IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Recebido em 21 de junho de 2023, aceito em 10 de julho de 2023, data de publicação em 21 de julho de 2023, data da versão atual em 28 de julho de 2023.

Identificador de Objeto Digital 10.1109/ACCESS.2023.3297651

APPLIED RESEARCH

Análise de Expressões Faciais para Estimar o Nível de envolvimento em palestras online

MIAO YOSHIYUKI [1], HARUKA KATO1, YASUHIRO HATORI1,2 RENJUN, SATO 2,3, E SATOSHI SHIOIRI1,2,3 1Escola de Pós-Graduação em Ciências da Informação, Universidade de Tohoku, Sendai, Miyagi 980-8577, Japão
2Instituto de Pesquisa de Comunicação Elétrica, Universidade de Tohoku, Sendai, Miyagi 980- 8577, Japão 3Advanced Institute for Yotta Informatics, Tohoku University, Sendai, Miyagi 980-8577, Japão Autor correspondente: Renjun Miao (miao.renjun.s1@dc.tohoku.ac.jp)

Este trabalho foi apoiado em parte pelo Programa de Projetos de Pesquisa do Centro de Pesquisa para Tecnologia da Informação do Século 21 (Centro IT-21), Instituto de Pesquisa de Comunicação Elétrica (RIEC), Universidade de Tohoku; e em parte pelo Projeto Yotta de Informática do Ministério da Educação, Cultura, Esportes, Ciência e Tecnologia (MEXT), Japão. O trabalho de Satoshi Shioiri foi apoiado pela Sociedade Japonesa para a Promoção da Ciência (JSPS) KAKENHI sob a concessão 19H01111.

RESUMO O presente estudo teve como objetivo desenvolver um método para estimar o estado de atenção dos alunos a partir de expressões faciais durante aulas on-line. Estimamos o nível de atenção enquanto os alunos assistiam a uma videoaula medindo o tempo de reação (TR) para detectar um som alvo que fosse irrelevante para a aula.
Assumimos que o TR para tal estímulo seria mais longo quando os participantes estavam concentrados na palestra, em comparação com quando não estavam. Procuramos estimar o quanto os alunos se concentram em uma palestra usando a medição de RT. No experimento, o rosto do aluno foi gravado por uma câmera de vídeo enquanto assistia a uma vídeo-aula. As características faciais foram analisadas para prever o RT para um estímulo irrelevante para a tarefa, que foi assumido como um índice do nível de atenção. Aplicamos um método de aprendizado de máquina, light Gradient Boosting Machine (LightGBM), para estimar RTs de características faciais extraídas como unidades de ação (AUs) correspondentes a movimentos musculares faciais por um software de código aberto (OpenFace). O modelo obtido usando LightGBM indicou que os RTs para estímulos irrelevantes podem ser estimados a partir de AUs, sugerindo que as expressões faciais são úteis para prever estados de atenção enquanto assiste a palestras. Reanalisamos os dados enquanto excluímos os dados do TR com rostos sonolentos dos alunos para testar se a diminuição da excitação geral causada pela sonolência era um fator significativo no prolongamento do TR observado no experimento. Os resultados foram semelhantes independentemente da inclusão de RTs com rostos sonolentos, indicando que a expressão facial pode ser usada para prever o nível de atenção dos alunos às videoaulas.

TERMOS DE INDEXAÇÃO Atenção, computação afetiva, engajamento, traços faciais, palestra online.

