


Classificando emoções e envolvimento online

Aprendizagem baseada em uma única expressão facial

Rede Neural de Reconhecimento

Andrei V. Savchenko , Lyudmila V. Savchenko e Ilya Makarov

Resumo—Este artigo analisa o comportamento dos alunos no ambiente de e-learning. O novo pipeline é proposto com base no processamento facial de vídeo. Num primeiro momento, são aplicadas técnicas de detecção de faces, rastreamento e agrupamento para extrair as sequências de faces de cada aluno. Em seguida, uma única rede neural eficiente é usada para extrair características emocionais em cada quadro. Esta rede é pré-treinada em identificação facial e ajustada para reconhecimento de expressão facial em imagens estáticas da AffectNet usando uma técnica de otimização robusta especialmente desenvolvida. É mostrado que as características faciais resultantes podem ser usadas para uma previsão rápida e simultânea dos níveis de envolvimento dos alunos (de descomprometidos a altamente engajados), emoções individuais (feliz, triste, etc.) e afeto em nível de grupo (positivo, neutro ou negativo). Este modelo pode ser utilizado para processamento de vídeo em tempo real até mesmo no dispositivo móvel de cada aluno, sem a necessidade de enviar seu vídeo facial para o servidor remoto ou PC do professor. Além disso, a possibilidade de preparar um resumo de uma aula é demonstrada ao salvar pequenos cliques de diferentes emoções e envolvimento de todos os alunos. O estudo experimental sobre os conjuntos de dados dos desafios EmotiW (Emotion Recognition in the Wild) mostrou que a rede proposta supera significativamente os modelos únicos existentes.

Termos de indexação—Aprendizagem on-line, e-learning, reconhecimento de expressões faciais baseado em vídeo, previsão de envolvimento, reconhecimento de emoções em nível de grupo, dispositivos móveis

C

1. INTRODUÇÃO

há um crescimento explosivo das tecnologias de educação e aprendizagem eletrônica (e-learning) devido ao impacto da pandemia da COVID-19 [1]. Muitos novos cursos online abertos e massivos (MOOCs) surgiram recentemente [2].

Além disso, muitas universidades e instituições de ensino em todo o mundo mudaram muitas aulas para um formato online. Existem muitos factores essenciais para um ambiente de e-learning eficaz, tais como diferentes elementos relacionados com a avaliação do valor do e-learning e a necessidade de ferramentas suficientes disponíveis na universidade para o ensino remoto [3].

Um dos principais desafios na aprendizagem on-line é a dificuldade de um professor controlar o envolvimento dos alunos em uma aula on-line, de forma semelhante à educação off-line tradicional [4].

Na verdade, todos os microfones, exceto o do educador, devem ser silenciados durante uma palestra, para que seja impossível fornecer

feedback interativo quando a maioria dos alunos fica desembaraçada e/ou barulhenta. Ao ministrar uma palestra off-line, um professor pode usar as emoções individuais do aluno em nível de grupo para desacelerar ou modificar a apresentação [5], mas a presença de pequenos vídeos faciais de cada aluno não pode ajudar um professor durante uma palestra on-line quando o número de ouvintes é relativamente alta [6]. Embora existam alguns estudos sobre a avaliação formativa do envolvimento do aluno baseada em dispositivos móveis [7], parece que apenas a detecção automática do envolvimento dos alunos é a solução mais apropriada [8].

É importante enfatizar que o engajamento não é o único fator essencial para um ambiente de e-learning.

Por exemplo, a dinâmica do estado emocional de cada aluno desempenha um papel importante no processo de aprendizagem [9]. A análise do afeto no nível do grupo pode ser importante para encontrar as partes difíceis ou estranhas da palestra [10]. Sabe-se que todas essas tarefas estão interligadas, pois o engajamento e o afeto estão ligados ao aumento do ganho de aprendizagem e da produtividade [6]. Além disso, os autores do artigo [5] demonstraram que as expressões faciais dos alunos estão significativamente correlacionadas com a sua compreensão da aula.

A maioria das tarefas mencionadas acima foram consideradas em vários subdesafios dos concursos Emotion Recognition in the Wild (EmotiW). Seus vencedores propuseram técnicas bastante precisas que normalmente são baseadas em grandes conjuntos de redes neurais convolucionais profundas (CNNs) [11], [12] e recursos multimodais de áudio, rostos e pose corporal [13], [14], [15]. Como resultado, eles não podem ser usados em muitas aplicações práticas com requisitos de processamento em tempo real em um ambiente com poucos recursos. Além disso, para garantir a privacidade do aluno, é preferível processar os vídeos faciais diretamente no seu dispositivo pessoal (muitas vezes móvel) [16].

Andrei V. Savchenko e Lyudmila V. Savchenko trabalham no Laboratório de Algoritmos e Tecnologias para Análise de Rede, Universidade HSE, 603155 Nizhny Novgorod, Rússia. E-mail: {avsavchenko, lsavchenko}@hse.ru.
Ilya Makarov trabalha no Instituto de Pesquisa em Inteligência Artificial (AIRI), 105064 Moscou, Rússia, e também no Centro de Pesquisa de Big Data, Universidade Nacional de Ciência e Tecnologia MISIS, 119991 Moscou, Rússia. E-mail: makarov@airi.net.

Artigo recebido em 6 de dezembro de 2021; revisado em 28 de abril de 2022; aceito em 29 de junho de 2022. Data de publicação 4 de julho de 2022; data da versão atual 15 de novembro de 2022.

A publicação foi apoiada pela bolsa para centros de pesquisa na área de IA fornecida pelo Centro Analítico do Governo da Federação Russa (ACRF) de acordo com o acordo sobre a concessão de subsídios (identificador do acordo 0D730321P5Q0002) e o convênio com a HSE University nº 70-2021-00139.

(Autor correspondente: Andrei V. Savchenko.)

Recomendado para aceitação por M. Mahoor.

Identificador de objeto digital nº. 10.1109/TAFFC.2022.3188390

Assim, o objetivo deste artigo é o desenvolvimento de soluções rápidas e técnica precisa para classificar emoções e envolvimento que pode ser implementada em software de aprendizagem online em laptops sem GPUs poderosas (processamento gráfico unidades) e/ou dispositivos móveis de alunos e professores. O principal contribuição consiste no seguinte:

FER leve (reconhecimento de expressão facial) modelos baseados nas arquiteturas EfficientNet e MobileNet para extração de características emocionais de imagens. Propõe-se tomar emprestada a ideia de robustez mineração de dados [17] para modificar a função de perda softmax para o treinamento deste modelo para prever emoções em imagens estáticas. Modelo de rede neural eficiente para simultânea detecção de envolvimento e reconhecimento de emoções individuais e de grupo em vídeos faciais. CNN leve do item anterior extrai unificado características emocionais de cada quadro no aluno dispositivo. As características de vários quadros são agregadas em um descritor de vídeo usando estatística (STAT) funções (média, desvio padrão, etc.) [18]. O modelos resultantes nos permitem chegar ao estado da arte resulta em diversas tarefas de reconhecimento de emoções e detecção de engajamento.

Uma nova estrutura tecnológica para tempo real classificação de emoções baseada em vídeo e envolvimento na aprendizagem on-line usando apenas a modalidade facial. O envolvimento e as emoções individuais de cada aluno são previstos no dispositivo de cada aluno. Os vetores de características emocionais obtidos podem ser enviado ao dispositivo do professor para classificar as emoções de todo o grupo de alunos. Se os rostos de alguns alunos na ferramenta de videoconferência online (Zoom, MS Teams, Google Meet, etc.) estão ativados, propõe-se agrupar adicionalmente essas faces e resumir suas emoções e envolvimento durante toda a lição em pequenos vídeos [19]. Esse ajuda os professores a compreenderem a sua própria fraqueza e a mudá-la [5]. As fontes de treinamento e testar código usando Tensorflow 2 e estruturas Pytorch junto com aplicativo Android de demonstração e vários modelos são disponibilizados publicamente¹.

A parte restante deste artigo está estruturada da seguinte forma.

A seção 2 contém um breve levantamento de artigos relacionados. O detalhes da estrutura proposta são fornecidos na Seção 3. A seção 4 fornece resultados experimentais de nossos modelos em EngageWild [8], AFEW (Expressão Facial Atuada no Wild) [20] e conjuntos de dados VGAF (Video-level Group Affect) [10] dos desafios EmotiW. Finalmente, comentários finais e trabalhos futuros são discutidos na Seção 5.

2 PESQUISA DE LITERATURA

2.1 Reconhecimento de emoção baseado em vídeo

O reconhecimento das emoções dos alunos pode ter um grande impacto na qualidade de muitos sistemas de e-learning. Os autores do revisão [9] afirmou que o reconhecimento de emoção multimodal baseado em uma fusão de expressões faciais, gestos corporais e

as mensagens do usuário proporcionam melhor eficiência do que as mensagens monomodais. Recursos semelhantes foram usados em [21] para offline aprendizagem e vídeos de ambientes de sala de aula. Isso é conhecido que as emoções faciais, que são uma forma de comunicação não-verbal, podem ser usadas para estimar o efeito de aprendizagem de um aluno e melhorar as atuais plataformas de e-learning [22].

Assim, neste artigo optou-se por tratar apenas de uma análise da modalidade facial.

Os modelos FER são normalmente pré-treinados em imagens únicas de um conjunto de dados bastante grande, como AffectNet [23]. Excelente resultados foram obtidos recentemente usando métodos supervisionados aprendizagem (SL) e aprendizagem auto-supervisionada (SSL) [24] de EfficientNets [25], transformadores visuais e seletivos de atenção fusão [26], transformadores com reconhecimento de relação (TransFER) [27] e os modelos leves com pré-treinamento cuidadoso no rosto conjuntos de dados de reconhecimento [28]. Emoção muito precisaGCN explora dependências emocionais entre expressões faciais e excitação de valência treinando as redes convolucionais de grafos na estrutura de aprendizagem multitarefa [29].

O progresso no FER baseado em vídeo é medido principalmente em várias versões do conjunto de dados AFEW da EmotiW

Desafios 2013-2019 [20]. Um dos melhores modelos individuais é obtido através do treinamento ruidoso do aluno usando linguagem corporal [30], enquanto o método antigo com a agregação STAT de características extraídas por três CNNs (VGG13, VGG16 e ResNet) [18] ainda é um dos melhores baseados em conjunto técnicas. A melhor precisão de validação é alcançada por os mecanismos de fusão de recursos intermodais de atenção que destaque um recurso emocional importante explorando o recurso concatenação e agrupamento bilinear fatorado (FBP) [31]. No entanto, o último modelo tem uma precisão ligeiramente inferior em o conjunto de testes quando comparado à fusão bimodal [32] de recursos de áudio e vídeo extraídos por quatro CNNs diferentes.

