

# Detectando o envolvimento do aluno em MOOCs usando Reconhecimento Automático de Expressão Facial

Abhilash Dubbaka

Departamento de Computação

Imperial College London

Londres, Reino Unido

abhilash.dabbaka18@imperial.ac.uk

Anandha Gopalan

Departamento de

Computação Imperial

College London Londres,

Reino Unido a.gopalan@imperial.ac.uk

**Resumo**—As taxas de evasão em cursos online abertos e massivos (MOOCs) são muito altas. Embora existam muitos fatores externos, como os usuários não terem tempo suficiente ou não pretenderem concluir o curso, existem alguns aspectos que os instrutores podem controlar para envolver seus alunos de maneira ideal. Para fazer isso, eles precisam saber como os alunos se envolvem durante a videoaula. Este artigo explora o uso de webcams para registrar as expressões faciais dos alunos enquanto eles assistem a material de vídeo educacional para analisar seus níveis de envolvimento do aluno. Redes neurais convolucionais (CNNs) foram treinadas para detectar unidades de ação facial, que foram mapeadas em duas medidas psicológicas, valência (estado emocional) e excitação (atenção), usando regressões de vetores de suporte. Esses valores de valência e excitação foram combinados de uma maneira nova, resultando em níveis de envolvimento do aluno. Além disso, uma nova abordagem foi utilizada para combinar CNNs com técnicas baseadas em características geométricas para melhorar o desempenho dos modelos. Dois experimentos foram conduzidos e descobriram que 9 em cada 10 modelos CNN alcançaram 95% de precisão em média na maioria dos assuntos, enquanto o detector de envolvimento do aluno foi capaz de identificar expressões faciais que se traduziram em níveis de envolvimento do aluno com sucesso. Estes resultados sugerem que esta abordagem é promissora, na medida em que o feedback sobre o envolvimento dos alunos com o curso pode ser fornecido ao instrutor. Pesquisas adicionais devem ser realizadas para comprovar ainda mais esses resultados e superar algumas limitações enfrentadas.

**Palavras-chave**—MOOCs, engajamento, expressões faciais, e-learning

## I. INTRODUÇÃO

A oferta de educação está sendo revolucionada pela rápida evolução da tecnologia. Está a fornecer novos métodos de aprendizagem aos alunos e a ajudar os instrutores, dando-lhes as ferramentas para ensinar de forma mais eficaz e atingir um público mais vasto através da utilização de Cursos Online Abertos e Massivos (MOOCs).

Nos últimos anos, os MOOCs têm desfrutado de um enorme aumento de interesse e a sua utilização cresceu de menos de 20 milhões de utilizadores em 2014 para 110 milhões de utilizadores em 2019 [1], [2]. O número de MOOCs também cresceu, em média, 34% ao ano, de 2015 a 2019. Em particular, os cursos de educação em engenharia consistiam em 6% de MOOCs em 2015 [1], que cresceu para 8% em 2019 [2]. Combinar isso com o crescimento dos MOOCs significa que os MOOCs de engenharia cresceram 42% a cada ano de 2015 a 2019. Esta é a área de curso que mais cresce, com os próximos dois tópicos mais rápidos sendo Negócios e Tecnologia, ambos com 39%, o que enfatiza a importância de MOOCs para engenharia

Educação. Dada a sua ascensão, é importante compreender a sua eficácia e se podem ser considerados como um veículo para a oferta em larga escala de ensino de engenharia.

Um dos principais problemas que os MOOCs enfrentam são as taxas de abandono muito altas [3]. Segundo [4], a taxa média de conclusão é de apenas 15%. Os números de alunos que utilizam a plataforma edX e concluem o curso são ainda mais baixos, com uma taxa média de conclusão pouco menos de 5% nos últimos 5 anos [5]. Há uma infinidade de razões para isso, desde os alunos não terem tempo suficiente até apenas experimentarem o curso. Existem certas etapas que os instrutores MOOC podem seguir para envolver os alunos de maneira ideal, desde alterar a estrutura do curso e o nível de dificuldade até dividir o conteúdo em partes menores para que sejam mais fáceis de gerenciar pelos alunos. No entanto, os instrutores não conseguem receber feedback sobre como seu conteúdo e sua entrega estão sendo percebidos pelo aluno. Num ambiente de sala de aula, isto pode ser feito avaliando uma série de pistas, tais como linguagem corporal, contacto visual, expressão facial, etc. Se os alunos parecerem entediados ou confusos, o estilo de ensino pode ser adaptado para os reengajar [6]. No entanto, num MOOC, os instrutores não conseguem ver os seus alunos e, portanto, o feedback é mais limitado e, na sua maioria, recebido após o facto.

Para fornecer insights aos instrutores MOOC, precisaríamos identificar uma forma de observar os alunos enquanto eles interagem com o conteúdo. Embora existam muitas dicas que poderíamos usar, decidimos começar explorando a análise de expressões faciais; muitos usuários assistem MOOCs em laptops que possuem webcams integradas, o que nos permite monitorar discretamente os rostos dos alunos enquanto eles interagem com o conteúdo. Assim, definimos o Engajamento do Aluno (LE) como uma medida da atenção (excitação) do aluno combinada com seu nível de emoções positivas/negativas (valência), conforme identificado pelo estudo de suas expressões faciais. Isto fornecerá um feedback inestimável aos instrutores, permitindo-lhes compreender os níveis de LE dos alunos que assistem às suas palestras, o que por sua vez os ajudaria a melhorar o seu conteúdo. Acreditamos que isso poderia levar a um melhor envolvimento e, portanto, a um aumento potencial nas taxas de conclusão.

## II. FUNDO

### A. Reconhecimento de Expressão Facial (FER)

As emoções desempenham um papel crucial na aprendizagem dos alunos [7]. Expressões faciais, movimentos corporais e reações fisiológicas

são formas de comunicação não verbal. As expressões faciais, em particular, são fundamentais para entender como um indivíduo está se sentindo [8]. Uma das ferramentas de análise de movimentos faciais mais abrangentes é o **Sistema de Codificação de Ação Facial (FACS)** proposto em [9] e posteriormente refinado em [10]. O FACS divide quase todos os movimentos faciais possíveis em unidades de ação facial (AUs), que são ações musculares faciais com base anatômica, sem analisar seu significado subjacente. Essas AUs podem ocorrer isoladamente ou ser combinadas para determinar a expressão facial. Essas UAs são mapeadas para identificar as emoções exibidas pelo sujeito. Além das emoções padrão iniciais (Raiva, Nojo, Medo, Felicidade, Tristeza e Surpresa), o Desprezo foi posteriormente incluído como expressão básica [11]. Usamos AUs mais comumente detectadas em emoções e interações sociais [12], que são mostradas nas Tabelas I e II.

Contudo, emoções mistas (por exemplo, surpresa feliz) ou emoções sutis não podem ser claramente mapeadas para as emoções padrão. Em particular, existem estados afetivos chave que estão ligados aos aspectos cognitivos da aprendizagem [14], como tédio, confusão, prazer, envolvimento/fluxo, frustração e surpresa [15]. O mapeamento das UAs para estes estados afetivos centrados na aprendizagem é limitado, uma vez que não foi realizada uma investigação extensa sobre este assunto. Uma extensão é o modelo circunplex de afeto [16]. Este é o modelo dimensional mais comum usado em pesquisas [17], [18]. Este modelo mapeia as emoções em um espaço 2D, ou seja, em valência e excitação. Os autores em [19] estenderam isso mapeando uma série de estados afetivos para pontos em um sistema 2D. Isto permite-nos compreender a relação entre UAs e valência e excitação – uma abordagem que adoptámos.

**Nos métodos FER convencionais, o rosto e os pontos de referência faciais são detectados, as características são extraídas do rosto e a expressão facial** é classificada usando técnicas de aprendizado de máquina, como SVMs. A extração de características é vista como uma etapa crucial [20] e muito trabalho foi realizado com os dois métodos de extração comuns sendo baseados em características geométricas e baseados em características de aparência [21]. O primeiro utiliza as propriedades geométricas de uma face, ou seja, a localização dos pontos faciais, como olhos, nariz, boca etc., a distância entre esses pontos faciais e a forma dos componentes faciais [21], [22]. Esta abordagem foi usada em [22] e [23]. Usamos uma abordagem semelhante, mas apenas como validador. Recentemente, houve um aumento no uso de aprendizagem profunda para FER. Em [24], os autores utilizaram uma CNN de 7 camadas para estimar a ocorrência e intensidade de AU.