I. INTRODUÇÃO

Compreender os níveis de envolvimento dos alunos durante o estudo é importante para melhorar os resultados da aprendizagem. Para melhorar a qualidade da educação, é crucial estimar o nível de envolvimento dos alunos nos seus estudos. No entanto, é difícil para os professores prestar atenção a todos os alunos, principalmente nas aulas online. A medição automatizada dos níveis de envolvimento pode ser útil para melhorar as condições de aprendizagem. Para aprendizagem online, webcams podem ser usadas para capturar expressões faciais dos alunos, que podem ser usadas para estimar seus estados mentais[1], [2], [3]. Por exemplo, Shioiri et al. imagem conduzida estimativa de preferência a partir de expressões faciais e descobriu que esta informação era útil para estimar julgamentos subjetivos de preferência de imagem. Em estudos relacionados à educação, Thomas e Jayagopi gravaram imagens faciais dos alunos em uma sala de aula enquanto estudavam com material de vídeo em uma tela e estimaram o nível de envolvimento a partir das expressões faciais dos alunos [4], [5]. Os autores conseguiram prever o envolvimento, sugerindo a utilidade das expressões faciais para estimar o nível de envolvimento. A frequência cardíaca também tem sido usada para estimar estados mentais durante a aprendizagem. Darnell e Krieg mostraram que as mudanças na frequência cardíaca estão relacionadas à atividade dos alunos durante uma aula [6]. Embora estudos anteriores tenham focado no engajamento, que é avaliado externamente, esta pesquisa também foi estendida para

O editor associado que coordena a revisão deste manuscrito e quem a aprovou para publicação foi Filbert Juwono.

a medição de estados internos, que pode ser investigada estimando estados internos. Nestes estudos, o estado mental usado como verdade fundamental é baseado em julgamentos subjetivos [4], [7]. Contudo, os estados mentais envolvem outros fatores além daqueles que podem ser avaliados subjetivamente. Os processos inconscientes, que não podem ser estimados subjetivamente, podem desempenhar papéis mais importantes do que os processos conscientes. Assim, é improvável que julgamentos subjetivos sejam adequados para uso como índices de estados mentais. Por exemplo, a alteração da frequência cardíaca é considerada um índice útil da atividade dos alunos e não está necessariamente relacionada com a estimativa subjetiva de atenção e envolvimento [6]. Como tal, é importante desenvolver métodos que envolvam medidas objetivas para estimar o nível de envolvimento. Um estudo anterior mostrou que as características faciais podem ser úteis para estimar o tempo de reação (TR) para cálculos mentais [8]. Este resultado sugere que o RT pode ser um bom índice de atenção se variar dependendo do foco na tarefa, como normalmente assumido em estudos de atenção para simples detecção, discriminação ou identificação de estímulos visuais.

Porém, esse tipo de medida não está disponível para palestras. Portanto, tentamos usar RT para estímulos irrelevantes para a tarefa.

Embora engajamento seja um termo usado com diferentes significados em diferentes contextos [7], [9], ele é frequentemente usado em relação à atenção [10], [11], [12], [13]. Acredita-se que a atenção às palestras, aulas e tarefas esteja intimamente relacionada ao engajamento. Aqui usamos o termo atenção para nos referirmos à facilitação do processamento sensorial pela intenção endógena ou estimulação exógena saliente, e consideramos isso um fator importante para o envolvimento. Deve-se notar que o engajamento também tem sido usado para indicar estados mentais de maior duração em alguns estudos anteriores, como uma palestra inteira [6], [7], [14], [15], [16]. Medimos os níveis de atenção como um índice de engajamento durante as palestras neste estudo.

Projetamos um experimento no qual os participantes foram solicitados a detectar um alvo auditivo enquanto assistiam a um vídeo de palestra. A tarefa principal do experimento era compreender a palestra, e a tarefa secundária era detectar o alvo.
O TR para o alvo auditivo foi utilizado como medida objetiva do nível de atenção nos vídeos das palestras. Aqui, assumimos que o tempo necessário para detectar um alvo que era irrelevante para a tarefa primária seria maior quando o participante se concentrasse mais na tarefa primária (ou seja, assistindo a aulas em vídeo neste experimento). Imagens faciais dos participantes foram gravadas enquanto assistiam aos vídeos, e as expressões faciais foram analisadas após o experimento. O objetivo do estudo foi estimar o RT das expressões faciais para desenvolver um método para estimar o nível de envolvimento a partir das imagens faciais dos alunos.