As emoções previstas podem ser usadas não apenas para compreender o comportamento de cada aluno, mas também para resumo visual de vídeos de sala de aula [19] ou classificação do

emoções em nível de grupo em vídeos. Esta última tarefa tornou-se estudado desde o aparecimento do conjunto de dados VGAF [10]. Uma precisão bastante alta é alcançada pelo reconhecimento de atividades e redes de injeção K [33], [34]. O vencedor do subdesafio de reconhecimento de emoções do grupo de áudio e vídeo Emo-tiW 2020 desenvolveu um conjunto de redes híbridas para áudio, emoção facial, vídeo, estatísticas de objetos ambientais e combatendo fluxos de detectores [14].

2.2 Detecção Automática de Engajamento em E-Learning Sistemas

Envolvimento dos pais, interação e envolvimento dos alunos são os principais fatores que podem influenciar a aprendizagem on-line efeitos [1]. Embora a maioria das técnicas de e-learning se concentre em melhorando a interação dos alunos, os algoritmos de análise comportamental e detecção de engajamento tornaram-se estudado recentemente em mineração de dados educacionais [35]. Pesquisadores não temos uma compreensão consistente da definição de envolvimento na aprendizagem e considerá-lo como um processo multidimensional conceito [36]. Neste artigo, um tipo especial de aluno considera-se o esforço persistente para realizar a tarefa de aprendizagem [8], ou seja, o engajamento emocional. Ele se concentra em a extensão das reações positivas e negativas, o sentimento de interesse por um tema específico e gostar de aprender

¹, <https://github.com/HSE-asavchenko/face-emotion-recognition> sobre isso [36].

Uma pesquisa [6] considerou as dependências dos existentes métodos de participação dos alunos e classificou-os em categorias automática, semiautomática (rastreamento de engajamento) e manual. O mais popular ainda é a última categoria. Inclui autorrelatos, listas de verificação observacionais e escalas de classificação e normalmente requer muito tempo e esforço dos observadores [36]. Como resultado, a pesquisa recente o foco mudou para a detecção automática de engajamento que infere as pistas sociais de engajamento/desengajamento a partir de expressões faciais, movimentos corporais e padrão de olhar [8]. É dada especial atenção aos métodos baseados em FER devido à simplicidade de seu uso [6]. Na verdade, o FER e as tarefas de previsão de engajamento estão fortemente correlacionadas [5]. Para Por exemplo, um professor usa as expressões faciais dos alunos como fontes valiosas de feedback. Além disso, as emoções dos palestrantes mantiveram os alunos motivados e interessados durante as palestras [37].

Uma das primeiras técnicas que aplicou aprendizado de máquina e FER para prever o envolvimento dos alunos, foi proposto em [38]. Seus experimentos com máquinas de vetores de suporte (SVM) com recursos Gabor e regressão para expressão os resultados da caixa de ferramentas de reconhecimento de expressão computacional provaram que os detectores de engajamento automatizados funcionam com precisão comparável à dos humanos. Tradicional visão computacional para FER foi usada em [36], onde os histogramas ponderados adaptativos de códigos cinza de oito bits calculados por Local Gray Code Patterns (LGCP) foram classificados por SVM. Os autores deste último artigo introduziram dois conjuntos de dados para detecção de envolvimento de aprendizagem com base em dados faciais e dados de movimento do mouse, mas eles não estão disponíveis publicamente.

Hoje em dia, o progresso do aprendizado profundo causou o uso generalizado de CNNs. Por exemplo, o Engajamento Médio O escore foi proposto em [4], analisando os resultados da avaliação facial detecção de pontos de referência, reconhecimento emocional e os pesos de uma pesquisa especial. A previsão de envolvimento sem contato em ambientes irrestritos é aplicada não apenas em e-learning, mas em outras tarefas interativas, como jogos [39]. A estrutura de avaliação do envolvimento na aprendizagem [2] é oportuna adquiriu as mudanças emocionais dos alunos usando um especial CNN treinada com base na adaptação de domínio, que é adequada para o cenário MOOC.

O rápido crescimento dos estudos em previsão de engajamento começou com a introdução do conjunto de dados EngageWild [8] nos desafios EmotiW 2018-2020. Este conjunto de dados contém vídeos faciais com rótulos de engajamento correspondentes de do usuário, enquanto ele assiste a vídeos educativos como como os dos MOOCs. O olhar, a postura da cabeça e a unidade de ação recursos de intensidades da biblioteca OpenFace [40] foram concatenado no descritor Gaze-AU-Pose (GAP) [41]. Sua classificação utilizando as redes GRU (gated recurrent unit) leva ao MSE (mean square error) na validação conjunto, que é 0,03 menor quando comparado à solução de linha de base para os recursos OpenFace [8]. O uso da dilatação Classificador de Rede Convolutiva Temporal (TCN) [42] para recursos semelhantes do OpenFace levaram a um MSE ligeiramente inferior de 0,0655. Os melhores resultados no conjunto de testes no desafio de 2018 foram obtido por descritor facial LBP-TOP adicional e C3D recursos de ação [43].

Os autores desta última abordagem melhoraram-na para o Desafio EmotiW 2019 usando o bootstrap clássico agregação e projeto de uma perda de classificação como uma regularização

que impõe uma margem de distância entre as características de pares de categorias distantes e pares de categorias adjacentes [44]. O estratégia anti-overfitting com treinamento sobreposto segmentos de vídeos de entradas foi proposto em [45]. Solução de os vencedores [46] usaram as características de comportamento facial extraídas por modelo OpenFace e ResNet-50 pré-treinado no rosto identificação no grande conjunto de dados VGGFace2 [47]. Esses os resultados foram melhorados no desafio de 2020 usando um GRU baseado em atenção e processamento de vídeo multi-taxa [13].

3 MATERIAIS E MÉTODOS

3.1 Abordagem Proposta

A maioria das técnicas mencionadas na seção anterior usou modelos de conjuntos complexos e vários conjuntos de recursos para aumentar seu desempenho. Infelizmente, cada modelo para um conjunto de recursos relatado nestes artigos não pode competir com as soluções finais. Assim, neste artigo, o uma nova estrutura tecnológica (Fig. 1) é proposta para analisar o comportamento dos alunos em aulas on-line.

Aqui cada aluno pode lançar uma aplicação no seu seu dispositivo para fornecer os resultados da análise do comportamento sem a necessidade de compartilhar o vídeo facial. Como resultado, a alta nível de privacidade de dados pode ser alcançado porque o vídeo de não é necessário enviar um rosto para o servidor remoto ou computador do professor. Neste caso, a maior região facial está localizada em cada t-ésimo quadro de vídeo na unidade “Detecção de rosto 1” por usando qualquer técnica rápida, como MTCNN (multi-tarefa CNN). A seguir, as características emocionais são extraídas de um rosto extraído são obtidos na unidade “Extração de características emocionais 2” [48] usando a CNN leve proposta, treinada para classificar emoções em imagens estáticas. Os detalhes sobre o treinamento desta rede neural serão fornecidos na próxima subseções. Finalmente, as características de vários quadros com duração de 5 a 10 segundos são agregados usando Funções STAT para classificar o nível de engajamento e individual emoções nas unidades “Previsão de engajamento 3” e “Reconhecimento de emoção 4”, respectivamente. Nível de envolvimento previsto (de descomprometido a altamente engajado) e emoções individuais (feliz, irritado, triste, neutro, etc.) para cada intervalo de tempo na saída das unidades 3 e 4 junto com o características emocionais na saída da unidade 2 são enviadas para o computador do professor. Como os modelos nas primeiras quatro unidades são muito eficientes, a inferência pode ser lançada mesmo em qualquer ambiente com poucos recursos, como um dispositivo móvel dos alunos [16], [28].

O processamento mais difícil é implementado no dispositivo do professor nas unidades 5-13. Essas etapas podem ser executadas off-line modo após obter a gravação de toda a palestra em a ferramenta de videoconferência online. Este vídeo é alimentado no “Detecção de rosto 5”, que funciona de forma semelhante à primeira unidade no dispositivo do aluno, mas pode retornar várias (K 1) faces de alunos que concordaram em transferir seus vídeos. Ainda é possível que várias imagens faciais extraídas tenham resolução muito baixa para reconhecimento preciso de emoções [49]. Neste artigo, a solução mais simples é implementada, de modo que as faces com tamanho inferior a um limite predefinido (64x64 pixels) são ignorado. A seguir, o vetor de características emocionais de cada k-ésima face são obtidas em “Extração de características emocionais 6”, que pode usar a mesma CNN da unidade 2 ou mais arquitetura complexa se o processamento for implementado em um