Eles usaram os conjuntos de dados BP4D [25] e SEMAINE [26]. Seus resultados sofreram de potencial overfitting e para evitar isso

TABELA I  
UNIDADES DE AÇÃO FACIAL SUPERIORES . ADAPTADO DE [13]











Neutro	AU1 Levantador de sobrancelha interno	AU2 Levantador de sobrancelha externo
		
AU4 Abaixador de sobrancelha	AU5 Levantador de tampa superior	AU6 Elevador de bochecha
		

TABELA II  
UNIDADES DE AÇÃO FACIAL INFERIORES . ADAPTADO DE [13]

Neutro	AU9 Enrugador de nariz	AU12 Extrator de canto labial
		
AU15 Depressor de canto labial	AU25 Parte dos lábios	AU26 Queixo caído
		

usamos três conjuntos de dados diferentes com muitos assuntos. Al-Darraj et al. [27] usaram 23 modelos CNN para identificar 20 UAs e três combinações de UA. Sua arquitetura alcançou uma alta taxa de precisão (> 90% para AUs e emoções), mas eles apenas treinaram e testaram no Extended Cohn-Kanade Dataset (CK+) [28] (um pequeno conjunto de dados) para detecção de AU. Nossa arquitetura é baseada nisso, mas sem a classificação final nos estados emocionais padrão. Além disso, usamos vários conjuntos de dados para treinar os classificadores AU para garantir que os modelos possam generalizar bem. Al-Hamadi et al. [17] extraíram características faciais de dados 3D e de imagem usando várias técnicas de extração de características e mapearam-nas no modelo circunplex 2D de valência e excitação usando redes neurais artificiais.

B. Detecção de envolvimento na educação

Como o FACS nos permite identificar movimentos faciais, ou seja, AUs, e portanto identificam o estado afetivo de uma pessoa, têm sido usados recentemente para detecção de engajamento em um contexto de aprendizagem [29]. Tem havido muita pesquisa sobre o mapeamento das UAs para as emoções padrão [30], mas não tanto em termos de mapeamento das UAs para os estados afetivos centrados na aprendizagem.

McDaniel et al. [31] gravaram os rostos de seus sujeitos enquanto eles interagiam com o AutoTutor [32]. Eles descobriram que algumas características faciais específicas podiam distinguir a confusão, o deleite e a frustração do estado neutro, mas parecia mais desafiador identificar o tédio. Isto indica que o uso de categorias de emoções discretas pode não ser o ideal. Whitehill et al. [33] utilizaram um sistema FER automático (apresentado por [34]). Eles descobriram que o nível de dificuldade de uma videoaula, bem como a velocidade preferida do vídeo para cada assunto, podem ser determinados por meio de expressões faciais. Em [35], eles descobriram que os rótulos de engajamento para um vídeo de 10 segundos poderiam ser estimados com segurança a partir dos rótulos médios dos quadros únicos do mesmo vídeo. Isto indicou que as expressões estáticas poderiam conter a maior parte das informações necessárias para classificar o engajamento. Usamos essa abordagem de tirar uma média de um certo número de quadros.

Grafsgaard et al. [29], [36] usaram CERT [37] e descobriram que ele permitia rastrear movimentos faciais refinados. Eles também descobriram que as UAs da face superior davam indicações de envolvimento, frustração e aprendizagem. Usamos isso para colocar mais foco nas UAs da face superior. Bosch et al. [38] usaram CERT e descobriram que através de AUs, confusão e frustração foram detectadas mais em comparação com tédio, neutro e engajado, o que é

semelhante ao encontrado em [31]. Em [39], [40], Bosch et al. considerou a detecção de estados afetivos na natureza e descobriu foi possível detectar essas emoções, mas houve muitos desafios devido a grandes desequilíbrios no conjunto de dados. Vail et al. [41] usaram AUs junto com técnicas, como postura corporal e condutância da pele para detectar o envolvimento dos alunos. Isto seria impraticável para o nosso propósito, uma vez que é improvável que os alunos queiram usar sensores de condutância da pele enquanto assistem aos MOOCs. Gupta et al. [42] apresentaram o DAiSEE, um conjunto de dados para o envolvimento do usuário na natureza, que contabiliza diferentes estados afetivos. Eles usaram CNNs para detectar esses estados e descobriram que eram mais precisos para classificar a frustração e a confusão do que o tédio e o envolvimento. Isso pode ser devido ao fato de eles não usarem AUs para categorizar expressões faciais, o que significa que movimentos musculares sutis, mas diferenciais, podem ter sido perdidos. Woolf et al. [43] descobriram que valência e excitação poderiam ser usadas para representar estados desejáveis e indesejáveis para a aprendizagem dos alunos. Sua pesquisa mostrou que os estados afetivos centrados na aprendizagem poderiam ser mapeados em valência e excitação, o que foi estendido por [15]. O primeiro utilizou expressões corporais faciais e físicas e comportamento verbal, enquanto o último utilizou sinais fisiológicos multicanais. Estas abordagens não são práticas para observar o envolvimento dos alunos nos MOOCs, pelo que o nosso artigo centra-se apenas nas expressões faciais.

III. METODOLOGIA

Para detectar as unidades de ação (UAs), foram construídos modelos de Redes Neurais Convolucionais (CNN). Eles podem detectar 10 UAs em assuntos novos e inéditos e que são então mapeados em valores contínuos (em vez de discretos) de valência e excitação usando regressões de vetores de suporte (SVRs). Esses valores são combinados de uma maneira inovadora para fornecer uma definição mais abrangente de Envolvimento do Aluno.

A. Coleta de dados

As CNNs foram treinadas usando três conjuntos de dados que continham imagens faciais com rótulos de unidades de ação: • Extended Cohn-Kanade Dataset (CK+) [28], que forneceu 593 sequências de imagens faciais frontais posadas e não colocadas com rótulos AU de 123 sujeitos masculinos e femininos de diferentes etnias.

• A Intensidade de Ação Facial Espontânea de Denver (DISFA) [12] fez com que 27 indivíduos de diferentes etnias assistissem a um videoclipe de 4 minutos, que provocou AUs espontâneos. Essas gravações consistiam em 130.000 quadros, que foram anotados manualmente quanto à presença e intensidade de AUs em uma escala de 0 a 5. Esse conjunto de dados continha um grande número de amostras de uma pequena variedade de sujeitos, o que significava que o uso de todo o conjunto de dados resultaria em overfitting para esses assuntos específicos. Portanto, para cada UA, de cada sujeito retiramos 40 frames que não exibiam aquela UA específica e 50 frames que exibiam. Observe que nem todos os sujeitos exibiram todas as UA que examinamos neste artigo, portanto o número total de quadros exibindo uma UA diferiu para cada UA. Além disso, para cada UA, queríamos garantir que a proporção entre as classes (ou seja, UA e

não AU) foi o mais equilibrado possível, garantindo pelo menos ou o mais próximo possível de 1.000 imagens por turma. • DISFA estendido (DISFA+) [44] fez com que 9 dos 27 sujeitos do conjunto de dados DISFA apresentassem expressões faciais específicas com AUs rotulados.

Para os SVRs, não havia conjuntos de dados acessíveis que tivessem imagens rotuladas de valência e excitação que pudessem ser usadas. Os conjuntos de dados rotulados para ambos consistiam principalmente em imagens selvagens, ou seja, não foram criados em um ambiente de laboratório, portanto nem sempre estavam voltados para a frente. Isso significava que eles não eram adequados para nossos modelos. Portanto, os SVRs foram treinados usando três conjuntos de dados que continham imagens faciais com rótulos de emoção: • CK+, que também possuía rótulos de emoção para 327 das 593 sequências de imagens com uma das sete emoções padrão. • O banco de dados Karolinska Directed Emotional Faces (KDEF) [45] continha 980 imagens frontais de 70 sujeitos, que exibiam uma expressão neutra e as seis emoções padrão (excluindo desprezo). • Radboud Faces Database (RaFD) [46] continha 1.608 imagens frontais de 67 indivíduos de etnia caucasiana e marroquina exibindo as sete emoções padrão e uma expressão neutra.

Para a excitação, os rótulos emocionais das imagens de todos os três conjuntos de dados foram traduzidos em valores de excitação, conforme mostrado na Tabela III. Para valência, apenas o conjunto de dados RaFD foi utilizado, pois cada imagem tinha valência rotulada em uma escala contínua de 1,5 a 4,83, onde uma classificação mais alta indicava um estado positivo.

B. Implementação de modelos CNN

Para treinar as CNNs, as imagens foram inspecionadas e quaisquer imagens com mais de um rosto foram removidas. Das imagens restantes, escolhemos um subconjunto de imagens a ser utilizado para evitar overfitting. Além disso, escolhemos apenas imagens que tivessem as UAs ativadas no nível de intensidade 3 ou mais, uma vez que níveis de intensidade inferiores a este tinham muito pouca presença das UAs. Em seguida, um rosto foi detectado na imagem, que foi redimensionada para um tamanho menor de 160 x 224 pixels (largura x altura) e depois convertida para escala de cinza.

