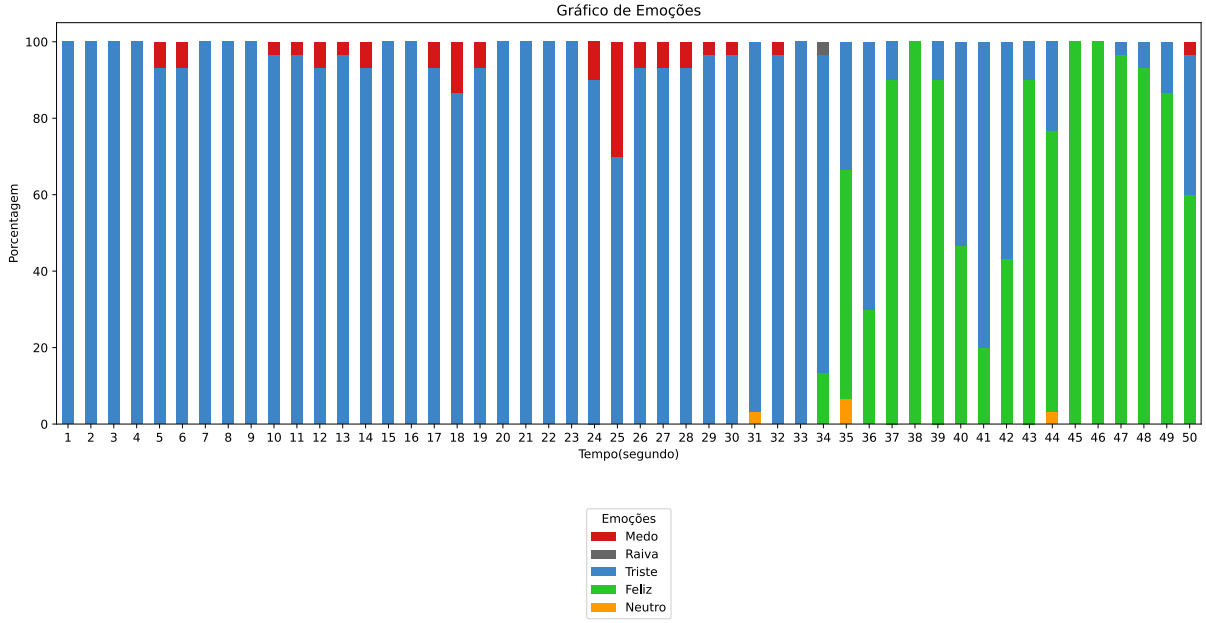
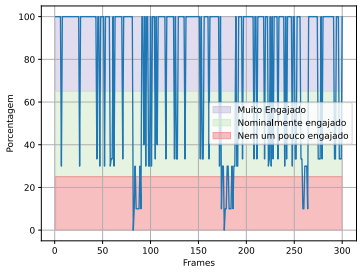
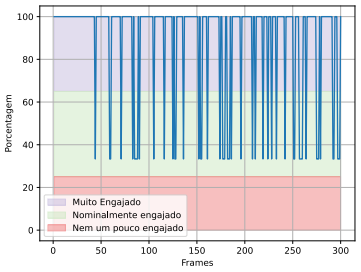


Figura 5.16: Emoções por segundo do Participante 2 no cenário 4 .

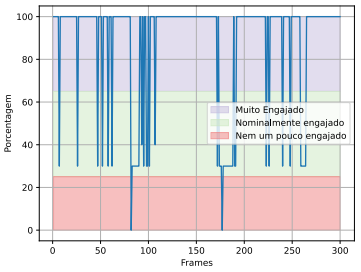
Em relação ao Participante 2, na Figura 5.16 é possível analisar que na primeira



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.17: Índice de concentração do Participante 1 em diferentes abordagens no cenário 4.

parte do tempo até o segundo 34 a emoção **Triste** ganhou maior destaque em relação as outras, porém após isso a emoção **Feliz** ganha predominância em alguns momentos finais do vídeo. Além disso, as emoções **Raiva**, **Medo** e **Neutro** também são captadas em alguns momentos sem predominância.