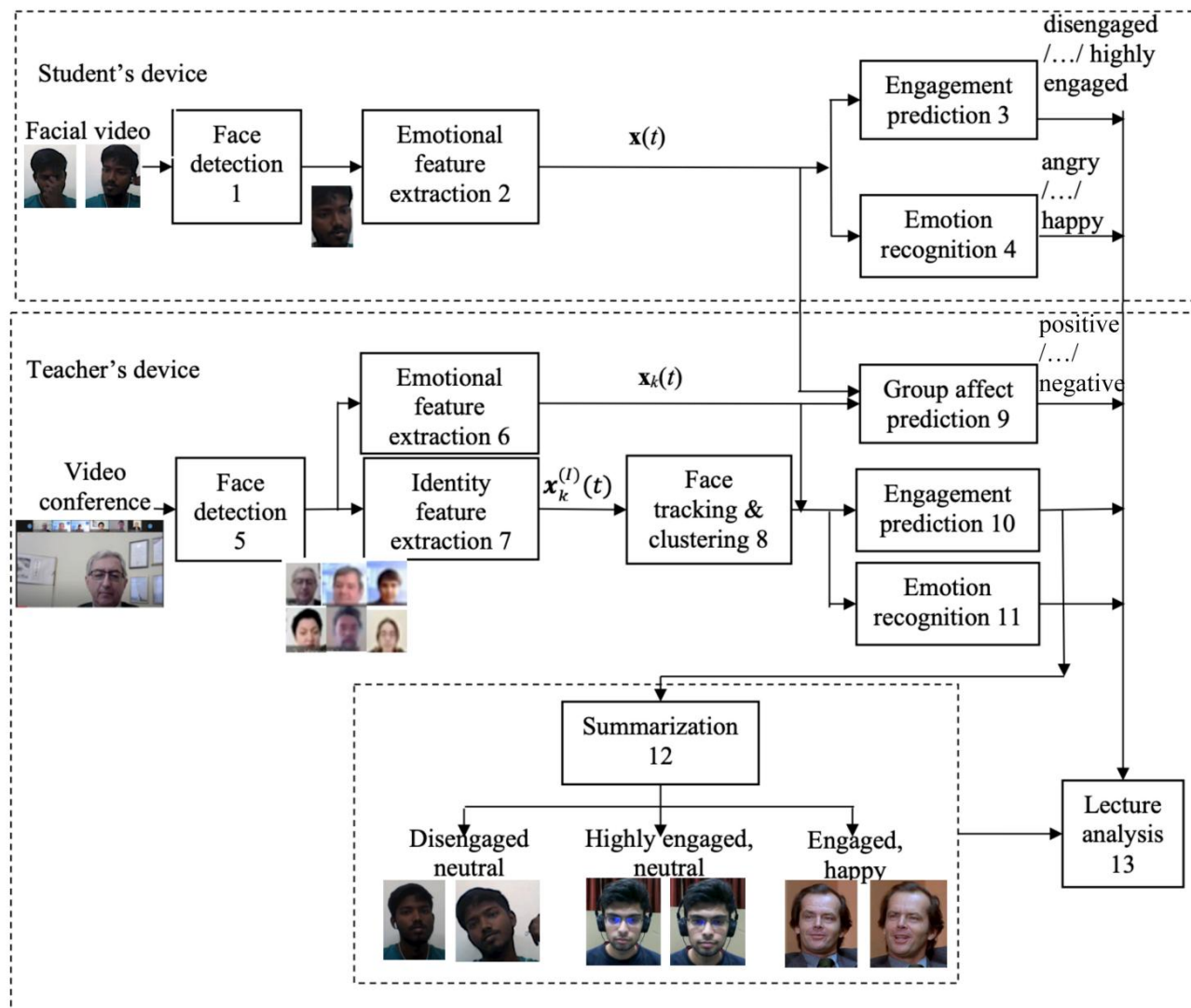


Figura 1. Pipeline proposto.

PC bastante poderoso. As características de identidade $x_k^{(I)}(t)$ extraído na unidade 7 usando reconhecimento facial apropriado CNN [28], [47], [50]. Os últimos recursos são usados para rastrear e agrupar as regiões faciais dos mesmos alunos da unidade "Rastreamento facial e agrupamento 8". As características emocionais $x_k(t)$ na saída da unidade 8 da mesma trilha são combinados para resolver as tarefas posteriores nas unidades 9, 10 e 11. A unidade "predição de efeito de grupo 9" primeiro agrega características emocionais de rostos do mesmo quadro em recursos de quadro único de todo o grupo de alunos. Em seguida, todos os recursos de quadro durante 5 a 10 segundos de um vídeo são combinados em um único descritor que pode ser alimentado em um classificador apropriado. A "Previsão de engajamento 10" e "Reconhecimento de emoção 11" funcionam de forma idêntica às unidades 3 e 4, mas repita o processamento para cada k-ésima face e cada grupo de quadros.

Finalmente, as emoções e o envolvimento de cada aluno podem ser resumidos em pequenos vídeos e visualizados na unidade "Resumo 12". Por exemplo, é possível tomar os pontos de tempo onde a emoção forte é previsto. Os resultados típicos de diversas conferências reais ou lições são apresentadas na Fig. 2. Outra oportunidade é o agrupamento de diferentes emoções com base no espaço 2D de Russel de afeto [51] que pode dar ao professor uma impressão inicial

sobre como os alunos são concentrados, afetados e inspirados durante a aula eletrônica. Por fim, um pequeno GIF com diferentes emoções e envolvimento durante uma aula pode ser enviado para um aluno de seus parentes para aumentar o poder parental envolvimento [1]. Além disso, esses cliques, juntamente com o gráficos de emoções previstas dos alunos, afeto de grupo e engajamento dependendo do tempo são armazenados na unidade "Análise de palestra 13" que pode ajudar os educadores online detectar com precisão o status de envolvimento de seus alunos on-line [6] e organizar melhor seus materiais. Isso é também é possível destacar os pontos de tempo com valores altos ou envolvimento muito baixo para encontrar as partes estranhas ou difíceis de a palestra. Esses dados permitem acompanhar a eficiência das aulas e aumentar a conversão de cursos online.

3.2 Otimização Robusta da Rede FER

Nesta subseção, vamos descrever o procedimento para treinar o Rede FER que extrai características emocionais robustas de fotos estáticas ou quadros de vídeo. A princípio, um peso leve A CNN é treinada para identificação facial em um conjunto de dados muito grande [47]. Em seguida, esta rede é ajustada para qualquer emoção conjunto de dados com fotos faciais estáticas [23]. Como existe emocional conjuntos de dados são normalmente altamente desequilibrados, o peso ponderado

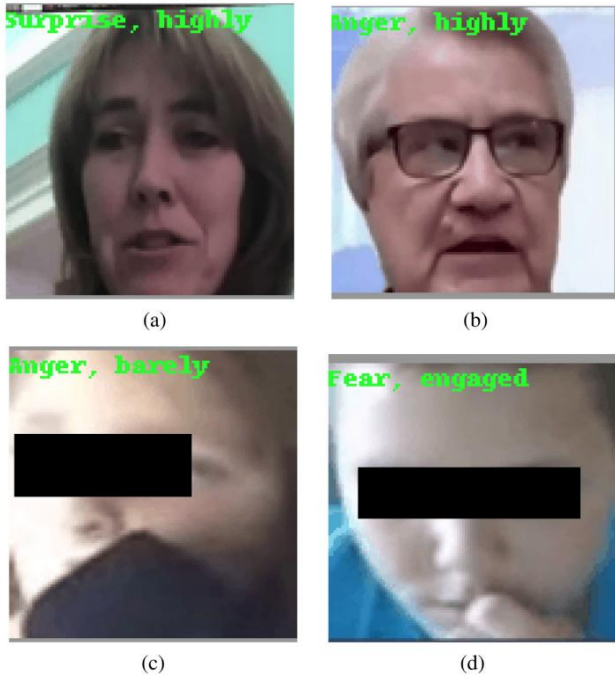


Figura 2. Exemplos de quadros da apresentação resumida da conferência (a), (b) e lição (c), (d). Os rostos das crianças estão parcialmente ocultos.

a perda de entropia cruzada categórica (softmax) é otimizada [23], [28]:

$$L_{\text{D}}(X; y) = -\sum_{i=1}^N \log \text{softmax}(\hat{y}_i) \quad (1)$$

onde u é o vetor de pesos de uma rede neural, X é o
imagem de treinamento, y 2 f1; ... ; C_g é seu rótulo emocional, $z_{y\delta X}$; $\text{você}P$
é a y -ésima saída da penúltima camada (logits) do
CNN com entrada X e softmax é a ativação do softmax
função.

Aqui os pesos das classes são definidos inversamente proporcionais
ao número total N_y de exemplos de treinamento da y -ésima classe

$$w_y = \frac{1}{N_y} \max_{c \in \{1, \dots, C_g\}} \frac{N_c}{C_g} \quad (2)$$

Para melhorar a qualidade dos modelos emocionais treinados, é
propôs usar mineração de dados robusta [17] e formular o
tarefa de otimização da seguinte forma:

$$\min_u \max_{j \in \{1, \dots, N\}} L_{\text{D}}(X; y) \quad (3)$$

Sabe-se que o vetor gradiente dá a direção de
aumento máximo de uma função. Portanto, a tarefa de otimização robusta
(3) pode ser simplificada da seguinte forma:

$$\min_u L_{\text{D}}(X; y) \quad (4)$$

onde r_{L-} é definido como o gradiente normalizado L2 do
função de perda:

$$r_{L-} = \frac{1}{\| \nabla L_{\text{D}} \|_2} \nabla L_{\text{D}} \quad (5)$$

A tarefa de otimização (4) é resolvida usando uma modificação
de qualquer método estocástico de descida gradiente. O completo

Uso licenciado autorizado limitado a: Universidade Federal de Alagoas. Baixado em 28 de março de 2024 às 19:15:40 UTC do IEEE Xplore. Restrições aplicadas.

procedimento de otimização é mostrado no Algoritmo 1. Ele modifica
o otimizador Adam [52] na notação do clássico
livro [11]. Além dos parâmetros de Adam [52], a saber,
taxa de aprendizagem h e taxas de decaimento exponencial para
estimativas de momento r_1 (valor padrão 0,9) e r_2 (0,999 por padrão), o
parâmetro é adicionado para controlar o nível de incerteza.

Algoritmo 1. Otimização Robusta Baseada em Adam de FER Rede neural

Exigir: Pesos de uma CNN pré-treinada em reconhecimento facial

tarefa, conjunto de treinamento \mathcal{X}_m ; y_m \mathcal{P}_g de imagens faciais com
rótulos emocionais

Garanta: Pesos u que otimizem a perda robusta (3)

```

1: Inicializar acumuladores  $s$ :  $\frac{1}{4}$  0;  $r$ :  $\frac{1}{4}$  0 e intervalo de tempo  $t$   $\frac{1}{4}$  1
2: para  $a$  época 2 f1; ... ; NumEpochsg fazer
3: para lote 2 f1; ... ; NumBatchesg fazer
4: Amostra de minilote de  $M$  exemplos  $\mathcal{X}_m$ ;  $y_m$   $\mathcal{P}_g$ 
5: para  $m$  2 f1; ... ;  $M$ g fazer
6: Alimente a imagem  $X_m$  em uma CNN com pesos  $u$  e obtenha
   a  $y_m$ -ésima saída  $p\delta 1P$  eu  $z_{y_m\delta X_m}$ ;  $\text{você}P$ 
7: fim para
8: Calcular gradiente usando backprop
    $g1$ :  $\frac{1}{4}$   $\frac{1}{M} \sum_{m=1}^M \nabla L_{\text{D}}(X_m; y_m)$  eu  $0$ 
9: Calcular pesos  $u$ :  $\frac{1}{4}$   $u$   $p$   $g1 = jg1$ 
10: para  $m$  2 f1; ... ;  $M$ g fazer
11: Alimente a imagem  $X_m$  em uma CNN com pesos  $u$  e obtenha
   a  $y_m$ -ésima saída  $p\delta 2P$  eu  $z_{y_m\delta X_m}$ ;  $\text{você}P$ 
12: fim para
13: Calcular gradiente usando backprop
    $g2$ :  $\frac{1}{4}$   $\frac{1}{M} \sum_{m=1}^M \nabla L_{\text{D}}(X_m; y_m)$  eu  $0$ 
14: Atribuir  $t$ :  $\frac{1}{4}$   $t + 1$ 
15: Atribuir  $s$ :  $\frac{1}{4}$   $r1$   $s$   $p$   $\delta 1$   $r1P$   $g2$ 
16: Atribuir  $r$ :  $\frac{1}{4}$   $r2$   $r$   $p$   $\delta 1$   $r2P$   $g2$   $g2$ 
17: Atualizar pesos  $u$ :  $\frac{1}{4}$   $u$   $\frac{hs = \delta 1 p r t 1 P}{dp r = \delta 1 p r t 2 P}$ 
18: fim para
19: fim para
20: retorne os pesos aprendidos para você