Em seguida, as imagens foram divididas em faces superiores, inferiores e inteiras. A face superior continha apenas a parte superior da imagem e tinha 160 x 144 pixels. A face inferior continha apenas a parte inferior da imagem e tinha 160 x 96 pixels. Todo o rosto tinha o mesmo tamanho da imagem redimensionada. Para AU1, AU2, AU4 e AU5 foi utilizada a face superior desde

TABELA III  
EMOÇÕES PADRÃO EXPRESSAS COM VALORES APROXIMADOS DE EXCITAÇÃO BASEADA NO MODELO CIRCUMPLEX DE [19], EXCETO PARA SURPRESA, QUE FOI DO CONJUNTO DE DADOS AFFECTNET [47]

Excitação Emocional	
Neutro	0,00
Raiva	0,79
Desprezo	0,66
Nojo	0,49
Temer	0,79
Feliz	0,17
Tristeza	-0,40
Surpresa	0,69

essas UAs referem-se apenas às sobranceiras e olhos, então não houve informações redundantes, como a boca incluída. De forma similar, a face inferior foi utilizada para AU12, AU15, AU25 e AU26 Modelos CNN. Para AU6 e AU9 foi utilizada toda a face. Os pixels da imagem foram dimensionados para ficar no intervalo de 0 a 1, por dividindo cada pixel por 255.

Para cada CNN, foram realizadas pesquisas aleatórias para encontre os hiperparâmetros ideais. A arquitetura de todos os dez As CNNs eram semelhantes, pois todas consistiam em três convoluções camadas, cada uma das quais seguida por um Linear Retificado Unit (ReLU) e uma camada de pooling máximo, e depois destes, há eram duas camadas totalmente conectadas, cada uma com uma ReLU e uma camada de abandono de 0,2 probabilidade. A saída do final totalmente camada conectada fluiu para uma camada de saída Softmax, que tinha dois neurônios, um exibindo a probabilidade da UA sendo ativado e o outro mostrando a probabilidade de não sendo ativado. As arquiteturas do modelo diferiam em termos de o número de camadas de normalização de lote (BN) e o número de filtros e neurônios usados na convolução (Conv) e camadas totalmente conectadas (FC), respectivamente. Essas diferenças são descrito na Tabela IV. Observe que todas as camadas BN estão antes do Camada ReLU, e a coluna da tabela 'Camadas BN' indica a Conv e camadas FC, que posteriormente possuem uma camada BN. Um exemplo da arquitetura final (para o modelo AU9) é mostrada na Fig. Todos os outros modelos são exatamente iguais, apenas com a entrada e dimensões da camada e as camadas de normalização de lote são diferentes.

Após treinar esses modelos, testamos sua capacidade preditiva sobre a presença de UAs em imagens inéditas de um dos autores retratando as 10 UAs e uma face neutra. Os 10 modelos tinham capacidades de previsão muito fortes, pois previram 0% para AU presente em imagens neutras e 95%+ para a probabilidade da UA estar presente nas imagens específicas da UA. Além disso, após testes preliminares sobre outro assunto inédito, descobrimos que o modelo AU2 previu AU2 como presente quando não era, o que provavelmente se devia ao fato de o sujeito ter sobranceiras naturalmente altas e a modelo vendo isso como um levantamento sobranceiras. Isso mostrou que a variância genética pode impactar no desempenho. Testamos um dos outros modelos AU2 que encontramos durante as pesquisas aleatórias e não prever a ativação de AU2, exceto nos momentos em que o sujeito realmente levantaram as sobranceiras. Este modelo quando testado em outras imagens AU2 não eram tão fortes na previsão, o que pode deveu-se a este modelo ser menos sensível ao aumento nas sobranceiras. Portanto, para ambos os modelos, decidimos usá-los dependendo da composição genética do sujeito sendo analisado. Observe que nos referimos ao modelo AU2 inicial como AU2a e este novo modelo como AU2b ao longo deste artigo.

Além disso, as CNNs foram combinadas com técnicas baseadas em recursos para melhorar seu desempenho. Esse decisão foi semelhante aos métodos híbridos que alguns artigos usado, onde combinaram abordagens baseadas em aparência e recursos geométricos [48], [49]. Uma pessoa que assiste a um vídeo pode não olhar diretamente para a tela o tempo todo, portanto, as CNNs autônomas (treinadas em imagens faciais frontais) foram limitado em seu desempenho. Esses recursos geométricos baseados validadores foram usados para verificar as previsões das CNNs para reduzir

TABELA IV  
DIFERENÇAS DE ARQUITETURA DE MODELO ENTRE OS MODELOS FINAIS DA CNN

	Nº de filtros em Camada de conversão 1, 2, 3	# de neurônios em Camada FC 1, 2	Camadas BN
AU1	8, 16, 32	256, 128	Todas as camadas de conversão e camada FC 1
AU2a	32, 64, 128	128, 128	Todas as camadas de conversão e camada FC 1
AU2b	32, 64, 128	512, 256	Todas as camadas de conversão e camada FC 1
AU4	8, 16, 32	512, 512	Todas as camadas de conversão e camada FC 1
AU5	32, 64, 128	512, 128	Todas as camadas de conversão e camada FC 1
AU6	16, 32, 64	256, 128	Todas as camadas de conversão e camada FC 1
AU9	32, 64, 128	64, 256	Todas as camadas de conversão e camada FC 1
AU12	32, 64, 128	128, 64	Todas as camadas de conversão e camadas FC
AU15	16, 32, 64	128, 128	Camadas conv 1 e 2 e todas as camadas FC
AU17	8, 16, 32	256, 128	Todas as camadas de conversão e camadas FC
AU20	32, 64, 128	256, 64	Todas as camadas de conversão e camadas FC
AU25	32, 64, 128	64, 128	Todas as camadas de conversão e camada FC 1
AU26	8, 16, 32	256, 128	Camadas conv 1 e 2 e todas as camadas FC

falso-positivo. Por exemplo, se uma CNN previsse que uma UA seria ativado, o validador geométrico verificaria a distância ou localização das características faciais relacionadas a essa UA para verificar se houve mudança suficiente da face neutra para garantir que uma UA seja ativada.

Para fazer isso, usamos Pantic [23] como guia para calcular nosso próprias características faciais geométricas. Eles usaram um modelo de 20 faciais pontos e os rastreou desde a posição neutra até a UA expressão facial. Usamos a abordagem mais flexível de 68 pontos conforme proposto por Gross et al. [50] conforme mostrado na Fig. 2. A ideia para cada validador da UA era calcular o valor euclidiano relevante distâncias entre pontos de referência faciais específicos e compará-los às distâncias em uma face neutra para ver se houve deslocamento relativo suficiente para constituir uma ativação da UA. Os seguintes validadores de métricas de distância foram aplicados:

- **AU1:** Aumento da distância entre o olho e a parte interna da sobranceira, calculado como a distância entre o ponto 22 e 41 para o olho esquerdo e pontos de referência 23 e 48 para o direito olho. Portanto, este validador confirma a ativação da UA se mudar em  $\text{Dist}(22, 41)$  ou  $\text{Dist}(23, 48) > \bar{y}$ . Este  $\bar{y}$  é explicado no final desta lista de validadores. Além disso, isso

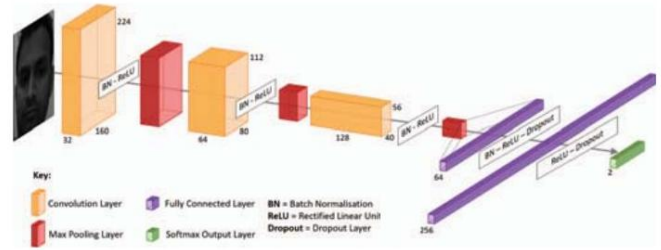
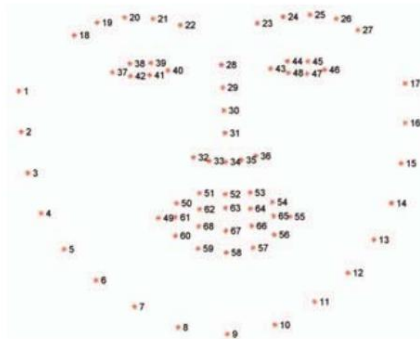


Figura 1. Arquitetura CNN para modelo AU9





2. Marcos faciais anotados em cada imagem, dados pela marcação Multi-PIE 68 pontos [50]. Reproduzido de [52]

AU pode ocorrer em conjunto com AU4, que não apresenta deslocamento significativo entre os olhos e a sobrancelha. Portanto, se o validador AU4 afirmou que AU4 estava ativado, o resultado do validador AU1 foi ignorado. • **AU2:** alteração em  $\text{Dist}(20, 42)$  ou  $\text{Dist}(25, 47) > \bar{y}$ .