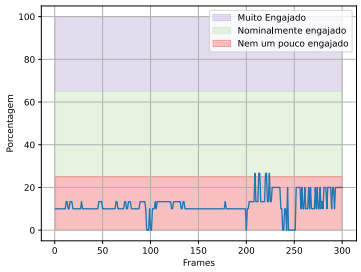
Tabela 5.4: Porcentagem do gênero encontrado em cada *frame* dos participantes 1 e 2 no cenário 3.

Participante	Mulher	Homem
1	**84%**	16%
2	**65,56%**	34,44%

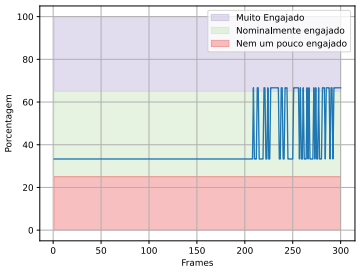
A Tabela 5.4 apresenta os resultados da classificação de gênero. Conforme previsto, o Participante 1 obteve a maior porcentagem para o gênero feminino, totalizando 77,81%. Já o Participante 2 foi predominantemente classificado como mulher na maior parte do tempo, alcançando a porcentagem de 65,56% para este gênero. Então neste cenário foi possível analisar que o classificador errou, na maior parte do tempo, o gênero do Participante 2.

A Figura 5.17 (a) mostra a evolução da concentração do Participante 1 baseado na expressão facial e no movimento dos olhos. É possível perceber que houve bastante concentração na classe **Muito Engajado**. Além disso, somente nos *frames* 83 e 178 o valor de **IC** foi igual a zero.

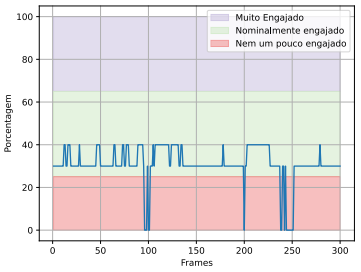
A Figura 5.17 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Nota-se que houve grande variação entre as classes **Muito engajado** e **Nominalmente engajado**. Vale ressaltar que nesse cenário o participante



(a) Índice de concentração baseado na expressão facial e no movimento dos olhos.



(b) Índice de concentração baseado apenas na predição da expressão facial.



(c) Índice de concentração baseado apenas nos movimentos dos olhos.

Figura 5.18: Índice de concentração do Participante 2 em diferentes abordagens no cenário 4.

obteve na maior parte do tempo predominância na emoção **Neutro** que corresponde ao peso 0,9, mas a emoção **Triste** com peso 0,3 também foi atingida em alguns *frames*.

A Figura 5.17 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Observa-se que houve grande concentração na classe **Muito engajado** e variações com a classe **Nominalmente engajado**. Apenas nos *frames* 83 e 178 o valor de **IC** foi igual a zero. Fazendo uma análise qualitativa nesse intervalo, o Participante 1 apresentou o olhar direcionado ao lado de sua câmera, que dificulta a captura dos olhos.

A Figura 5.18 (a) mostra a evolução da concentração do Participante 2 baseado na expressão facial e no movimento dos olhos. É possível perceber que em nenhum momento o IC atingiu a classe **Muito Engajado**. O participante apresenta grande predominância na classe **Nem um pouco engajado**. Porém, no intervalo de 210 até 225 ocorre uma pequena variação com a classe **Nominalmente engajado**. Comparado ao Participante 1 nesse cenário, o Participante 2 apresentou um baixo nível de concentração.

A Figura 5.18 (b) mostra a evolução da concentração do participante baseado apenas na expressão facial. Há uma grande concentração na classe **Nominalmente engajado**. A partir do *frame* 210, o valor de IC_E apresenta algumas variações com a classe **Muito engajado**. Vale destacar que nesses intervalos o participante atingiu a emoção **Triste** com peso 0,3 e **Feliz** com peso 0,6.

A Figura 5.18 (c) mostra a evolução da concentração do participante baseado apenas nos movimentos dos olhos. Houve algumas variações em que o participante esteve

em alguns momentos **Nominalmente engajado** ou **Nem um pouco engajado**. Ao analisar o vídeo é possível perceber que o participante está conversando porém a sua câmera está posicionada na lateral dos seus olhos o que dificulta a captura dos olhos.