Alguns dos resultados deste estudo com um número menor de participantes foram publicados em um livro pós-conferência como um relatório preliminar [14]. Aqui, relatamos análises de expressões faciais mais detalhadamente com dados de um número maior de participantes para considerar as contribuições de características faciais específicas, o efeito da variação individual e o efeito do nível geral de excitação ou sonolência.

II. EXPERIMENTO

Realizamos um experimento para investigar a relação entre o nível de atenção e a expressão facial enquanto assistimos a aulas em vídeo. Para estimar o nível de atenção em videoaulas, medimos o RT para um alvo auditivo irrelevante para a palestra. Assumimos que o TR para um estímulo irrelevante seria mais longo quando os participantes estavam concentrados na palestra, em comparação com quando não estavam. Foi sugerido que o efeito nas respostas cerebrais a estímulos irrelevantes é capaz de estimar a atenção à tarefa primária. Por exemplo, Kramer et al. realizaram medições de eletroencefalografia (EEG) e relataram que a resposta relacionada a eventos (ERP) a um estímulo irrelevante para a tarefa muda com a dificuldade de uma tarefa primária [18]. Mudanças semelhantes eram esperadas com as medições de RT porque tanto o ERP quanto o RT foram usados para estimar a atenção em geral [19]. No presente estudo, tentamos usar imagens faciais gravadas para prever o RT.

O alvo auditivo que usamos foi o desaparecimento do ruído branco contínuo em vez do aparecimento de um estímulo sonoro, enquanto experimentos anteriores para medir a atenção normalmente usavam um estímulo de pulso [20], [21]. A razão para usar o desaparecimento do som foi evitar a influência da atenção de baixo para cima sobre um estímulo saliente, como um pulso auditivo. A atenção de baixo para cima a um estímulo saliente pode ser forte o suficiente para mascarar o efeito da atenção à palestra. Na verdade, o efeito da atenção de cima para baixo não pode ser detectado quando há apenas uma estimulação transitória, enquanto um alvo é discriminado pela atenção de cima para baixo entre muitos estímulos transitórios [22], [23].

Quinze participantes (idade média de 23,1 anos) participaram do experimento. Os participantes tinham visão normal ou corrigida para normal e audição normal. Os participantes foram orientados a assistir a uma série de nove vídeo-aulas e a responder perguntas ao final de cada vídeo (fig. 1).
Os participantes também foram orientados a pressionar uma tecla ao perceberem o alvo auditivo (o súbito desaparecimento do ruído branco) enquanto assistiam à vídeo-aula. Os participantes foram instruídos de que a palestra era a tarefa principal do experimento, enquanto a detecção do alvo era uma tarefa secundária, e eles eram obrigados a responder perguntas no final do experimento. O TR do alvo foi medido para estimar o nível de atenção dos participantes no momento da apresentação do alvo.

Os materiais de aprendizagem eram de um curso introdutório sobre uma linguagem de computador, PHP, que foi postado no YouTube [24]. Os vídeos foram exibidos em uma tela de computador (MacBook Pro, Apple, Califórnia) com fones de ouvido (MDR-7506, Sony, Tóquio) em uma sala com iluminação de escritório (483,2 lx na mesa onde o computador foi colocado e 211,8 lx na a localização do rosto do participante). O volume médio da voz do palestrante foi de 70 db e o do ruído branco foi de 0,66 db.