No entanto, esta UA também poderia ocorrer em conjunto com AU4, então aplicamos uma condição semelhante à de AU1 • **AU4:** mudança em  $\text{Dist}(22, 23) < \bar{y}$  • **AU5:** Nenhuma característica geométrica foi usada desde o modelo CNN para esta UA tiveram muito poucas previsões incorretas • **AU6:** Diminuição da distância entre as pálpebras superiores e inferiores, calculada usando a relação de aspecto dos olhos (EAR) [51], que calcula o nível de abertura dos olhos. Portanto, este validador confirmou a ativação da UA se mudança em  $\text{EAR} < \bar{y}$  • **AU9:** mudança em  $\text{Dist}(22, 29)$  ou  $\text{Dist}(23, 29) < \bar{y}$  • **AU12:** Nenhuma

característica geométrica foi usada desde o modelo CNN para este UA teve muito poucas previsões incorretas • **AU15:** mudança em  $\text{Dist}(49, 8)$  ou  $\text{Dist}(55, 10) < \bar{y}$  • **AU25:** mudança em  $\text{Dist}(63, 67) > \bar{y}$  • **AU26:** mudança em  $\text{Dist}(34, 9) > \bar{y}$

Para a lista de validadores, o parâmetro de limite mínimo  $\bar{y}$  foi calculado usando o conjunto de dados CK+. Para cada UA, as

imagens relevantes ativadas pela UA e suas respectivas faces neutras foram selecionadas e a distância entre os pontos de referência relevantes para as faces neutras e ativadas pela UA foi calculada e comparada. Essas mudanças na distância foram usadas para calcular  $\bar{y}$ .

Além disso, a posição da cabeça também impactou a capacidade de previsão dos modelos CNN. A posição da cabeça pode ser dissecada em três componentes: inclinação, rotação e guinada. Os números de inclinação e guinada para cada imagem foram usados apenas para verificar se uma face estava significativamente virada para cima ou para baixo (figura de inclinação) ou para a esquerda ou direita (figura de guinada); nesse caso, as previsões da UA foram todas substituídas por 0, uma vez que os modelos da CNN não conseguiam prever bem em faces não frontais. Para contrabalançar a figura do rolo, a imagem foi girada para que o rosto ficasse na vertical.

### C. Implementação de Modelos SVR Os

modelos SVR para valência e excitação utilizaram os resultados dos preditores AU, ou seja, os resultados dos modelos CNN após validação baseada em características geométricas. Estas foram 10 entradas, uma para cada UA (apenas uma de AU2a e AU2b foi selecionada após

o método de validação geométrica). Para treinar o Valence SVR, os rótulos contínuos de valência fornecidos pelo conjunto de dados RaFD foram todos dimensionados para um intervalo de  $(-1, 1)$  usando escalonamento de recursos. Para o SVR de excitação, os rótulos de emoções dos três conjuntos de dados foram traduzidos para excitação conforme a Tabela III. Para ambos os SVRs, cada imagem foi enviada através dos preditores AU para obter as previsões que foram normalizadas, de modo que todos os recursos em cada amostra foram dimensionados proporcionalmente para ter norma unitária, ou seja, todos os recursos na amostra somam 1. Em seguida, pesquisas em grade foram realizadas para encontrar os hiperparâmetros ideais que produziram modelos com os menores erros quadráticos médios e o maior R2.

Um gráfico de dispersão das previsões do modelo de Valência nos dados de teste é mostrado na Fig. 3. Isso mostrou que o modelo foi bom em classificar as previsões da UA em valência com base neste conjunto de dados, mas não foi tão decisivo em torno da valência 0, ou seja, neutro. Isto pode ser devido a dados limitados para emoções sutis.

Um gráfico de dispersão das previsões do modelo de excitação nos dados de teste é mostrado na Fig. 4. Isso foi muito discreto, pois os rótulos de entrada para a excitação estavam em torno das sete emoções padrão. O modelo foi geralmente bom na classificação das previsões da UA quanto à excitação com alguns valores discrepantes, mas teve problemas com pontos de excitação de -0,4, que se referiam à emoção triste. Isto provavelmente ocorreu porque o conjunto de dados rotulou todos os rostos tristes como excitação de -0,4, mas a extensão das ativações de UA nessas imagens variou muito.

Os parâmetros finais do modelo SVR foram: •

**Valência:** C de 0,5,  $\gamma$  de 0,1,  $\bar{y}$  de 1 e um polinômio kernel com 4 graus •

**Arousal:** C de 5,  $\gamma$  de 0,1,  $\bar{y}$  de 1 e um kernel RBF

### D. Detector de envolvimento do aluno

A partir dos valores de valência e excitação, o nível LE teve que ser calculado. Até onde sabemos, não houve nenhuma pesquisa que combinasse valência e excitação para detectar

envolvimento, mas ficou claro que a valência e a excitação afetaram o estado de aprendizagem e estavam interligadas. Gomes et al. [53] descobriram que tanto a valência positiva quanto a negativa impactavam a recordação da memória, mas isso dependia do nível de excitação, sugerindo que a excitação era um pouco mais importante. Além disso, Rowe e Fitness [54] descobriram que emoções negativas, como o tédio, geralmente impactavam negativamente a aprendizagem, mas algumas emoções negativas ativas (ou seja, valência negativa, mas alta excitação), como a frustração, provaram ser benéficas para a aprendizagem sob algumas condições. circunstâncias. Isto sugeriu que se a valência fosse negativa, então o nível de excitação teria um impacto maior sobre

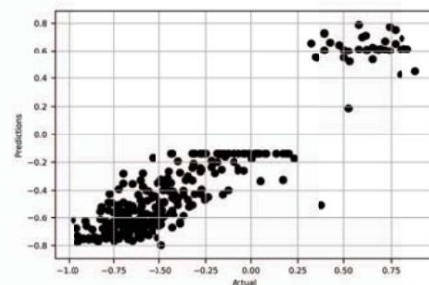


Figura 3. Valores reais de valência versus previsões do modelo SVR final

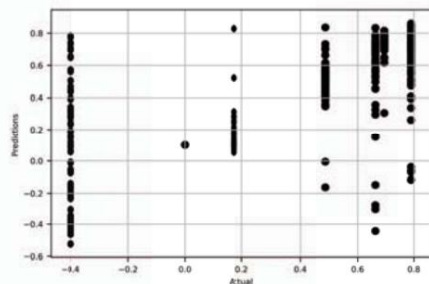


Figura 4. Valores reais de excitação versus previsões do modelo SVR final

aprendizagem, enquanto a valência positiva é considerada benéfica para a aprendizagem. Assim, podemos ter uma definição mais abrangente para o Envolvimento do Aluno (LE) como uma combinação de valência e excitação. Estas ideias influenciaram-nos a criar a função de ponderação que utiliza excitação mais valência para calcular o nível LE.

Para mapear os resultados do SVR em LE, decidimos suavizar esses resultados tomando médias móveis ponderadas exponencialmente (EWMAs). Dado que as emoções de uma pessoa são maioritariamente retratadas pelas expressões mais recentes, foi utilizado um EWMA durante 5 segundos, que foi um período suficientemente longo para garantir que os resultados não são voláteis e é mostrada uma tendência sustentada. Esses EWMAs são convertidos em uma escala de 0 a 10 antes de serem usados para calcular o nível de LE, onde 5 é neutro e 0 e 10 são LE baixo e alto, respectivamente.

#### E. Projeto do Sistema

Uma visão geral do sistema final construído é mostrada na Fig. 5: 1)

Uma webcam gravou um aluno que assistia a um vídeo educacional, utilizando o pacote OpenCV [55] em Python.

2) A gravação foi extraída quadro a quadro e cada quadro foi enviado para o primeiro processo de detecção facial.

3) Este detector facial recuperou os pontos de referência faciais e a rotação do rosto na imagem. Isso foi realizado usando o pacote Dlib [56].

Usando esses pontos de referência, foi encontrada uma aproximação para o ângulo de rotação da face e toda a imagem foi girada nesse ângulo. 4a) Os pontos de referência faciais foram usados para

calcular validadores baseados em características geométricas usados na Etapa 7. 4b) As novas coordenadas de posição

da face na imagem girada foram detectadas usando o detector facial de rede neural profunda (DNN) embutido no OpenCV.

5) Esta imagem é passada para o detector LE, onde foi transformada em escala de cinza e reduzida para as dimensões de entrada de 160 × 224 pixels. Este foi então dividido em face superior, inferior e inteira.

6) As imagens divididas do rosto foram passadas para os preditores da CNN da UA.

7) Os resultados dessas CNNs foram primeiro validados pelo métricas baseadas em recursos geométricos.

8) Estas saídas validadas foram passadas para os dois SVRs.

9) Foram utilizadas as saídas dos SVRs de valência e excitação para calcular o nível LE final.

## 4. EXPERIMENTOS

### A. Estímulo e configuração

Para testar a generalização dos nossos modelos e se a detecção de envolvimento durante a visualização de cursos online poderia ser alcançada, conduzimos dois experimentos. Os primeiros pediram aos sujeitos que assistissem a um vídeo educativo e suas expressões faciais foram analisadas para detectar os níveis de envolvimento do aluno. O segundo experimento testou a generalização dos modelos de UA, de modo que cada sujeito fez expressões faciais específicas e foram tiradas fotos dessas expressões para testar os modelos de UA.