```

3.3 Detalhes de Treinamento do Modelo de Processamento Facial

Vamos resumir todo o procedimento de treinamento do FER

modelos neste artigo. Várias arquiteturas leves

foram treinados, nomeadamente, MobileNet v1, EfficientNet-B0 e

EfficienteNet-B2 [25]. Eles foram treinados em duas etapas com

(1) pré-treinamento em reconhecimento facial; e (2) ajuste fino em

classificação das emoções. Os detalhes do procedimento de treinamento

para a primeira etapa foram descritos nos artigos anteriores [28],

[50]. É importante ressaltar que as CNNs foram

treinadas nas faces cortadas sem quaisquer margens, de modo que

a maior parte do fundo, cabelos, etc. não é apresentada.

Como resultado, a precisão do reconhecimento facial pode ligeiramente

degradar, mas as características faciais aprendidas são mais adequadas

para análise emocional.

A segunda etapa de treinamento foi implementada da seguinte forma. A

um conjunto de treinamento altamente desequilibrado de 287.651 imagens

foi utilizado a partir do conjunto de dados AffectNet [23] anotado com $C = 8$

expressões básicas (Raiva, Desprezo, Nojo, Medo, Felicidade, Neutro,

Tristeza e Surpresa). A validação oficial

Um conjunto de 4.000 imagens (500 por turma) foi utilizado para fins de

teste. Os rostos foram cortados com as caixas delimitadoras

Uso licenciado autorizado limitado a: Universidade Federal de Alagoas. Baixado em 28 de março de 2024 às 19:15:40 UTC do IEEE Xplore. Restrições aplicadas.

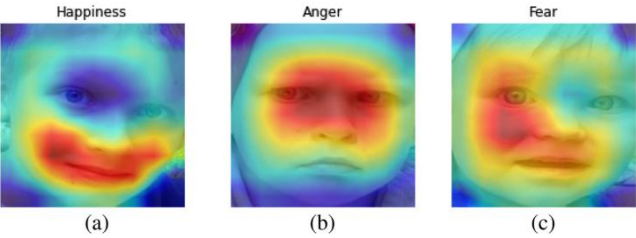


Figura 3. Visualização das emoções infantis previstas pelo EfficientNet-B2.

fornecido pelos autores do AffectNet. O impacto do pré-processamento adicional foi examinado, no qual as imagens faciais recortadas foram giradas para alinhá-las com base na posição dos olhos de forma semelhante a [29], mas sem aumento de dados.

A última camada da rede pré-treinada em VGGFace2 foi substituída pela nova cabeça (camada totalmente conectada com saídas C e ativação softmax), de forma que a penúltima camada com neurônios D possa ser considerada como um extrator de características faciais. O modelo foi treinado totalmente em 8 épocas pelo otimizador robusto baseado em Adam (Algoritmo 1). Em particular, a taxa de aprendizagem foi definida em 0,001 nas três primeiras épocas e ajustou-se apenas aos pesos da última camada da CNN pré-treinada. Finalmente, toda a rede foi treinada com uma taxa de aprendizagem de 0,0001 nas últimas cinco épocas. O parâmetro de incerteza é igual a 0,05 em todos os experimentos.

Embora o AffectNet contenha principalmente fotos de adultos, os modelos FER resultantes podem classificar emoções mesmo para crianças pequenas. Por exemplo, a Figura 3 contém as classes emocionais previstas e a visualização GradCAM de fácil interpretação da decisão da CNN.

Sabe-se que muitos artigos existentes [29], [53] relatam o desempenho de seus métodos apenas para 7 emoções básicas (sem Desprezo), para cada uma existem 283.901 imagens de treinamento e 3.500 imagens de validação. Assim, foram estudadas duas opções para comparar os modelos propostos com resultados existentes para 7 estados emocionais (sem Desprezo). A primeira opção é treinar a CNN no conjunto de treinamento completo com 8 classes, mas usar apenas 7 previsões da última camada Softmax para 3.500 imagens de validação, de modo que a saída que corresponde à classe Desprezo seja simplesmente ignorada. A segunda opção é treinar novamente o modelo com 7 saídas no conjunto de treinamento reduzido. Esta abordagem leva a uma melhor precisão do que a anterior, embora o uso do modelo universal de 8 classes seja desejável se a emoção de desprezo puder ser usada no futuro.

As CNNs obtidas foram aplicadas na extração de características emocionais faciais para processamento de vídeo. Em particular, a última camada Softmax foi removida, e as saídas dos neurônios D na penúltima camada foram usadas como um vetor de características D-dimensional em experimentos posteriores. O modelo multitarefa completo proposto para reconhecer as emoções e o envolvimento dos alunos em um vídeo é mostrado na Fig. 4. Ele explora a ligação conhecida entre uma expressão facial e o nível de compreensão que ajuda os professores a melhorar seu estilo de acordo e mantém os alunos interessados e entusiasmados durante as palestras virtuais [5].

Aqui, duas EfficientNets [25] são aplicadas para extrair identidade facial e características emocionais, respectivamente. A segunda CNN é obtida ajustando a primeira CNN em um grande conjunto de dados FER. Dois módulos estatísticos (codificação STAT) são usados para agregar as características de todos os rostos em um vídeo. O uso licenciado autorizado limita-se a: Universidade Federal de Alagoas. Baixado em 28 de março de 2024 às 19:15:40 UTC do IEEE Xplore. Restrições aplicadas.

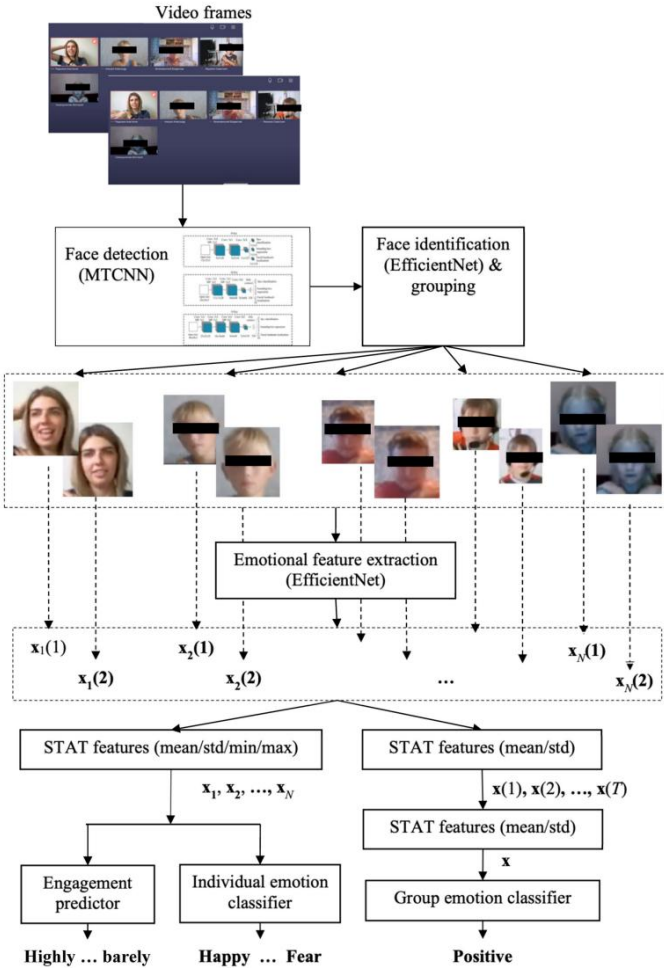


Figura 4. Modelo proposto baseado nas funcionalidades do EfficientNet.

A primeira unidade combina o resultado de diversas funções estatísticas (mínimo, máximo, média e desvio padrão) calculadas para cada característica de todos os quadros de vídeo. O descritor normalizado em L2 é classificado para prever a emoção de cada aluno. O desvio padrão das características emocionais de cada rosto é usado para prever o nível de envolvimento.

Finalmente, a média e o desvio padrão das características emocionais de todos os rostos em cada quadro são concatenados e agregados em um único descritor de vídeo, e o efeito geral de todo o grupo é reconhecido.

4 RESULTADOS EXPERIMENTAIS

4.1 Extração de Características Emocionais

Nesta subseção, o reconhecimento de emoções é estudado em imagens estáticas para o conjunto de dados AffectNet. Os resultados nos conjuntos de validação para os modelos FER propostos em comparação com os resultados do estado da arte são mostrados na Tabela 1. A fonte em negrito indica a melhor pontuação para experimentos com 8 emoções (Raiva, Desprezo, Nojo, Medo, Felicidade, Neutro, Tristeza e Surpresa) e 7 categorias emocionais básicas (as mesmas sem Desprezo).