5.6 Discussão dos resultados

Ao examinar os gráficos de emoções, observa-se que, em cenários que envolviam interação direta com a câmera, como nos casos dos Cenários 1 e 4, onde os participantes precisaram se comunicar, as emoções predominantes foram, na maior parte do tempo, **Feliz**, **Triste** ou **Neutro**. Em contrapartida, quando o contexto mudou para momentos de introspecção, destacaram-se as emoções **Neutro** e **Triste**.

Avaliando o desempenho do classificador de gênero, observa-se sua eficácia na identificação correta do gênero dos participantes na maioria dos cenários. No entanto, é importante ressaltar que ocorreu uma exceção no Cenário 4, onde o classificador apresentou um equívoco ao identificar o gênero do Participante 2. Isso ressalta a importância de considerar as nuances e desafios na classificação de gênero, reconhecendo que, embora o classificador tenha acertado na maioria dos casos, ainda pode apresentar erros significativos.

Os resultados do Participante 1 revelaram que, na maioria dos *frames*, as classes **Muito engajado** e **Nominalmente engajado** foram predominantes. É relevante destacar que, no Cenário 1, onde houve interação direta com a câmera, o participante atingiu o estado **Muito engajado** quando o IC_E foi baseado na predição da expressão facial. Apesar da câmera estar posicionada um pouco acima dos olhos, o Participante 1 obteve níveis de concentração mais altos em comparação ao Participante 2.

Em contrapartida, o Participante 2 apresentou predominantemente um baixo nível de engajamento na maior parte dos cenários. Essa observação pode ser atribuída à posição lateral da câmera, comprometendo a captura eficaz da região dos olhos. Vale destacar que na maior parte dos cenários o participante obteve predominância na emoção triste que possui o peso de 0,3. Essa condição influenciou negativamente o seu desempenho em cenários que demandavam comunicação direta com a câmera, como ocorreu no Cenário 1.

O Participante 3 também demonstrou um bom nível de engajamento, com a classe **Muito engajado** predominando na maioria dos *frames*. Além disso quando o valor de IC_E se baseou na expressão facial o participante atingiu grande concentração na classe **Muito engajado**. Vale ressaltar que a Participante 3 utilizou a câmera embutida em seu computador, facilitando a captura da região dos olhos e contribuindo para o seu desempenho positivo.

Observou-se também que situações em que o usuário não interage diretamente com a câmera podem resultar em níveis de concentração mais baixos devido à limitação do peso atribuído aos olhos. Isso pode ser atribuído ao posicionamento da câmera, pois a pessoa pode estar olhando para a tela, mas não diretamente para a câmera, caso a câmera esteja em uma posição lateral. Dessa forma, essa metodologia para estimativa do índice de concentração pode não ser adequada para avaliar determinadas situações.

6 Conclusão

O presente trabalho destacou uma análise abrangente de um classificador de expressões faciais, gênero e nível de engajamento, ampliando sua aplicabilidade para contextos que ultrapassam os ambientes educacionais *online*. A implementação desse classificador ofereceu benefícios notáveis em termos de supervisão e controle, proporcionando uma compreensão mais profunda do engajamento dos participantes do outro lado da tela. Essa abordagem visa mitigar a desconexão que pode ocorrer nesses ambientes, ao mesmo tempo em que permite a coleta de informações valiosas sobre emoções e gênero.

Por meio dos diversos experimentos conduzidos, observou-se a habilidade do classificador em identificar não apenas as expressões faciais, mas também o movimento dos olhos. Este estudo envolveu a coleta de um conjunto de 9 vídeos, cada um com 50 segundos de duração, nos quais foram criados quatro cenários relevantes para análise. As avaliações realizadas abrangeram a classificação de emoção e gênero, indo além ao considerar não apenas o nível de engajamento, mas também outros dois fatores: o movimento dos olhos e as emoções.

Além disso, a capacidade do classificador em avaliar o nível de engajamento revelou-se uma ferramenta valiosa. Em um cenário de ensino remoto, onde manter a atenção dos alunos pode ser desafiador, a capacidade de medir e compreender o engajamento oferece oportunidades significativas para ajustar as estratégias pedagógicas e melhorar a eficácia do ensino *online*. Apesar disso, é importante ressaltar que o classificador de engajamento ainda apresenta algumas limitações na detecção dos olhos, especialmente em situações em que não há interação direta com a câmera ou se a câmera não estiver centralizada.