Um pequeno vídeo composto por cinco cliques do YouTube (todos sob a licença Creative Commons [57]) foi criado para o primeiro experimento. Isso simulou uma variedade de métodos de ensino sobre diversos tópicos. A duração de todo o vídeo foi de 4 minutos e 55 segundos, com cada clipe tendo pelo menos 30 segundos. Maiores detalhes são fornecidos na Tabela V. No início do vídeo indicamos que haverá algumas dúvidas no final do vídeo. Isso garantiu que os sujeitos prestassem atenção em todo o vídeo.

Para este experimento, cada sujeito sentou-se em uma sala silenciosa de sua casa ou universidade em frente a um laptop com uma webcam externa montada em cima do laptop. Suas expressões faciais foram gravadas usando uma webcam Logitech C920 em formato 720p a 30 quadros por segundo sob condições normais de iluminação. Para assistir ao vídeo, foi criada uma GUI especial que podia reproduzir, pausar e parar o vídeo com a possibilidade de ajustar o volume. Assim que o assunto pressionou o play, ele começou automaticamente a gravar o assunto em segundo plano. Os participantes foram informados que poderiam pausar o vídeo para fazer perguntas, mas não pausar para dedicar mais tempo ao conteúdo. Embora a capacidade de pausar, pular e retroceder sejam aspectos disponíveis em MOOCs do mundo real, para o propósito deste experimento, decidimos não ativá-la. Depois de assistir ao vídeo, os sujeitos foram informados sobre o experimento e fizeram algumas perguntas sobre sua formação, como idade, sexo, etnia e escolaridade.

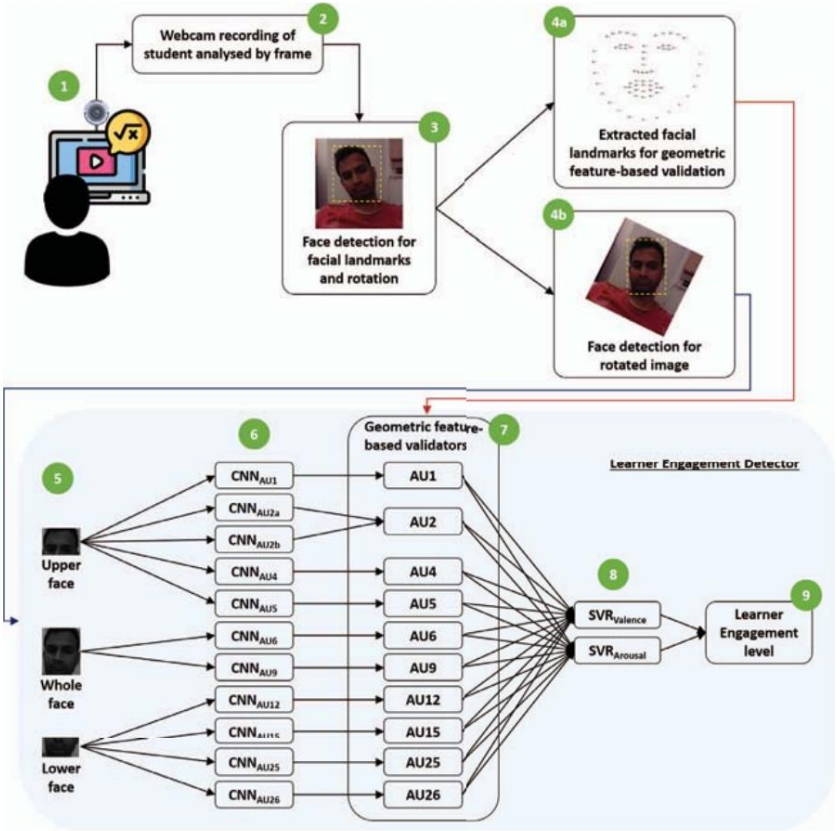
Após este experimento, os sujeitos foram solicitados a exibir expressões faciais de acordo com alguns exemplos de imagens do conjunto de dados CK+ que tinham as AUs rotuladas para o segundo experimento. A foto deles foi tirada usando a mesma webcam.

Quando os sujeitos não conseguiam fazer determinadas expressões, eram solicitados a passar para a próxima.

### B. Sujeitos Os

sujeitos foram cinco adultos (duas mulheres e três homens) na faixa dos 20 anos, estudantes universitários ou recém-formados.

Essa faixa etária foi escolhida porque geralmente assistem ou assistiram recentemente a videoaulas e tutoriais. Em termos de etnia, três eram brancos e dois eram do Leste/Sudeste Asiático. Dois deles tinham formação em Engenharia e os demais tinham formação em Economia ou Ciência da Computação. Nenhum dos sujeitos estava ciente do propósito do experimento no início, nem sabia que suas expressões faciais seriam gravadas. Isso foi para garantir que eles não fossem constrangidos ou tendenciosos ao fazer expressões faciais. Eles foram informados apenas de que assistiriam a um vídeo por cinco minutos e que posteriormente seriam feitas perguntas.



5. Design final do sistema (ícones da Etapa 1 feitos por Freepik em <https://www.flaticon.com>)

TABELA V

DETALHES DOS CLIPES DE VÍDEO DO EXPERIMENTO. A ESCALA DE RÓTULO A5 FOI USADA PARA VALÊNCIA ALVO (MUITO NEGATIVO, NEGATIVO, NEUTRO, POSITIVO, MUITO POSITIVO) E EXCITAÇÃO (MUITO BAIXO, BAIXO, MODERADO, ALTO, MUITO ALTO) DESTINADO A ELICITAR DO GRAMPO

Clípe #	Descrição	Link	Iniciar	Parar	Alvo Valência	Alvo Excitação
1	Um clipe mencionando recursos gratuitos para aprender UX e Design de Interface do usuário	<a href="https://www.youtube.com/watch?v=LupF26Zs5Y">https://www.youtube.com/watch?v=LupF26Zs5Y</a>	00:00	01:10	Neutro	Moderado
2	Uma explicação muito rápida da teoria pitagórica-orem	<a href="https://www.youtube.com/watch?v=QcAUy784UT8">https://www.youtube.com/watch?v=QcAUy784UT8</a>	00:00	01:00	Negativo	Baixo
3	Um quebra-cabeça matemático muito difícil de 30 segundos com a resposta mostrada no final (o som foi silenciado para não distrair o sujeito)	<a href="https://www.youtube.com/watch?v=h6lrjqSpzE">https://www.youtube.com/watch?v=h6lrjqSpzE</a>	00:09	00:52	Muito Negativo	Alto
4	Um quebra-cabeça matemático muito fácil de 30 segundos com a resposta mostrada no final (o som foi silenciado para não distrair o sujeito)	<a href="https://www.youtube.com/watch?v=h6lrjqSpzE">https://www.youtube.com/watch?v=h6lrjqSpzE</a>	01:02	01:39	Positivo	Muito baixo
5	Um experimento científico agradavelmente surpreendente	<a href="https://www.youtube.com/watch?v=lihszmzDqXKU">https://www.youtube.com/watch?v=lihszmzDqXKU</a>	00:24	01:27	Muito Positivo	Muito Alto

V. RESULTADOS

A. Experiência 1 – Envolvimento do Aluno

Após a realização dos experimentos, as gravações foram enviadas através do pipeline do detector LE. Após a análise dos vídeos, percebemos que os sujeitos 2 e 4 posicionaram suas cabeças de forma que nem sempre ficassem voltadas para frente e assim, em alguns momentos, o rosto não podia ser detectado. Portanto, estas previsões da UA eram voláteis, apesar dos validadores baseados em características geométricas e, portanto, não foram incluídas nesta análise.

Para cada um dos outros três assuntos, criamos gráficos de níveis LE a partir das saídas do detector LE e cada gráfico foi analisado para compreender os desvios do neutro. Um

Um exemplo deste gráfico para o Sujeito 1 é apresentado na Figura 6. Neste gráfico, os principais desvios foram destacados com um número, que são explorados em detalhes a seguir. As linhas pontilhadas verticais laranja representavam o início de cada videoclipe, então V1 representa o início do videoclipe 1 do vídeo experimental. Observe que o gráfico LE foi ampliado para ficar entre 3 e 7 para ver mais do impacto e 5 representa o valor neutro.