Como se pode notar, o melhor modelo EfficientNet-B2 melhora a precisão do estado da arte anteriormente conhecida [55] para validação completa definida de 62,09% a 63,025%. É ligeiramente (0,1%) menos preciso que o EmotionGCN [29] para 7 classes, embora a arquitetura da rede não tenha sido modificada e o

TABELA 1
Precisão para o conjunto de validação AffectNet

Modelo	Precisão, %	
	8 aulas	7 aulas
Linha de base (AlexNet) [23]	58,0	-
Perda profunda do centro de atenção [53]	-	65,20
Estudante destilado [54]	61,60	65,40
EfficientNet-B2 (SL + SSL in-panting-pl) [24]	61,32	EfficientNet-B0 -
(SL + SSL in-panting-pl) [24]	61,72	DAN [55] -
	62,09	65,69
Transferir [27]	-	66,23
EmoçãoGCN [29]	-	66,46
Nosso MobileNet-v1	60,20	64,71
Nosso EficienteNet-B0	61,32	65,74
Nosso EficienteNet-B2	63,03	66,34

MESA 2
Precisão em nível de classe de reconhecimento de emoções em fotos Statis, conjunto de dados AffectNet

Emoção	MobileNet-v1	EfficientNet-B0	EfficientNet-B2
Raiva	62,8	61,4	54,2
Desprezo	48,0	60,4	66,0
Nojo	51,8	50,0	65,4
Temer	66,8	66,2	63,8
Felicidade	81,8	78,0	74,6
Neutro	58,6	53,4	54,6
Tristeza	61,8	59,4	65,4
Surpresa	56,0	61,8	60,2

processo de treinamento foi simples. A precisão do Mobile-Net e EfficientNet-B0 é menor, mas ainda comparável com o resultados mais conhecidos relatados para este conjunto de dados. É importante enfatizar que embora a precisão média dos valores mais profundos O modelo EfficientNet-B2 é maior, a precisão da classe para cada O tipo de emoções às vezes é menor quando comparado ao EfficientNet-B0 e ao MobileNet (Tabela 2), de modo que todos os nossos modelos pode ser útil em diferentes tarefas posteriores.

O estudo detalhado de ablação de experimentos para AffectNet é apresentado nas Tabelas 3 e 4. Nesta última tabela, o maior A precisão de 8 e 7 classes para cada linha (modelo) é marcado por negrito.

Aqui foram examinados dois conjuntos de dados para pré-treinamento, nomeadamente, (1) ImageNet convencional; e (2) VGGFace2 [47] para aprender as incorporações faciais adequadas para reconhecimento facial. Os modelos oficiais pré-treinados no ImageNet foram tirados dos modelos de imagem Tensorflow 2 e PyTorch (timm) para a abordagem anterior. Esta última técnica foi implementada conforme descrito na Seção 3.3. Como se pode notar, tal pré-treinamento leva a uma precisão muito melhor do FER, embora características de identidade facial não devem depender do emocional estado [28]. Além disso, a Tabela 4 demonstra que o robusto otimização (Algoritmo 1) permite aumentar o precisão. É especialmente perceptível para o melhor modelo EfficientNet-B2, que estabeleceu um novo resultado de última geração para conjunto de validação completo com 8 classes.

4.2 Previsão de engajamento

Nesta subseção, os resultados no EngageWild [8] são relatado. Este conjunto de dados contém 147 treinamentos e 48 validações

TABELA 3
Estudo de ablação dos modelos propostos, conjunto de dados AffectNet

Modelo	Precisão, %	
	Conjunto pré-treino	8 aulas 7 aulas
MobileNet-v1	ImageNet 56,88	ImageNet 60,4
EficienteNet-B0	57,55	ImageNet 60,28 60,8
EficienteNet-B2	VGGFace2 58,70	64,3
SENet-50		62,31
Nosso MobileNet-v1, 8 classes	VGGFace2 60,25	Nosso MobileNet-v1, 7 classes 64,71
VGGFace2 Nosso EfficientNet-B0, 8 classes		64,57
VGGFace2 Nosso EfficientNet-B2, 8 classes	VGGFace2	- 65,74
63,03 Nosso EfficientNet-B2, 7 classes	VGGFace2	66,29
		- 66,34

TABELA 4
Estudo de ablação de otimizadores, conjunto de dados AffectNet

CNN	Precisão de 8 classes, %		Precisão de 7 classes, %	
	Adam Robusto (3)-(5)	Adam Robusto (3)-(5)		
MobileNet-v1 59,87	EfficientNet-B0 60,94	EfficientNet-B2 62,11	60,25	64,54 64,71
			61,32	65,46 65,74
			63,03	65,89 66,34

vídeos com duração média de 5 minutos. Cada vídeo é associado a um dos 4 níveis de engajamento 0, 0,33, 0,66 e 1 representando engajamento mapeado para desengajado, mal engajado, engajado e altamente engajado.

As imagens do quadro foram extraídas do vídeo usando a ferramenta FFmpeg, e as regiões faciais foram encontradas em cada quadro usando o detector MTCNN. Se nenhum rosto for detectado, o quadro foi ignorado. A seguir, os modelos emocionais desenvolvidos foram utilizados para extrair características da maior região facial. O descritor final de todo o vídeo foi calculado como um desvio padrão das características faciais em termos de quadro semelhantes a a linha de base [20]. Tentamos usar outros recursos STAT (média, máximo, mínimo), mas não observaram melhorias em o MSE (erro quadrático médio) medido na validação definir. O descritor de vídeo obtido foi inserido na regressão de crista do pacote MORD porque a tarefa inicial de previsão de engajamento pode ser formulada como uma tarefa ordinal.

regressão. Os resultados das melhores tentativas em comparação com os resultados dos participantes do desafio EmotiW no o conjunto oficial de treinamento e validação é mostrado na Tabela 5.

É importante ressaltar que os melhores resultados normalmente são alcançados por modelos de conjunto que utilizam diversos recursos de áudio e vídeo diferentes. Portanto, os melhores resultados de modelos únicos são apresentados aqui para comparar francamente o métodos que usam apenas uma CNN. Apesar disso, o MSE 0,0563 para recursos do EfficientNet-B0 é o melhor quando comparado a qualquer método existente. A matriz de confusão do a melhor regressão ordinal é mostrada na Fig. 5. Seu MSE é menor do que o melhor modelo único [13] até 0,01 (15% relativo melhoria).

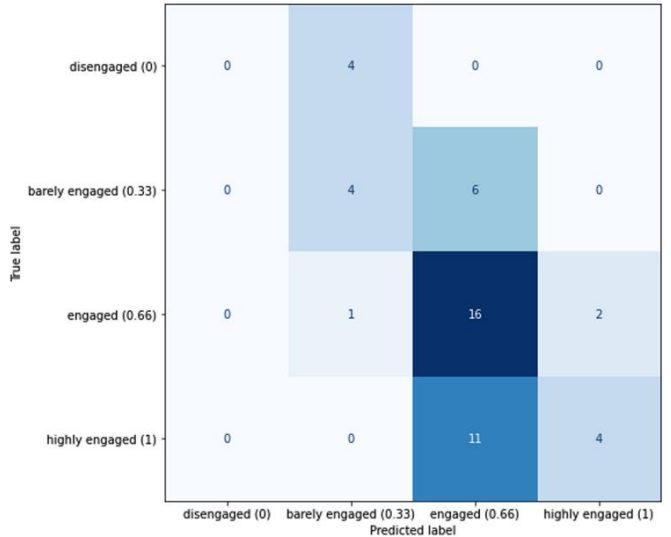
Contudo, este ponto deve ser esclarecido. Os participantes do desafio Emotion Engagement in the Wild verificado que alcançar melhores resultados no conjunto de validação não levar a excelente qualidade no conjunto de testes. Por exemplo, o vencedor do primeiro desafio (EmotiW 2018) tem alta

TABELA 5
MSE para o conjunto de validação EngageWild

Modelo	MSE
Atenção+multitaxa+híbrido [13]	0,0609
Ensemble 4 modelos para 2 características de comportamento [45]	0,0572
GAP+LBP-TOP [41]	0,0569
Linha de base (OpenFace + LSTM) [20]	0,10
LGCP [36]	0,0884
Bootstrap (OpenPose + LSTM) [44]	0,0717
LACUNA [41]	0,0671
TCN dilatado para unidades de ação [42]	0,0655
VGG [13]	0,0653
Nosso MobileNet, regressão de cume	0,0722
Nosso EfficientNet, regressão de cume	0,0563

MSE de validação de 0,0717 [43]. Assim, muitos pesquisadores [44], [45] ajustaram os hiperparâmetros de seus modelos em diferentes divisões de vídeos dos conjuntos unidos de treinamento e validação fornecido pelos autores do conjunto de dados EngageWild [8]. Por isso, no estudo de ablação foram utilizadas duas divisões do conjunto de dados: (1) a divisão oficial; e (2) a nova divisão aleatória com equilíbrio conjuntos de treinamento/validação, que têm o mesmo tamanho dos conjuntos da divisão original. Além da regressão ordinal de crista, foram testados modelos de regressão do scikit-learn, ou seja, floresta aleatória (RF), regressão vetorial de suporte (SVR) com núcleos lineares e RBF (função de base radial). Além disso, vários modelos de sequência foram implementados: (1) um GRU com 128 unidades e uma camada de saída totalmente conectada, e (2) atenção em nível de quadro único [56] com uma camada de saída densa. Esses modelos de sequência foram aplicados a dois entradas, ou seja, características iniciais de cada quadro extraído por desenvolveu a CNN emocional, e sua concatenação com um descritor único (desvio padrão por componente de recursos de quadro) usados nos modelos de regressão. O os resultados são apresentados na Tabela 6. A fonte em negrito indica o MSE mais baixo para cada divisão e a arquitetura da CNN.

Aqui a nova divisão causa resultados mais explicáveis. Para por exemplo, uma regressão de crista simples tem 0,01-0,02 menor MSE do que RF/SVR mais amplamente utilizado para a divisão original.



5. Matriz de confusão para previsão de engajamento, regressão ordinal de cume, recursos EfficientNet-B0.
Uso licenciado autorizado limitado a: Universidade Federal de Alagoas. Baixado em 28 de março de 2024 às 19:15:40 UTC do IEEE Xplore. Restrições aplicadas.

TABELA 6
Estudo de ablação dos modelos propostos, conjunto de dados EngageWild

		Validação MSE	
Nossa CNN	Classificador	Divisão original	Nossa divisão
MobileNet v1 GRU (somente quadro)	RF	0,0844	0,0511
	LinearSVR	0,0895	0,0588
	SVRRBF	0,0759	0,0526
	Regressão de cume	0,0722	0,0547
	GRU (quadro + padrão)	0,0981	0,0680
	Atenção (apenas quadro)	0,0970	0,0585
EfficientNet-B0 GRU (somente quadro)	Atenção (quadro + padrão)	0,0977	0,0618
	RF	0,0882	0,0530
	LinearSVR	0,0738	0,0540
	SVRRBF	0,0758	0,0560
	Regressão de cume	0,0778	0,0543
	GRU (quadro + padrão)	0,0563	0,0593
GRU EfficientNet-B2 (somente quadro)	Regressão de cume	0,0970	0,0668
	GRU (quadro + padrão)	0,0761	0,0445
	Atenção (apenas quadro)	0,0882	0,0626
	Atenção (quadro + padrão)	0,0682	0,0494
	RF	0,0882	0,0635
	LinearSVR	0,0897	0,0599
	SVRRBF	0,0868	0,0592
	Regressão de cume	0,0702	0,0642
	GRU (quadro + padrão)	0,1065	0,0777
	Atenção (apenas quadro)	0,0850	0,0672
	Atenção (quadro + padrão)	0,0997	0,0715
		0,0914	0,0652

No entanto, nossa divisão estratificada leva a resultados aproximadamente iguais MPEs. Além disso, os modelos de sequência têm significativamente melhor resultados. Por exemplo, o MSE mais baixo para o EfficientNet-B0 recursos foram obtidos usando atenção por quadro. Como um Como resultado, decidimos treinar os modelos para a aplicação de demonstração usando a nova divisão que parece ser mais consistente com estudos existentes de técnicas de regressão.