O ruído branco desaparecia ocasionalmente, sendo o alvo da tarefa secundária. O intervalo entre dois

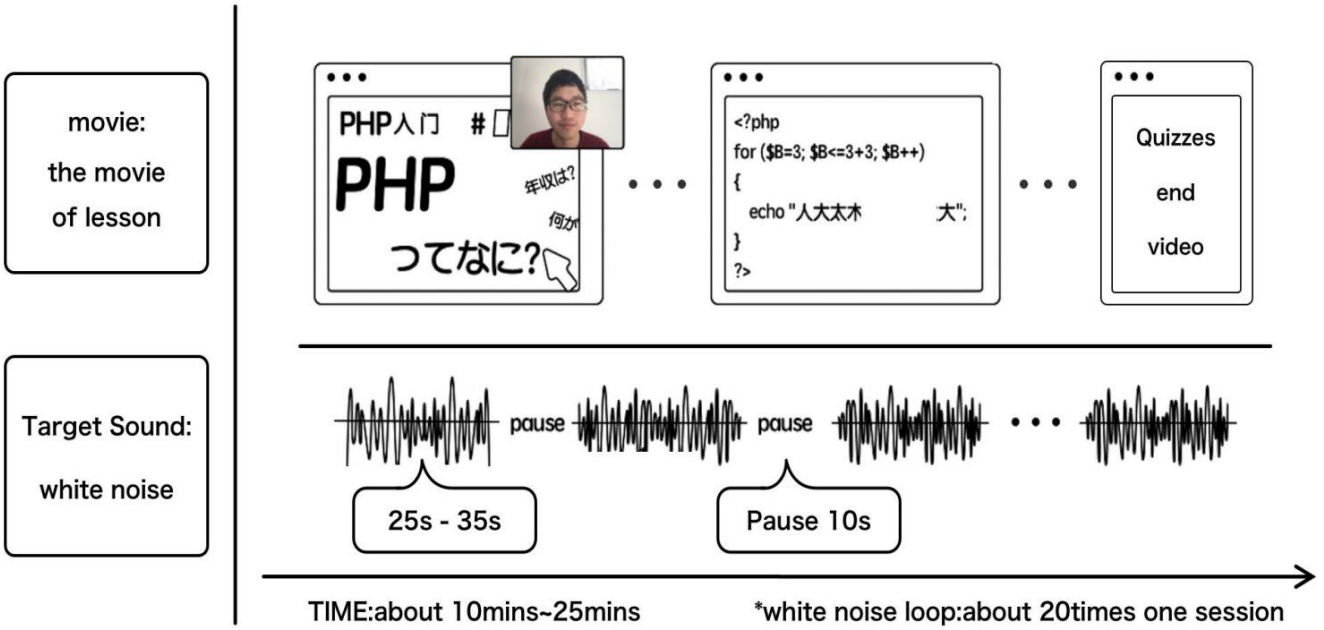

FIGURA 1. Desenho experimental. Enquanto assistia a uma aula em vídeo de um curso introdutório de PHP em uma sessão, sinais auditivos de ruído branco foram adicionados à trilha auditiva original do vídeo do palestrante. Ao final da sessão, após assistir ao vídeo, os participantes responderam a diversos questionários sobre a palestra.

os alvos, que era um período de apresentação de ruído branco, foram selecionados aleatoriamente entre 25 e 35 segundos. O ruído branco começou novamente imediatamente após o pressionamento da tecla para indicar a detecção, ou após um período de 10 segundos se nenhuma tecla tivesse sido pressionada. Cada palestra durou entre 10 e 20 minutos, dependendo do conteúdo. Ao final de cada palestra, foram disponibilizadas oito questões em formato de formulário google. Para cada pergunta, os participantes selecionaram uma das quatro opções como resposta. Assistir a uma palestra é uma sessão do experimento. Foram nove palestras.

Além das sessões expositivas, ocorreram duas sessões de controle para medir o TR para a tarefa de detecção sem prestar atenção à palestra, de modo que o número total de sessões foi onze. Nas sessões controle, foram utilizados dois vídeos das mesmas videoaulas para que os participantes conhecessem o conteúdo e tivessem pouco ou nenhum motivo para se sentirem atraídos pelo conteúdo. Os participantes foram solicitados a focar no ruído branco e informados de que não precisavam prestar atenção ao conteúdo do vídeo na tela. A primeira sessão de controle foi conduzida como a 6ª sessão com o primeiro vídeo da aula, e a segunda sessão de controle foi conduzida como a 11ª sessão com o 6º vídeo da aula usado na 7ª sessão. O experimento foi conduzido durante 2 dias. Cinco sessões teóricas e a primeira sessão controle foram realizadas no primeiro dia, e o restante das sessões (quatro palestras e uma sessão controle) foram realizadas no segundo dia.