Para o Sujeito 1 foram encontrados os seguintes desvios:

- 1) O sujeito apertou o play logo após explicarmos o vídeo, quando estava sorrindo. O segundo golpe foi do sujeito, confirmando que o vídeo estava sendo reproduzido.
- 2) Houve uma breve ativação do AU12 no final do vídeo

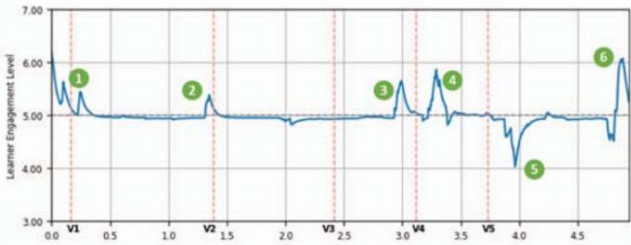


Figura 6. Gráfico de detecção de nível LE do sujeito 1

- clipe 1, onde o instrutor do vídeo fez uma piada.
- 3) No final do videoclipe 3 (um quebra-cabeça matemático difícil), um cronômetro na tela começou a contagem regressiva. Isso resultou no sujeito dizendo o quebra-cabeça em voz alta e resultou em Ativação do AU25, o que significou aumento da excitação.
- 4) Quando o videoclipe 4 começou, qual era a matemática fácil quebra-cabeça, o assunto foi envolvido imediatamente, possivelmente para evitar ficar sem tempo como no clipe anterior. Eles comecei a falar em voz alta imediatamente e depois de encontrar o responder rapidamente, eles voltaram para uma expressão neutra.
- 5) AU1 ativado no início da explicação ao vídeo científico, possivelmente devido a algum nervosismo quando o nome do experimento “cachorro latindo” foi declarado assim eles não sabiam o que esperar.
- 6) Houve uma breve inclinação da cabeça quando o sujeito estava tentando olhar mais de perto a tela, o que foi imediatamente seguido pelo sorriso do sujeito (ativação AU12 e AU6) quando o experimento atingiu seu clímax emocionante.

Portanto, parece que no geral a pessoa foi muito neutra durante todo o vídeo de acordo com os gráficos, o que foi confirmado assistindo ao vídeo. Os movimentos sustentados nas expressões faciais foram bem capturadas com o software, em particular, os níveis LE correspondiam ao assunto, como em pontos 2, 3, 4 e 6, eles se engajaram mais com o conteúdo e no ponto 5, possivelmente sentiram nervosismo, o que poderia ser visto como alguém se tornando menos engajado. Portanto, há foram sinais de que o envolvimento do aluno poderia ser detectado através da análise automática de expressão facial.

Os resultados deste experimento para os três sujeitos mostrou que o detector LE foi capaz de identificar e analisar expressões faciais para gerar níveis de envolvimento do aluno com sucesso. A Tabela VI resume os resultados para todos os assuntos. Em geral, as expressões faciais foram captadas e o unidades de ação mais proeminentes ativadas ao longo do estímulo materiais foram AU1, AU6 com AU12 e AU25. AU1 resultou em um declínio na LE, que Bosch et al. [38] encontrado como um indicador de confusão, o que pode resultar em desligamento nossos resultados concordam com isso. AU12, que ocorreu com AU6, resultou em um aumento no LE, o que concordou com os resultados de Vail et al. [41]. O AU25 por si só não estava presente em anteriores estudos, mas descobrimos que é um indicador positivo de LE, o que fazia sentido para o Sujeito 1, que estava lendo em voz alta o conteúdo de vídeo, o que significa que eles estavam engajados.

No entanto, havia algumas limitações, como assuntos movendo significativamente a cabeça e, em geral, parece

TABELA VI  
RESUMO DOS RESULTADOS DO PRIMEIRO EXPERIMENTO PARA CADA SUJEITO

Sub. 1	detector LE capturou as partes neutras do vídeo, como bem como o envolvimento ativo. Houve alguma correlação com o Sujeito 5, pois ambos riram da mesma piada final do videoclipe 1, que os manteve envolvidos com um LE mais alto do que neutro.
Sub. 2	Nenhum achado apresentado devido a grandes mudanças na postura da cabeça.
Sub. 3	O sujeito manteve um estado neutro durante todo o vídeo, o que foi capturado pelo detector LE. Portanto, identificou que não houve alteração nas expressões faciais (desvios das saídas LE do neutro foram devidas à pose da cabeça movimentos) e, portanto, nenhuma alteração no LE.
Sub. 4	Nenhum achado apresentado devido a grandes mudanças na postura da cabeça.
Sub. 5	Em grande parte em torno de pontos neutros, mas teve alguma volatilidade Níveis de LE devido a movimentos nas previsões de excitação, que foi o resultado dos dados discretos de treinamento para excitação. No entanto, o detector LE foi capaz de contabilizar todos os expressões faciais exibidas ao longo do vídeo. Lá houve alguma correlação com o Sujeito 1, pois ambos riram a mesma piada no final do videoclipe 1.

ser menos comum as pessoas expressarem emoções negativas, o que também foi identificado por McDaniel et al. [31]. Portanto, há uma questão de que as pessoas são mais propensas a expressar sentimentos positivos emoções do que emoções negativas. Além disso, não houve um tamanho de amostra grande o suficiente para obter uma visão agregada do vídeos para entender se havia envolvimento comum ou partes desconectadas do vídeo. No entanto, vimos uma correlação entre os Sujeitos 1 e 5. Isso mostra que este sistema poderia pegar correlações se amostras suficientes forem fornecidas e isso poderia dar feedback útil ao instrutor com base nisso, dando orientação sobre os níveis LE ao longo do vídeo.

B. Experimento 2 – Modelos de Unidades de Ação

Para o segundo experimento, havia imagens invisíveis de todos cinco assuntos que poderiam ser usados para avaliar o desempenho dos modelos AU CNN (sem os validadores baseados em características geométricas). As imagens dos sujeitos para cada UA foram enviado através do respectivo modelo da UA e previsões da UA foram coletadas ativações, que são exibidas na Tabela VII.

A partir desta tabela, ficou claro que a maioria dos modelos previu corretamente as ativações da UA para a maioria dos indivíduos. Todos as previsões acima de 0,7 (70%) de probabilidade dessa UA ser ativados nessa imagem são destacados em verde. Observe que não todos os sujeitos poderiam realizar todas as UA específicas. Os estímulos O material revelou-se difícil de compreender em alguns casos. Para AU15, três dos cinco sujeitos fizeram caretas de emoção triste ao

TABELA VII  
PREVISÕES DOS MODELOS AU DOS SUJEITOS DO EXPERIMENTO 2

	Sub. 1	Sub. 2	Sub. 3	Sub. 4	Sub. 5		
AU1	0,79	AU2a	1,00	0,98	0,03	1,00	1,00
0,05 AU2b	0,77		0,90	1,00	0,95	1,00	1,00
AU4	0,70	AU5 AU6	1,00	1,00	1,00	1,00	1,00
1,00 AU9	0,99					1,00	1,00
AU12	-	-				-	0,98
1,00 AU15	AU25	1,00				1,00	1,00
1,00 AU26	0,94	1,00				1,00	1,00
		1,00				0,99	1,00
	-	-	-			0,22	0,27
		1,00	1,00			0,99	1,00
		1,00	0,72			0,93	1,00



aproximando os lábios, sem o abaixador de canto labial conforme exigido para AU15. Para AU5, eles foram solicitados a fazer uma careta de surpresa, o que também capturou AU2 e AU26. No entanto, três deles não abriram mais os olhos do que o normal. Assim, essas imagens não foram utilizadas para avaliar o modelo.

A grande maioria foi prevista com mais de 95% de probabilidade, o que mostrou a força dos modelos. No entanto, os resultados da AU2a mostraram que para os sujeitos 1 e 2, o modelo previu ativações de AU muito baixas. Isto estava em forte contraste com AU2b. Anteriormente, observou-se que o AU2a funcionava melhor para aqueles com sobranças mais próximas dos olhos, em oposição ao AU2b. Após a inspeção das imagens AU2, notou-se que não havia uma grande diferença na distância entre as sobranças e os olhos em todos os indivíduos. No entanto, os sujeitos 1 e 2 levantaram as sobranças sem enrugar a testa, enquanto todos os outros sujeitos tinham rugas na testa. Portanto, pareceu que o AU2a se concentrou mais nessas rugas como uma característica necessária, mas o AU2b não as considerou necessárias. Além disso, para AU15, os dois sujeitos que produziram imagens válidas tiveram previsões abaixo de 30%. Olhando para as imagens, ficou claro que eles não puxaram os cantos dos lábios para baixo tanto quanto nas imagens de treinamento. Portanto, é possível que o modelo tenha sido treinado para captar expressões mais exageradas. No geral, parece que os modelos foram geralmente bem treinados para trabalhar com dados não vistos, já que em nove entre dez modelos da CNN, houve uma média de pontuações de precisão de 95% na maioria dos assuntos.

## VI. CONCLUSÃO

Neste artigo, estudamos uma abordagem de uso de webcams para monitorar os rostos dos alunos que assistem a MOOCs e traduzir suas expressões faciais em níveis de envolvimento do aluno. Estes dariam um feedback inestimável aos instrutores sobre como adaptar o seu material, o que, por sua vez, poderia ajudar a um melhor envolvimento nos MOOCs e a uma maior taxa de conclusão.