Notavelmente, o EfficientNet-B0 com menor número de parâmetros tem desempenho ligeiramente melhor quando comparado ao EfficientNet-B2, embora este último tenha uma emoção muito maior precisão de reconhecimento para imagens estáticas (Tabela 3). É importante mencionar que a concatenação dos recursos do quadro

com o desvio padrão das características do todo o vídeo funciona muito melhor na maioria dos casos, exceto nas redes MobileNet e GRU mais simples.

4.3 Reconhecimento de emoção baseado em vídeo

O reconhecimento de emoções baseado em vídeo é estudado em dois conjuntos de dados do desafio EmotiW [20]. Primeiro, o conjunto de dados AFEW com 773 trems e 383 amostras de validação foram examinadas. Isto contém pequenos cliques de áudio e vídeo coletados de filmes e Seriado de TV com expressões espontâneas, poses diversas e iluminações. Eles são rotulados usando um semi-automático abordagem. A tarefa é atribuir um único rótulo de emoção ao videoclipe das seis emoções universais (raiva, nojo, Medo, Felicidade, Tristeza e Surpresa) e Neutro.

O pré-processamento de todos os videoclipes do anterior subseção foi usada. No entanto, a média pontual, max, min e o desvio padrão de seus descritores de quadros foram concatenado [57]. Assim, a dimensionalidade do vídeo descritor é 4 vezes maior que a dimensionalidade D de o rosto incorporações emocionais. Se o rosto não foi encontrado em

TABELA 7
Precisão para o conjunto de validação AFEW

Modelo		Precisão, %
Conjunto, áudio + vídeo	Fusão bimodal de 4 CNNs [32]	54,3
	VGG13+VGG16+ResNet [18]	59,42
	5 modelos FBP [31]	65,5
Modelo único	LBP-TOP (linha de base) [20]	38,90
	Rede de atenção de quadro (FAN) [56]	51,18
	Aluno barulhento com treinamento iterativo [30]	55,17
	Nosso MobileNet-v1	55,35
	Nosso EficienteNet-B0	59,27
	Nosso EficienteNet-B2	59,00

o vídeo de treinamento, foi ignorado, mas os vídeos de validação com faces perdidas foram associados a zero descritores [28]. Os descritores normalizados em L2 foram classificados usando Line-arSVC do scikit-learn com parâmetros de regularização encontrado usando validação cruzada no conjunto de treinamento. A precisão da classificação é apresentada na Tabela 7. A maior precisão para modelos conjuntos e individuais está marcada em negrito.

Aqui os modelos propostos são até 5% mais precisos quando comparado a outros modelos individuais. Até o MobileNet é 0,1% mais preciso do que a técnica mais conhecida com o ResNet que foi treinado iterativamente como um aluno barulhento [30]. A matriz de confusão (Fig. 6) do melhor modelo Efficient-Net demonstra que embora muitas emoções sejam previsto com precisão, a precisão para pelo menos Nojo e As categorias de medo devem ser melhoradas. Embora o melhor conjuntos [31] ainda são muito mais precisos, nossa abordagem é muito mais rápido e pode ser implementado para processamento em tempo real das emoções de um aluno, mesmo em seu celular dispositivo.

A tarefa de reconhecimento de emoções em grupo de vídeo foi estudada usando o conjunto de dados VGAF [10] que contém vídeos de grupo baixado do YouTube com licença Creative Commons. Os dados têm muitas variações em termos de contexto, número

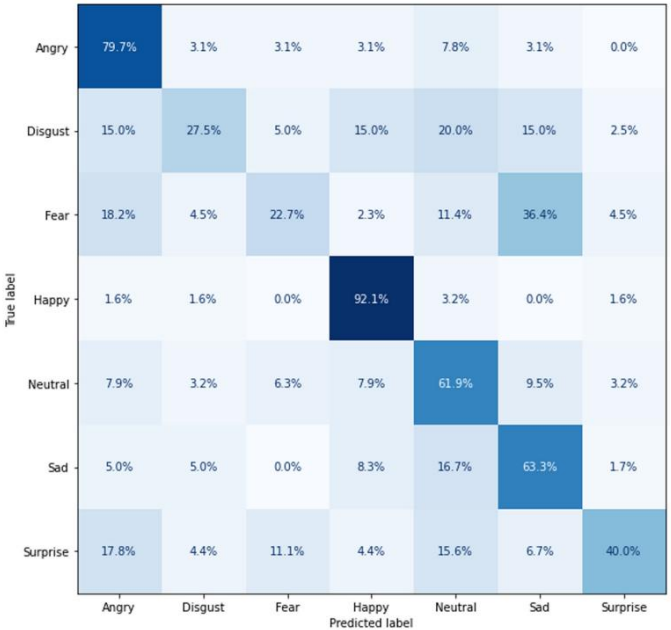


Figura 6. Matriz de confusão para reconhecimento de emoções individuais baseado em vídeo, Recursos do EfficientNet-B0.

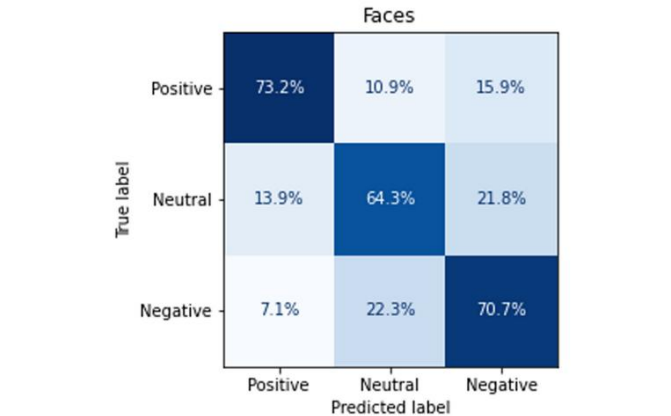
TABELA 8
Precisão para o conjunto de validação VGAF

Modelo		Precisão, %
Ensemble, VGAFNet (face + holístico + áudio) [10]	áudio+	61,61
	Vídeo de redes de injeção K [34]	66,19
	Fusão de 14 modelos [58]	71,93
Modelo único	Redes Híbridas [14]	74,28
	VGAFNet (rostos) [10]	60,18
	DenseNet-121 (redes híbridas) [14]	64,75
	Rede de injeção K de autoatenção [34]	65,01
	Lento e rápido [58]	68,57
	Nosso MobileNet-v1	68,92
	Nosso EficienteNet-B0	66,80
	Nosso EficienteNet-B2	70,23

de pessoas, qualidade de vídeo, etc. O conjunto de treinamento fornecido por os organizadores do desafio contém 2.661 clipes, enquanto 766 vídeos estão disponíveis para validação. A tarefa é classificar cada vídeo em 3 classes - positivo, neutro e negativo.

Como o quadro neste conjunto de dados geralmente contém vários regiões faciais, a média e o desvio padrão da face incorporações de emoção em cada quadro foram concatenadas para obter suas características. O descritor final de todo o vídeo foi calculado como uma concatenação da média e padrão desvio dos recursos do quadro. Os rostos desaparecidos em todos os vídeos foram processados de forma semelhante aos experimentos anteriores com o conjunto de dados AFEW: os vídeos vazios foram removidos o conjunto de treinamento, mas associado a zero descritores para comparar a precisão da validação com modelos existentes [28]. O os descritores de vídeo obtidos pelas CNNs propostas são classificados por SVC com kernel RBF. Os principais resultados são apresentados na Tabela 8.

A abordagem proposta leva novamente aos melhores resultados contra modelos únicos. O melhor EfficientNet-B2 melhora a mais conhecida análise de vídeo Slowfast [58] até 1,7%. Isso é matriz de confusão é apresentada na Fig. 7. Pode-se notar a comportamento natural deste classificador, ou seja, a pior precisão para a categoria Neutra, que normalmente é classificada erroneamente como afeto positivo ou negativo. Até o MobileNet é 0,35% mais preciso que o uso do Slowfast. Embora o grandes grupos de vencedores deste desafio [14] estão mais preciso, os modelos propostos ainda são competitivos mesmo com a fusão de 14 CNNs profundas aplicadas a vários recursos de áudio e vídeo [58].



7. Matriz de confusão para reconhecimento de emoções em nível de grupo baseado em vídeo, recursos do EfficientNet-B2.

5. CONCLUSÃO

Embora existam muitas detecções precisas de engajamento e técnicas de classificação de emoções [31], [41], seu uso no aprendizado on-line ainda é muito limitado porque a maioria dos alunos não desejam compartilhar seus vídeos faciais devido a questões de privacidade. É especialmente difícil obter permissão para tratamentos faciais análise de crianças. Para lidar com esta questão, o nova estrutura (Fig. 1) foi proposta neste artigo para análise baseada em vídeo do envolvimento emocional dos alunos. Esta estrutura pode ser integrada no e-learning existente ferramentas para avaliação rápida e precisa das emoções dos alunos e compreensão. Acreditamos que os principais usuários deste enquadramento serão os gestores e especialistas técnicos de plataformas on-line ou especialidades/graus on-line que precisam analisar o sucesso de seus cursos e encontrar os fatores-chave melhorar a qualidade dos cursos on-line, reduzir a rotatividade de clientes, etc. Os alunos precisam apenas ligar seus câmeras e professores precisam gravar o vídeo de seus webinars. Para salvar a privacidade do aluno, os vídeos faciais podem ser processado até mesmo em seu dispositivo móvel. A professora recebe apenas a média prevista de envolvimento e pontuações emocionais de todo o grupo. O neural desenvolvido os modelos exigem que as imagens faciais tenham resolução de 224x224 ou 300x300, para que a qualidade de qualquer câmera frontal seja adequada. Além disso, nossos experimentos preliminares demonstram que é possível extrair características emocionais de 1 quadro por segundo sem degradação significativa da precisão. Como um Como resultado, mesmo smartphones ou laptops muito baratos podem ser usado por um aluno com qualquer formação econômica.