O intervalo entre o primeiro e o segundo dia foi de 1 semana. A duração total do experimento, 11 sessões, foi de aproximadamente 130 minutos.

III. ANÁLISE DE CARACTERÍSTICAS FACIAIS

Os rostos dos participantes foram gravados enquanto assistiam a vídeos de palestras, e suas características faciais foram analisadas após o experimento. Analisamos imagens faciais gravadas 3 segundos antes da apresentação do alvo (desaparecimento do ruído branco), utilizando OpenFace [25] para extrair as características faciais.

Para realizar análises de expressões faciais usando OpenFace, o primeiro passo é coletar imagens faciais ou dados de vídeo. A partir de cada quadro de vídeo, o OpenFace detecta um rosto (vários rostos podem ser detectados enquanto havia apenas um rosto em nosso experimento) e o localiza no quadro. Em seguida, ele faz a aparência facial como orientação facial e faz pontos de referência faciais, como

limites dos olhos, sobrancelhas e boca. Ao analisar as mudanças de posição dos pontos de referência faciais e da aparência facial, o OpenFace avalia o grau de atividade muscular facial como unidades de ação (AUs). AUs são atribuídos a movimentos musculares relacionados a expressões faciais com base no Facial Action Coding System (FACS) [26]. Por exemplo, AU1 indica elevação da parte interna das sobrancelhas, AU4 indica abaixamento das sobrancelhas e AU5 indica elevação das pálpebras superiores (Tabela 1). OpenFace oferece diversas vantagens de pesquisa para análise facial. Em primeiro lugar, aproveitando técnicas de aprendizagem profunda, particularmente redes neurais convolucionais (CNNs), o OpenFace alcança alta precisão em tarefas de reconhecimento facial e extração de características. Isto é crucial para projetos de pesquisa que exigem identificação e comparação precisas de características faciais. Em segundo lugar, o OpenFace não só permite o reconhecimento facial, mas também facilita a extração de características faciais, como expressões e poses. Isso amplia suas aplicações em áreas de pesquisa como facial

TABELA 1. Os significados das UAs também estão listados.

Action Unit	Description	Facial Muscle
AU1	Inner Brow Raiser	Frontalis, pars medialis
AU2	Outer Brow Raiser	Frontalis, pars lateralis
AU4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Currugator
AU5	Upper Lid Raiser	Levator palpebrae superioris
AU6	Cheek Raiser	Orbicularis oculi, pars orbitalis
AU7	Lid Tightener	Orbicularis oculi, pars palpebralis
AU9	Nose Wrinkler	Levator labii superioris alaquae nasi
AU10	Upper Lip Raiser	Levator Labii Superioris, Caput infraorbitalis
AU12	Lip Corner Puller	Zygomatic Major
AU14	Dimpler	Buccinator
AU15	Lip Corner Depressor	Depressor anguli oris (Triangularis)
AU17	Chin Raiser	Mentalis
AU20	Lip stretcher	Risorius
AU23	Lip Tightener	Orbicularis oris
AU25	Lips part	Depressor Labii, Relaxation of Mentalis (AU17), Orbicularis Oris
AU26	Jaw Drop	Masetter; Temporal and Internal Pterygoid relaxed
AU28	Lip Suck	Orbicularis oris
AU45	Blink	Relaxation of Levator Palpebrae and Contraction of Orbicularis Oculi, Pars Palpebralis.