No entanto, existem limitações à nossa capacidade de prever com sucesso os níveis de envolvimento na aprendizagem em determinadas condições. Por exemplo, pessoas que movem significativamente a cabeça ou têm obstruções no rosto (por exemplo, óculos) limitam o desempenho dos modelos. Os validadores geométricos melhoraram o desempenho, porém mesmo estes não foram infalíveis. A melhor maneira de melhorar a precisão nessas condições é treinar em uma gama mais ampla de imagens que inclua várias poses de cabeça e obstruções faciais. Trabalhos futuros poderiam envolver o treinamento com conjuntos de dados adicionais e o uso de técnicas de aprendizagem profunda mais avançadas, como a combinação de memória de longo e curto prazo (LSTM) com CNNs, o que poderia melhorar o desempenho, pois esses modelos teriam conhecimento de previsões anteriores.

Também precisamos ter em mente que a intensidade com que as emoções são demonstradas nas expressões faciais pode variar com base na cultura, nas circunstâncias pessoais e em outros fatores. Além disso, certas emoções, como a frustração, são difíceis de prever apenas através das expressões faciais. Para contrariar esta situação, no futuro, teríamos de ter em conta outros sinais, como a linguagem corporal (especialmente a postura e o contacto visual), a atividade no computador, etc. Isso abre novos e excitantes caminhos de pesquisa.

## REFERÊNCIAS

- [1] D. Shah, "Pelos números: Moocs em 2015", obtido em <https://www.classcentral.com/report/moocs-2015-stats/>, 2015, acessado em: 03/06/2019.
- [2] —, "Pelos números: Moocs em 2019", obtido em <https://www.classcentral.com/report/mooc-stats-2019/>, 2019, acessado em: 06/12/2019.
- [3] L. Rothkrantz, "Taxas de abandono de cursos regulares e moocs", na 8ª Conferência Internacional sobre Educação Apoiada por Computador, CSEDU 2016, ser. Comunicações em Ciência da Computação e Informação, vol. 739. Springer, Cham, abril de 2016, pp. 25–46, doi: 10.1007/978-3-319-63184-4\_3.
- [4] K. Jordan, "Taxas de conclusão do Mooc: os dados", obtido em <http://www.katyjordan.com/MOOCproject.html>, 2015, acessado em: 03/06/2019.
- [5] J. Reich e JA Ruiperez-Valiente, "O pivô mooc", Science, vol. 363, pp. 130–131, janeiro de 2019, doi: 10.1126/science.aav7958.
- [6] LB Krithika e LGG Priya, "Sistema de reconhecimento de emoções do aluno (sers) para melhoria do e-learning com base na métrica de concentração do aluno", Procedia Computer Science, vol. 85, pp. 739.
- [7] MH Immordino-Yang e A. Damasio, "Sentimos, portanto aprendemos: A relevância da neurociência afetiva e social para a educação", Mind, Brain, and Education, vol. 1, não. 3–10, 2007, doi: 10.1111/j.1751-228X.2007.00004.x.
- [8] M. Pantic e L. Rothkrantz, "Sistema especialista para análise automática de expressões faciais", Image and Vision Computing, vol. 881–905, 2000, doi: 10.1016/S0262-8856(00)00034-2.
- [9] P. Ekman e W. Friesen, Sistema de codificação de ação facial: uma técnica para medir o movimento facial. Palo Alto, Califórnia: Consulting Psychologists Press, 1978.
- [10] P. Ekman, WV Friesen e JC Hager, Sistema de codificação de ação facial: o manual em CD ROM. Um rosto humano. Salt Lake City, Utah: Research Nexus, 2002.
- [11] D. Matsumoto, "Mais evidências para a universalidade de uma expressão de desprezo", Motivation and Emotion, vol. 16, não. 363–368, dezembro de 1992, doi: 10.1007/BF00992972.
- [12] SM Mavadati, MH Mahoor, K. Bartlett, P. Trinh e JF Cohn, "Disfa: Um banco de dados de intensidade de ação facial espontânea", IEEE Transactions on Affective Computing, vol. 4, não. 2, pp. 151–160, abril de 2013, doi: 10.1109/T-AFFC.2013.4.
- [13] Y.-I. Tian, T. Kanade e JF Cohn, "Reconhecendo unidades de ação para análise de expressões faciais", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, não. 2, pp. 97–115, fevereiro de 2001, doi: 10.1109/34.908962.
- [14] RW Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cav-allo, T. Machover, M. Resnick, D. Roy e C. Strohecker, "Aprendizagem afetiva - uma manifesto", BT Technology Journal, vol. 22, não. 253–269, outubro de 2004, doi: 10.1023/B:BTJT.0000047603.37042.33.
- [15] MS Hussain, O. AlZoubi, RA Calvo e SK D'Mello, "Detecção de efeitos da fisiologia multicanal durante sessões de aprendizagem com autotutor", em Inteligência Artificial na Educação. AIED 2011. Notas de aula em Ciência da Computação, G. Biswas, S. Bull, J. Kay e A. Mitrovic, Eds., vol. 131–138, ISBN: 978-3-642-21869-9, doi: 10.1007/978-3-642-21869-9\_19.
- [16] JA Russell, "Um modelo circunpleto de afeto", Journal of Personality and Social Psychology, vol. 39, não. 6, pp. 1161–1178, dezembro de 1980, doi: 10.1037/h0077714.
- [17] A. Al-Hamadi, A. Saeed, R. Niese, S. Handrich e H. Neumann, "Traço emocional: Mapeamento da expressão facial para o espaço de excitação de valência", British Journal of Applied Science & Technology, vol. 16, não. 6, pp. 1–14, janeiro de 2016, doi: 10.9734/BJAST/2016/27294.
- [18] S. Zafeiriou, D. Kollias, MA Nicolaou, A. Papaioannou, G. Zhao e I. Kotsia, "Aff-wild: Valência e excitação no desafio selvagem", na Conferência IEEE de 2017 sobre Visão Computacional e Workshops de reconhecimento de padrões, julho de 2017, pp. 1980–1987, doi: 10.1109/CVPRW.2017.248.
- [19] G. Paltoglou e M. Thelwall, "Vendo estrelas de valência e excitação em postagens de blog", IEEE Transactions on Affective Computing, vol. 4, não. 116–123, janeiro de 2013, doi: 10.1109/T-AFFC.2012.36.
- [20] I. Revina e WS Emmanuel, "Uma pesquisa sobre técnicas de reconhecimento de expressão facial humana", Journal of King Saud University - Computer and Information Sciences, 2018, doi: 10.1016/j.jksuci.2018.09.002.
- [21] B. Jiang, M. Valstar, B. Martinez e M. Pantic, "Uma abordagem de descritor de aparência dinâmica para modelagem temporal de ações faciais", IEEE

Transações sobre Cibernética, vol. 44, não. 2, pp. 161–174, fevereiro de 2014, doi: 10.1109/TCYB.2013.2249063.

[22] MF Valstar e M. Pantic, "Reconhecimento totalmente automático das fases temporais das ações faciais", IEEE Transactions on Systems, Man, and Cybernetics, Parte B (Cybernetics), vol. 42, não. 28–43, fevereiro de 2012, doi: 10.1109/TSMCB.2011.2163710.

[23] M. Pantic e I. Patras, "Detectando ações faciais e seus segmentos temporais em sequências de imagens faciais de visão quase frontal", em 2005 Conferência Internacional IEEE sobre Sistemas, Homem e Cibernética, vol. 4, outubro de 2005, pp.

[24] A. Gudi, HE Tasli, TM den Uyl e A. Maroulis, "Ocorrência de unidade de ação facs baseada em aprendizagem profunda e estimativa de intensidade", em 2015, 11ª Conferência Internacional IEEE e Workshops sobre Reconhecimento Automático de Rosto e Gestos (FG), vol. 06, maio de 2015, pp. 1–5, doi: 10.1109/FG.2015.7284873.

[25] X. Zhang, L. Yin, JF Cohn, SJ Canavan, M. Reale, A. Horowitz, P. Liu e JM Girard, "Bp4d-espontâneo: um banco de dados de expressão facial dinâmica 3D espontânea de alta resolução ", Computação de imagem e visão, vol. 32, não. 692–706, 2014, doi: 10.1016/j.imavis. 2014.06.002.

[26] G. McKeown, M. Valstar, R. Cowie, M. Pantic e M. Schroder, "O banco de dados semaine: registros multimodais anotados de conversas emocionalmente coloridas entre uma pessoa e um agente limitado", IEEE Transactions on Affective Computing , vol. 3, não. 5–17, janeiro de 2012, doi: 10.1109/T-AFFC.2011.20.