O componente principal do pipeline proposto são os modelos neurais leves (Fig. 4) aprendidos usando a modificação robusta do otimizador Adam (Algoritmo 1). O melhor modelo supera os resultados do estado da arte conhecidos (Tabela 6) para previsão de 4 níveis de envolvimento (desenvolvido, pouco engajado, engajado e altamente engajado) no EngageWild conjunto de dados que contém vídeos faciais da observação do aluno vídeos educativos como os dos MOOCs [8]. Infelizmente, até onde sabemos, não existem informações publicamente disponíveis conjuntos de dados para reconhecimento de emoções do aluno. Portanto, alguns os experimentos foram feitos em conjuntos de dados que não pertencem para e-learning. Foi demonstrado que as representações faciais obtidos pelos modelos desenvolvidos podem ser usados para reconhecimento rápido e simultâneo de emoções individuais e de grupo. Nossa melhor CNN supera o estado da arte conhecido resultados de modelos únicos para reconhecimento de 8 emoções em fotos estáticas do AffectNet (Tabela 1), reconhecimento de 7 emoções básicas no conjunto de dados AFEW (Tabela 7) e classificação de 3 afetos (positivos, neutros e negativos) no VGAF conjunto de dados (Tabela 8). No entanto, afirmamos que as características emocionais propostas podem ser usadas para um reconhecimento preciso das emoções em outros conjuntos de dados, incluindo domínio de e-learning. Por exemplo, os mesmos modelos permitem que o primeiro autor deste artigo assuma o terceiro lugar na competição de aprendizagem multitarefa sobre Análise do Comportamento Afetivo na natureza (ABAW) [48].

É necessário mencionar que a abordagem proposta é menos preciso quando comparado ao multimodal mais conhecido conjuntos nos conjuntos de dados AFEW e VGAF. No entanto, já que os vídeos em grupo em sistemas de e-learning normalmente não contém a voz dos alunos e até a pose não é clara, muitas partes desses conjuntos são inúteis se for necessário

estimar a emoção geral de todo o grupo. Nisso

Neste caso, a análise das regiões faciais é a mais preferível para reconhecimento de emoções em aulas on-line porque leva a um precisão de classificação de emoções muito maior quando comparada ao desempenho de modelos únicos existentes.

No futuro, será necessário utilizar dados de vídeo adicionais para melhorar a qualidade do mecanismo de previsão de engajamento, que agora é limitado devido ao uso de um pequeno conjunto de treinamento. Em segundo lugar, como a resolução dos rostos detectados ainda pode ser baixa, vamos estudar as técnicas conhecidas de FER de baixa resolução [49] para melhorar a qualidade da detecção de engajamento e emoção reconhecimento modificando os modelos propostos que foram treinado em imagens faciais de 224x224 e 300x300. Terceiro, pode É importante prever a excitação e a valência [51], [59] além das expressões faciais para dar ao professor informações adicionais sobre a atitude de cada aluno. Outra pesquisa

direção é a melhoria do agrupamento de rostos aplicando técnicas de detecção e reconhecimento de texto para obter o nome de cada participante da videoconferência e tratamento facial em grupo regiões localizadas próximas ao nome detectado. Finalmente, é necessário examinar o potencial da técnica proposta e dos recursos emocionais em outras tarefas de e-learning, como supervisão on-line.

REFERÊNCIAS

- [1] X. Wang, T. Liu, J. Wang e J. Tian, "Compreendendo a intenção de continuidade do aluno: uma comparação entre aprendizagem por vídeo ao vivo, aprendizagem por vídeo pré-gravada e aprendizagem por vídeo híbrida em COVID-19 pandemia", *Int. J. Hum.-Computação. Interagir.*, vol. 38, não. 3, pp. 2022.
- [2] J. Shen, H. Yang, J. Li e Z. Cheng, "Avaliando o envolvimento na aprendizagem com base no reconhecimento de expressões faciais no cenário MOOC," *Sistema Multimídia*, vol. 28, pp. 469–478, 2022.
- [3] Tomczyk, K. Potyra»a, N. Demeshkant e K. Czerwicz, "Professores universitários e e-learning em crise: resultados de um piloto polaco estudo sobre: Atitudes em relação ao e-learning, experiências com e-learning e antecipação do uso de soluções de e-learning após a pandemia", em *Proc. IEEE 16ª Conferência Ibérica. Inf. Sist. Tecnologia*, 2021, pp.
- [4] P. Bhardwaj, P. Gupta, H. Panwar, MK Siddiqui, R. Morales-Menendez e A. Bhaik, "Aplicação de aprendizagem profunda no envolvimento do aluno em ambientes de e-learning", *Comput. Eleger. Eng.*, vol. 93, 2021, art. não. 107277.
- [5] M. Sathik e SG Jonathan, "Efeito das expressões faciais em reconhecimento da compreensão do aluno em ambientes educacionais virtuais", *SpringerPlus*, vol. 2, não. 1, pp. 1–9, 2013.
- [6] MAA Dewan, M. Murshed e F. Lin, "Detecção de engajamento na aprendizagem on-line: uma revisão", *Smart Learn. Meio Ambiente*, vol. 6, não. 1, pp. 1–20, 2019.
- [7] J. Bacca-Acosta e C. Avila-Garzon, "Envolvimento dos alunos com sistemas de avaliação baseados em dispositivos móveis: uma análise de sobrevivência", *J. Comput. Ajude. Aprenda.*, vol. 37, não. 158–171, 2021.
- [8] A. Kaur, A. Mustafa, L. Mehta e A. Dhali, "Predição e localização do envolvimento dos alunos na natureza", em *Proc. IEEE Dígito. Computação de imagem: Techn. Appl.*, 2018, pp.
- [9] M. Imani e GA Montazer, "Uma pesquisa de reconhecimento de emoções métodos com ênfase em ambientes de E-learning", *J. Netw. Computação. Ap.*, vol. 147, 2019, art. não. 102423.
- [10] G. Sharma, A. Dhali e J. Cai, "Grupo audiovisual automático análise de impacto", *IEEE Trans. Computação Afetiva.*, a ser publicado, doi: [10.1109/TAFFC.2021.3104170](https://doi.org/10.1109/TAFFC.2021.3104170).
- [11] I. Goodfellow, Y. Bengio e A. Courville, *Aprendizado Profundo*. Cambridge, MA, EUA: MIT Press, 2016.
- [12] T. Liu, J. Wang, B. Yang e X. Wang, "NGDNet: Não uniforme Aprendizagem de distribuição de rótulo gaussiano para estimativa de pose de cabeça infravermelha e compreensão do comportamento na tarefa na sala de aula", *Neurocomputação*, vol. 436, pp.
- [13] B. Zhu, X. Lan, X. Guo, KE Barner e C. Boncelet, "Multi-taxa modelo GRU baseado em atenção para previsão de engajamento", em *Proc. Internacional Conf. Interação Multimodal.*, 2020, pp.