Como um benefício adicional, o classificador também demonstrou sua aptidão na classificação de gênero. Essa funcionalidade não apenas amplia a compreensão dos participantes, mas também adiciona uma camada adicional de personalização, permitindo adaptações específicas com base nas características individuais. É importante ressaltar que, apesar dos acertos, o classificador de gênero ainda apresenta erros significativos.

Como análise futura, seria possível realizar uma revisão na literatura com o intuito de aprimorar o modelo atribuído a esse classificador.

Como trabalhos futuros, há algumas áreas importantes a serem exploradas para o classificador. Primeiramente, a melhoria contínua da acurácia do classificador de gênero pode envolver a exploração de técnicas mais avançadas de aprendizado de máquina, o refinamento dos conjuntos de dados utilizados no treinamento e a consideração de características adicionais para aprimorar a precisão nas predições de gênero.

Em segundo lugar, é relevante explorar a possibilidade de testar o classificador utilizando outros modelos de aprendizado de máquina. A experimentação com diferentes arquiteturas e abordagens pode oferecer *insights* valiosos sobre como otimizar ainda mais o desempenho do classificador, proporcionando uma base sólida para futuras iterações e aprimoramentos. Além disso, é relevante explorar outras abordagens para o cálculo do nível de engajamento, a fim de aprimorar o desempenho do classificador e superar as limitações encontradas.

Por último a integração efetiva do classificador em plataformas *online* já existentes. Ao incorporar essa tecnologia em ambientes educacionais virtuais, por exemplo, os educadores poderiam colher benefícios imediatos ao monitorar o engajamento dos alunos e compreender suas expressões faciais, contribuindo para uma experiência de aprendizado mais personalizada e eficaz.

Ainda assim, apesar dos trabalhos futuros recomendados para o aprimoramento do classificador, o mesmo se apresentou muito eficaz dentro da sua proposta, podendo ser inclusive, facilmente usado em outros segmentos, de acordo com as suas necessidades.

Bibliografia

ALVES, G. *Entendendo Redes Convolucionais (CNNs)*. 2018. Acesso em: 03/12/2023. Disponível em: ⟨https://medium.com/neuronio-br/entendendo-redes-convolucionais-cnns-d10359f21184⟩.

ARRIAGA, O. *Face classification and detection*. 2020. Acesso em: 08/12/2023. Disponível em: ⟨https://github.com/oarriaga/face_classification⟩.

ARRIAGA, O.; VALDENEGRO-TORO, M.; PLÖGER, P. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017.

BARSOUM, E.; ZHANG, C.; FERRER, C. C.; ZHANG, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *Proceedings of the 18th ACM international conference on multimodal interaction*. [S.l.: s.n.], 2016. p. 279–283.

BOUHLAL, M.; AARIKA, K.; ABDELOUAHID, R. A.; ELFILALI, S.; BENLAHMAR, E. Emotions recognition as innovative tool for improving students' performance and learning approaches. *Procedia Computer Science*, Elsevier, v. 175, p. 597–602, 2020.

BYUN, A. *Convolutional Neural Networks (CNNs / ConvNets)*. 2023. Acesso em: 03/12/2023. Disponível em: ⟨https://cs231n.github.io/convolutional-networks/⟩.

EKMAN, P.; OSTER, H. Facial expressions of emotion. *Annual review of psychology*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 30, n. 1, p. 527–554, 1979.

GOODFELLOW, I. J.; ERHAN, D.; CARRIER, P. L.; COURVILLE, A.; MIRZA, M.; HAMNER, B.; CUKIERSKI, W.; TANG, Y.; THALER, D.; LEE, D.-H. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. [S.l.], 2013. p. 117–124.

GUPTA, S.; KUMAR, P.; TEKCHANDANI, R. K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications*, Springer, v. 82, n. 8, p. 11365–11394, 2023.

KIM, K. G. Book review: Deep learning. *Healthcare informatics research*, Korean Society of Medical Informatics, v. 22, n. 4, p. 351–354, 2016.

KING, D. *Real-Time Face Pose Estimation*. 2014. Acesso em: 03/12/2023. Disponível em: ⟨https://blog.dlib.net/2014/08/real-time-face-pose-estimation.html⟩.