[27] S. Al-Darraj, K. Berns e A. Rodi, "Reconhecimento de expressão facial baseado em unidade de ação usando aprendizagem profunda", em Advances in Robot Design and Intelligent Control: Proceedings of the 25th Conference on Robotics in Alpe -Região Adria-Danúbio (RAAD16), vol. 540, novembro de 2017, pp.

[28] P. Lucey, JF Cohn, T. Kanade, J. Saragih, Z. Ambadar e I. Matthews, "O conjunto de dados cohn-kanade estendido (ck +): Um conjunto de dados completo para unidade de ação e expressão especificada por emoção, "em 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, junho de 2010, pp.

[29] JF Grafsgaard, JB Wiggins, KE Boyer, E. Wiebe e JC Lester, "Reconhecendo automaticamente a expressão facial: Prevendo engajamento e frustração", em Anais da 6ª Conferência Internacional sobre Mineração de Dados Educacionais, 2013, pp. 50.

[30] B. Martinez, MF Valstar, B. Jiang e M. Pantic, "Análise automática de ações faciais: uma pesquisa", IEEE Transactions on Affective Computing, pp. 2017.2731763.

[31] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp e A. Graesser, "Características faciais para detecção de estado afetivo em ambientes de aprendizagem", Anais da 29ª Reunião Anual do Sociedade de Ciência Cognitiva, pp. 467–472, janeiro de 2007.

[32] AC Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz e o Grupo de Pesquisa de Tutoria, "Autotutor: Uma simulação de um tutor humano", Cognitive Systems Research, vol. 1, não. 35–51, dezembro de 1999, doi: 10.1016/S1389-0417(99)00005-4.

[33] J. Whitehill, M. Bartlett e J. Movellan, "Reconhecimento automático de expressão facial para sistemas de tutoria inteligentes", em 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Work-shops, junho de 2008, pp. 6, doi: 10.1109/CVPRW.2008.4563182.

[34] MS Bartlett, GC Littlewort, MG Frank, C. Lainscsek, IR Fasel e JR Movellan, "Reconhecimento automático de ações faciais em expressões espontâneas", Journal of Multimedia, vol. 1, não. 6, pp. 22–35, setembro de 2006, doi: 10.4304/jmm.1.6.22-35.

[35] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster e J. Movellan, "As faces do envolvimento: reconhecimento automático do envolvimento do aluno a partir de expressões faciais", IEEE Transactions on Affective Computing, vol. 5, não. 86–98, abril de 2014, doi: 10.1109/TAFFC.2014.2316163.

[36] JF Grafsgaard, JB Wiggins, KE Boyer, E. Wiebe e JC Lester, "Reconhecendo automaticamente indicadores faciais de frustração: uma análise centrada na aprendizagem", em 2013, Conferência da Associação Humaine sobre Computação Afetiva e Interação Inteligente, ACII 2013, Setembro de 2013, pp. 159–165, doi: 10.1109/ACII.2013.33.

[37] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan e M. Bartlett, "A caixa de ferramentas de reconhecimento de expressão de computador (cert)," em 2011 IEEE International Conference on Automatic Workshops e reconhecimento facial e de gestos, FG 2011, março de 2011, pp. 298–305, doi: 10.1109/FG.2011.5771414.

[38] N. Bosch, Y. Chen e S. D'Mello, "Está escrito em seu rosto: Detectando estados afetivos de expressões faciais enquanto aprende programas de computador

programação", em ITS 2014: Intelligent Tutoring Systems, ser. Notas de aula em Ciência da Computação, S. Trausan-Matu, KE Boyer, M. Crosby e K. Panourgia, Eds., vol. 8474. cham: Springer International Publishing, junho de 2014, pp. 39–44, doi: 10.1007/978-3-319-07221-0 5.

[39] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang e W. Zhao, "Detecção automática de estados afetivos centrados na aprendizagem na natureza ,," em Anais da 20ª Conferência Internacional sobre Interfaces de Usuário Inteligentes, ser. IUI '15. Nova York, NY, EUA: ACM, março de 2015, pp. 379–388, doi: 10.1145/2678025.2701397.

[40] N. Bosch, SK D'Mello, RS Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang e W. Zhao, "Detectando emoções dos alunos em salas de aula habilitadas para computador", em Proceedings of a Vigésima Quinta Conferência Conjunta Internacional sobre Inteligência Artificial, ser. IJCAI'16. AAAI Press, 2016, pp.

[41] AK Vail, JF Grafsgaard, KE Boyer, EN Wiebe e JC Lester, "Prevendo a aprendizagem a partir da resposta afetiva do aluno às perguntas do tutor", em Anais da 13ª Conferência Internacional sobre Sistemas Tutores Inteligentes - Volume 9684, ser. ITS 2016. Nova York, NY, EUA: Springer-Verlag New York, Inc., 2016, pp.

[42] A. Gupta, R. Jaiswal, S. Adhikari e V. Balasubramanian, "Daisee: Conjunto de dados para estados afetivos em ambientes de e-learning", CoRR, 2018, pré-impressão arXiv: arXiv:1609.01885.

[43] B. Woolf, W. Bursleson, I. Arroyo, T. Dragon, D. Cooper e R. Picard, "Tutores conscientes do afeto: reconhecendo e respondendo ao afeto do aluno", Int. J. Tecnologia de Aprendizagem, vol. 4, não. 3/4, pp. 129–164, setembro de 2009, doi: 10.1504/IJLT.2009.028804.

[44] M. Mavadati, P. Sanger e MH Mahoor, "Conjunto de dados disfa estendido: Investigando expressões faciais posadas e espontâneas", na Conferência IEEE de 2016 sobre Workshops de Visão Computacional e Reconhecimento de Padrões, junho de 2016, pp. : 10.1109/CVPRW.2016.182.

[45] D. Lundqvist, A. Flykt e A. hman, The Karolinska Directed Emotional Faces - KDEF. CD ROM do Departamento de Neurociência Clínica, seção de Psicologia, Karolinska Institutet, ISBN 91-630-7164-9, 1998.

[46] O. Langner, R. Dotsch, G. Bijlstra, DHJ Wigboldus, ST Hawk e A. van Knippenberg, "Apresentação e validação do banco de dados de rostos radboud", Cognition and Emotion, vol. 24, não. 8, pp.

[47] A. Mollahosseini, B. Hasani e MH Mahoor, "Affectnet: Um banco de dados para expressão facial, valência e computação de excitação na natureza", IEEE Transactions on Affective Computing, vol. 10, não. 1, pp. 18–31, janeiro de 2019, doi: 10.1109/TAFFC.2017.2740923.

[48] Y. Zhu, F. De la Torre, JF Cohn e Y.-J. Zhang, "Cascatas dinâmicas com inicialização bidirecional para detecção de unidades de ação em comportamento facial espontâneo", IEEE Transactions on Affective Computing, vol. 2, não. 2, pp. 79–91, julho de 2011, doi: 10.1109/T-AFFC.2011.10.

[49] I. Kotsia, S. Zafeiriou, N. Nikolaidis e I. Pitas, "Fusão de informações de textura e forma para reconhecimento de unidade de ação facial", na Primeira Conferência Internacional sobre Avanços na Interação Computador-Humano, fevereiro de 2008, 77–82, doi: 10.1109/ACHI.2008.26.

[50] R. Gross, I. Matthews, J. Cohn, T. Kanade e S. Baker, "Multi-torta", Computação de Imagem e Visão, vol. 28, não. 5, pp. 807–813, 2010, doi: 10.1016/j.imavis.2009.08.002.

[51] T. Soukupova e J. Cech, "Detecção de piscar de olhos em tempo real usando pontos de referência faciais", no 21ª Workshop de Inverno de Visão Computacional, 2016, pp.

[52] Intelligent Behavior Understanding Group (IBUG), "Facial point annotations", obtido em <https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/>, nd, acessado em: 2019-08- 31.

[53] CAF Gomes, CJ Brainerd e LM Stein, "Efeitos da valência emocional e excitação na recordação lembrada e não lembrada", Journal of experimental Psychology. Aprendizagem, memória e cognição, vol. 39, não. 663–677, 2013, doi: 10.1037/a0028578.

[54] A. Rowe e J. Fitness, "Compreendendo o papel das emoções negativas na aprendizagem e realização de adultos: uma perspectiva social funcional," Ciências Comportamentais, vol. 8, não. 2, pág. 27, 2018, doi: 10.3390/bs8020027.

[55] OpenCV, "Opencv (biblioteca de visão computacional de código aberto) é uma biblioteca de software de visão computacional e aprendizagem de máquina de código aberto", obtido em <https://opencv.org/>, 2019, acessado em: 05/06/2019.

[56] Dlib, "Dlib é um kit de ferramentas c++ moderno contendo algoritmos e ferramentas de aprendizado de máquina", obtido em <http://dlib.net/>, 2019, acessado em: 05/06/2019.

[57] YouTube, "Creative commons", obtido em <https://support.google.com/youtube/answer/2797468?hl=en-GB>, 2019, acessado em: 15/08/2019.