- [14] C. Liu, W. Jiang, M. Wang e T. Tang, "Reconhecimento de emoções de áudio e vídeo em nível de grupo usando redes híbridas", em Proc. Internacional Conf. Interação Multimodal., 2020, pp.
- [15] T. Liu, J. Wang, B. Yang e X. Wang, "Método de reconhecimento de expressão facial com aprendizagem de distribuição multi-rótulo para compreensão do comportamento não verbal na sala de aula", *Infrared Phys. Tecnologia*, vol. 112, 2021, art. não. 103594.
- [16] AV Savchenko, KV Demochkin e IS Grechikhin, "Previsão de preferência baseada em uma análise de galeria de fotos com reconhecimento de cena e detecção de objetos", *Pattern Recognit.*, vol. 121, 2022, art. não. 108248.
- [17] P. Xanthopoulos, PM Pardalos e TB Trafalis, *Robust Data Mining*. Berlim, Alemanha: Springer, 2012.
- [18] SA Bargal, E. Barsoum, CC Ferrer e C. Zhang, "Reconhecimento de emoções na natureza a partir de vídeos usando imagens", em Proc. Internacional Conf. Interação Multimodal., 2016, pp.
- [19] H. Zeng et al., "EmotionCues: Resumo visual orientado para a emoção de vídeos de sala de aula", *IEEE Trans. Vis. Computação. Gráficos*, vol. 27, não. 7, pp. 3168–3181, julho de 2021.
- [20] A. Dhali, "EmotiW 2019: Tarefas automáticas de previsão de emoção, engajamento e coesão", em Proc. Internacional Conf. Interação Multimodal., 2019, pp.
- [21] T. Ashwin e RMR Guddeti, "Banco de dados afetivo para e-learning e ambientes de sala de aula usando rostos, gestos de mãos e posturas corporais de estudantes indianos", *Future Gener. Computação. Sistema*, vol. 334–348, 2020.
- [22] BE Zakka e H. Vadapalli, "Estimando o efeito da aprendizagem do aluno usando emoções faciais", em Proc. IEEE 2º Int. Inf. Multidisciplinar. *Tecnologia. Eng. Conf.*, 2020, pp.
- [23] A. Mollahosseini, B. Hasani e MH Mahoor, "AffectNet: Um banco de dados para expressão facial, valência e computação de excitação na natureza", *IEEE Trans. Computação Afetiva*, vol. 10, não. 1, pp. 18–31, janeiro–março. 2019.
- [24] M. Pourmirzaei, GA Montazer e F. Esmaili, "Usando tarefas auxiliares auto-supervisionadas para melhorar a representação facial refinada", 2021, arXiv:2105.06421.
- [25] M. Tan e Q. Le, "EfficientNet: Repensando o escalonamento de modelos para redes neurais convolucionais", em Proc. Internacional Conf. Mach. Aprenda., 2019, pp.
- [26] F. Ma, B. Sun e S. Li, "Reconhecimento de expressão facial com transformadores visuais e fusão seletiva de atenção", *IEEE Trans. Computação Afetiva.*, a ser publicado, doi: [10.1109/TAFFC.2021.3122146](https://doi.org/10.1109/TAFFC.2021.3122146).
- [27] F. Xue, Q. Wang e G. Guo, "TransFER: Aprendendo representações de expressões faciais com reconhecimento de relação com transformadores", em Proc. Internacional IEEE/CVF. Conf. Computação. Vis., 2021, pp.
- [28] AV Savchenko, "Expressão facial e reconhecimento de atributos com base na aprendizagem multitarefa de redes neurais leves", em Proc. IEEE 19º Int. Simp. Intel. Sist. Informar., 2021, pp.
- [29] P. Antoniadis, PP Filintisis e P. Maragos, "Explorando dependências emocionais com redes convolucionais de grafos para reconhecimento de expressões faciais", em Proc. 16º Int. IEEE. Conf. Automático. Reconhecimento de gestos faciais, 2021, pp.
- [30] V. Kumar, S. Rao e L. Yu, "Treinamento barulhento de alunos usando conjunto de dados de linguagem corporal melhora o reconhecimento de expressões faciais", em Proc. EUR. Conf. Computação. Vis., 2020, pp.
- [31] H. Zhou et al., "Explorando recursos emocionais e estratégias de fusão para reconhecimento de emoções de áudio e vídeo", em Proc. Internacional Conf. Interação multimodal., 2019, pp.
- [32] S. Li et al., "Fusão de bimodalidade para reconhecimento de emoções na natureza", em Proc. Internacional Conf. Interação Multimodal., 2019, pp.
- [33] JR Pinto et al., "Classificação audiovisual da valência emocional do grupo usando redes de reconhecimento de atividade", em Proc. IEEE 4º Int. Conf. Processo de imagem. Apl. Sistema, 2020, pp.
- [34] Y. Wang, J. Wu, P. Heraclous, S. Wada, R. Kimura e S. Kuri-hara, "Modelo audiovisual de atenção cruzada injetável de conhecimento implícito para reconhecimento de emoções de grupo", em Proc. Internacional Conf. Interação multimodal., 2020, pp.
- [35] IP Ratnapala, RG Ragel e S. Deegalla, "Análise comportamental dos alunos em um ambiente de aprendizagem online usando mineração de dados", em Proc. IEEE 7º Int. Conf. Inf. Automático. Sustentabilidade, 2014, pp.
- [36] Z. Zhang, Z. Li, H. Liu, T. Cao e S. Liu, "Detecção de envolvimento de aprendizagem online baseada em dados por meio de expressão facial e tecnologia de reconhecimento de comportamento do mouse", *J. Educ. Computação. Res.*, vol. 58, não. 1, pp. 63–86, 2020.
- [37] T. Dragon, I. Arroyo, BP Woolf, W. Burleson, RE Kaliouby e H. Eydgahi, "Visualizando o afeto e a aprendizagem dos alunos por meio da observação em sala de aula e sensores físicos", em Proc. Internacional Conf. Intel. Sistema de tutoria, 2008, pp.
- [38] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster e JR Movellan, "As faces do envolvimento: Reconhecimento automático do envolvimento dos alunos a partir de expressões faciais", *IEEE Trans. Com-put Afetivo.*, vol. 5, não. 1, pp. 86–98, janeiro–março. 2014.
- [39] X. Chen, L. Niu, A. Veeraraghavan e A. Sabharwal, "FaceEngage: estimativa robusta de envolvimento de jogo a partir de vídeos contribuídos por usuários (YouTube)", *IEEE Trans. Com-put Afetivo.*, vol. 13, não. 2, pp. 651–665, abril–junho. 2022.
- [40] T. Baltrusaitis, A. Zadeh, YC Lim e L.-P. Morency, "OpenFace 2.0: Kit de ferramentas de análise de comportamento facial", em Proc. IEEE 13º Int. Conf. Automático. Reconhecimento de gestos faciais, 2018, pp.
- [41] X. Niu et al., "Previsão automática de engajamento com recurso GAP", em Proc. Internacional Conf. Interação Multimodal., 2018, pp.
- [42] C. Thomas, N. Nair e DB Jayagopi, "Prevenção da intensidade do engajamento na natureza usando rede convolucional temporal", em Proc. Internacional Conf. Interação Multimodal., 2018, pp.
- [43] J. Yang, K. Wang, X. Peng e Y. Qiao, "Aprendizagem profunda recorrente de múltiplas instâncias com recursos espaço-temporais para previsão de intensidade de engajamento", em Proc. Internacional Conf. Interação Multimodal., 2018, pp.
- [44] K. Wang, J. Yang, D. Guo, K. Zhang, X. Peng e Y. Qiao, "Conjunto de modelos Bootstrap e perda de classificação para regressão de intensidade de engajamento", em Proc. Internacional Conf. Interação Multimodal., 2019, pp.
- [45] J. Wu, Z. Zhou, Y. Wang, Y. Li, X. Xu e Y. Uchida, "Aprendizagem de múltiplos recursos e múltiplas instâncias com estratégia anti-overfitting para previsão de intensidade de engajamento," em Proc. Internacional Conf. Interação Multimodal., 2019, pp.
- [46] VT Huynh, S.-H. Kim, G.-S. Lee e H.-Y. Yang, "Previsão da intensidade do envolvimento com características de comportamento facial", em Proc. Internacional Conf. Interação Multimodal., 2019, pp.
- [47] Q. Cao, L. Shen, W. Xie, OM Parkhi e A. Zisserman, "VGGFace2: Um conjunto de dados para reconhecer rostos através de pose e idade", em Proc. IEEE 13º Int. Conf. Automático. Reconhecimento de gestos faciais, 2018, pp.
- [48] AV Savchenko, "Análise facial em nível de quadro baseada em vídeo do comportamento afetivo em dispositivos móveis usando EfficientNets", em Proc. Conferência IEEE/CVF. Computação. Vis. Reconhecimento de padrões. Oficinas, 2022, pp.
- [49] Y. Yan, Z. Zhang, S. Chen e H. Wang, "Reconhecimento de expressão facial de baixa resolução: Uma perspectiva de aprendizagem de filtro", *Signal Process.*, vol. 169, 2020, art. não. 107370.
- [50] AV Savchenko, "Representações faciais eficientes para idade, gênero e reconhecimento de identidade na organização de álbuns de fotos usando ConvNet de múltiplas saídas", *PeerJ Comput. Ciência*, vol. 5º de 2019, art. não. e197.
- [51] JA Russell, LM Ward e G. Pratt, "Qualidade afetiva atribuída a ambientes: Um estudo analítico de fator", *Environ. Comportamento*, vol. 13, não. 3, pp. 259–288, 1981.
- [52] DP Kingma e J. Ba, "Adam: Um método para otimização estocástica mização", 2014, arXiv:1412.6980.
- [53] AH Farzaneh e X. Qi, "Reconhecimento de expressão facial na natureza por meio da perda profunda do centro de atenção", em Proc. Conferência de Inverno IEEE/CVF. Apl. Computação. Vis., 2021, pp.
- [54] L. Schoneveld, A. Othmani e H. Abdelkawy, "Aproveitando os avanços recentes na aprendizagem profunda para o reconhecimento de emoções audiovisuais", *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021.
- [55] Z. Wen, W. Lin, T. Wang e G. Xu, "Distraia sua atenção: rede de atenção cruzada com várias cabeças para reconhecimento de expressão facial", 2021, arXiv:2109.07270.
- [56] D. Meng, X. Peng, K. Wang e Y. Qiao, "Redes de atenção de quadro para reconhecimento de expressão facial em vídeos", em Proc. Internacional IEEE. Conf. Processo de imagem., 2019, pp.
- [57] P. Demochkina e AV Savchenko, "MobileEmotiFace: representações eficientes de imagens faciais no reconhecimento de emoções baseado em vídeo em dispositivos móveis", em Proc. Reconhecimento de padrões. Internacional Desafio de Workshops, 2021, pp.
- [58] M. Sun et al., "Fusão multimodal usando recursos espaço-temporais e estáticos para reconhecimento de emoções de grupo", em Proc. Internacional Conf. Interação Multimodal., 2020, pp.
- [59] V. Skaramagkas et al., "Uma abordagem de aprendizado de máquina para prever a excitação emocional e a valência dos recursos extraídos do olhar", em Proc. IEEE 21º Int. Conf. Bioinf. Bioeng., 2021, pp.



Andrey V. Savchenko recebeu o diploma de bacharelado em matemática aplicada e informática de Nizhny Universidade Técnica Estadual de Novgorod, Nizhny Novgorod, Rússia, em 2006, o título de PhD em modelagem matemática e ciência da computação pela Escola Superior de Economia da Universidade Estadual, Moscou, Rússia, em 2010, e o grau de DrSc em análise de sistema e processamento de informações de Universidade Técnica Estadual de Nizhny Novgorod, em 2016. Desde 2008, ele trabalha na HSE University, Nizhny Novgorod, onde atualmente é

professor titular do Departamento de Sistemas e Tecnologias de Informação. Ele também é pesquisador líder do Laboratório de Algoritmos e Tecnologias para Análise de Redes da Universidade HSE. Ele tem

é autor ou coautor de uma monografia e mais de 50 artigos. Dele os interesses de pesquisa atuais incluem reconhecimento estatístico de padrões, imagem classificação e biometria.



Lyudmila V. Savchenko recebeu o doutorado licenciatura em análise de sistemas e processamento de informações pela Voronezh State Technical University, em 2017, e o grau de especialista em matemática e informática pela Nizhny Nov-gorod State Technical University, Nizhny Nov-gorod, Rússia, em 2008. Desde 2018, ela tem esteve na Universidade HSE, Nizhny Novgorod, onde atualmente é professora associada com o Departamento de Sistemas de Informação e Tecnologias. Ela também é pesquisadora sênior do

Laboratório de Algoritmos e Tecnologias para Redes.

Análise, Universidade HSE. Seus atuais interesses de pesquisa incluem discurso sistemas de processamento e e-learning.



Ilya Makarov recebeu o título de PhD em ciência da computação pela Universidade de Ljubljana, Liubliana, Eslovênia. De 2011 a 2022, ele foi professor em tempo integral na HSE University, Escola de Análise de Dados e Inteligência Artificial. Ele é pesquisador sênior da AIRI, HSE University – Nizhniy Novgorod, e pesquisador da Centro Conjunto de IA Samsung-PDMI, São Petersburgo Departamento do Instituto Steklov de Matemática, Academia Russa de Ciências, São Petersburgo, Rússia. Sua carreira educacional em ciência de dados

cobre cargos de diretor de programa da BigData Academy MADE de VK, professor sênior do Instituto de Física e Tecnologia de Moscou, e engenheiro de aprendizado de máquina e chefe do Data Science Tech Master programa em PNL, Universidade Nacional de Ciência e Tecnologia MISIS.

" Para obter mais informações sobre este ou qualquer outro tópico de computação, visite nossa Biblioteca Digital em www.computer.org/csdl.