KIURU, N.; SPINATH, B.; CLEM, A.-L.; EKLUND, K.; AHONEN, T.; HIRVONEN, R. The dynamics of motivation, emotion, and task performance in simulated achievement situations. *Learning and Individual Differences*, Elsevier, v. 80, p. 101873, 2020.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017.

KWONG, J. C. T.; GARCIA, F. C. C.; ABU, P. A. R.; REYES, R. S. Emotion recognition via facial expression: utilization of numerous feature descriptors in different machine learning algorithms. In: IEEE. *TENCON 2018-2018 IEEE Region 10 Conference*. [S.l.], 2018. p. 2045–2049.

LIU, W.; ANGUELOV, D.; ERHAN, D.; SZEGEDY, C.; REED, S.; FU, C.-Y.; BERG, A. C. Ssd: Single shot multibox detector. In: SPRINGER. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. [S.l.], 2016. p. 21–37.

MARTINEZ, J. Take this pandemic moment to improve education. *Edu Source*, 2020.

MEHENDALE, N. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, Springer, v. 2, n. 3, p. 446, 2020.

MINAEE, S.; MINAEI, M.; ABDOLRASHIDI, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, mdpi, v. 21, n. 9, p. 3046, 2021.

MISHRA, L.; GUPTA, T.; SHREE, A. Online teaching-learning in higher education during lockdown period of covid-19 pandemic. *International Journal of Educational Research Open*, Elsevier, v. 1, p. 100012, 2020.

MISHRA, M. *Convolutional Neural Networks, Explained*. 2020. Acesso em: 04/12/2023. Disponível em: ⟨https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939⟩.

NOURAEY, P.; BAVALI, M.; BEHJAT, F. A post-pandemic systematic review of e-learning: A cross-cultural study. *International Journal of Society, Culture & Language*, Katibeh-ILCRG, p. 1–18, 2023.

PERES, M. d. L. L. *Aprenda a Criar e Treinar Uma Rede Neural Convolucional (CNN)*. 2021. Acesso em: 03/12/2023. Disponível em: ⟨https://www.insightlab.ufc.br/aprenda-a-criar-e-treinar-uma-rede-neural-convolucional-cnn/⟩.

ROTHE, R.; TIMOFTE, R.; GOOL, L. V. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, Springer, v. 126, n. 2-4, p. 144–157, 2018.

SHARMA, P.; JOSHI, S.; GAUTAM, S.; MAHARJAN, S.; KHANAL, S. R.; REIS, M. C.; BARROSO, J.; FILIPE, V. M. de J. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In: SPRINGER. *International Conference on Technology and Innovation in Learning, Teaching and Education*. [S.l.], 2022. p. 52–68.

SHEN, J.; YANG, H.; LI, J.; CHENG, Z. Assessing learning engagement based on facial expression recognition in mooc's scenario. *Multimedia Systems*, Springer, p. 1–10, 2022.

TALEGAONKAR, I.; JOSHI, K.; VALUNJ, S.; KOHOK, R.; KULKARNI, A. Real time facial expression recognition using deep learning. In: *Proceedings of international conference on communication and information processing (ICCIP)*. [S.l.: s.n.], 2019.

TIAN, Y.-I.; KANADE, T.; COHN, J. F. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 23, n. 2, p. 97–115, 2001.

TORRES, I. I. *Emotional Needs of Online Students: a Phenomenological Study of Graduate Level, Nontraditional Students*. Tese (Doutorado) — The Chicago School of Professional Psychology, 2020.

WENG, L. Object detection part 4: Fast detection models. lilianweng. github. io/lil-log, 2018. *URL http://lilianweng. github. io/lil-log/2018/12/27/object-detection-part-4. html*, 2018.

YAGOUB, K. A. *Concentration Index Generator*. 2023. Acesso em: 03/12/2023. Disponível em: ⟨https://github.com/CaedenZ/distractionModel⟩.

ZHENG, X.; HASEGAWA, S.; TRAN, M.-T.; OTA, K.; UNOKI, T. Estimation of learners' engagement using face and body features by transfer learning. In: SPRINGER. *Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*. [S.l.], 2021. p. 541–552.

ZOU, Z.; CHEN, K.; SHI, Z.; GUO, Y.; YE, J. Object detection in 20 years: A survey. *Proceedings of the IEEE*, IEEE, 2023.