reconhecimento de emoções, rastreamento facial e análise de atributos faciais. Em terceiro lugar, sendo um kit de ferramentas de código aberto, o OpenFace permite aos investigadores modificá-lo e personalizá-lo de acordo com as suas necessidades específicas. Essa flexibilidade permite ajustes e melhorias adaptadas aos objetivos individuais de pesquisa e aos vários cenários de aplicação. Em quarto lugar, o OpenFace suporta o processamento de grandes conjuntos de dados de imagens e vídeos faciais. Isto é particularmente valioso para projetos de investigação que envolvam o tratamento de dados extensos, como o reconhecimento facial em sistemas de videovigilância ou o estabelecimento de bases de dados de imagens faciais. Todas essas vantagens são importantes para nós, especialmente quando aplicamos os resultados da pesquisa em ocasiões práticas. Escolhemos arbitrariamente o período de tempo entre 3 segundos e 0 segundos antes da apresentação do alvo como a janela de tempo durante a qual o efeito da atenção pode ser refletido na detecção do alvo, mas os usos de 1 ou 5 segundos mostraram resultados semelhantes (ver Fig. 5) .

As características das expressões faciais foram extraídas como AUs do vídeo feito para cada apresentação alvo usando OpenFace, bem como as posições e ângulos da cabeça e dos olhos. Os significados das UAs são mostrados na Tabela 1. Dois tipos de índices de AU estão disponíveis no OpenFaces: um valor contínuo entre 0 e 5 para 17 UAs (referido como AUr) e um valor binário de 0 ou 1 (ausência ou presença) para 18 UAs (referidas como AUc), que são 17 UAs e AU28 para Lip Suck. Como coletamos dados por um período de 3 segundos para cada alvo, usamos recursos estatísticos dos valores que variam no tempo: mínimo, máximo, média, desvio padrão e três níveis de percentis (25%, 50% e 75%) para AUR e média e desvio padrão para AUc. O número de parâmetros foi 155 no total de variáveis no total.

Talvez existam recursos estatísticos melhores de dados sequenciais do que os que usamos aqui. No entanto, foram suficientes para mostrar a utilidade das UAs para prever o RT (ver mais adiante).

Para uma melhor previsão no futuro, poderíamos investigar características temporais mais complexas.

Para investigar a relação das expressões faciais com o RT de detecção de alvos, tentamos prever o RT de AUs usando um método de aprendizado de máquina chamado LightGBM [27].

LightGBM é um modelo de aumento de gradiente, que opera rapidamente e exibe desempenho relativamente preciso em geral. LightGBM é um modelo de árvore de decisão com aumento de gradiente, no qual o nó das árvores cresce para minimizar os resíduos. Como os dados de treinamento com resíduos grandes são usados preferencialmente, o aprendizado prossegue de forma eficiente, o que é uma técnica poderosa de aprendizado de máquina que pode ser usada tanto para tarefas de regressão quanto de classificação. Funciona combinando vários alunos fracos (árvores de decisão simples) em um aluno forte, que é capaz de fazer previsões precisas sobre novos dados. Neste estudo, dois métodos diferentes foram testados para previsões de RT de AUs. Um método era treinar um modelo com dados agrupados de todos os participantes (modelo de dados agrupados), e o outro era treinar um modelo com todos os participantes, exceto um, e testar com o participante restante (em modelos de teste individuais). O último método consistia em investigar diferenças individuais. Se as diferenças individuais forem pequenas, o modelo construído com outros participantes deverá ser capaz de prever os RT do participante testado. No entanto, variações individuais podem impedir a construção de um modelo geral que possa ser utilizado por qualquer pessoa cujos dados não sejam utilizados para construir o modelo.

Para a avaliação dos modelos foi utilizado um método de validação cruzada de 15 vezes. Todos os dados foram divididos em 15 grupos aleatoriamente para o modelo de dados agrupados, 14 dos quais foram utilizados para treinamento e o grupo restante foi utilizado para teste. O processo foi repetido 15 vezes, um teste para cada grupo, e a média foi utilizada como desempenho do modelo. Para o modelo de teste entre indivíduos, os dados de 14 dos 15 participantes foram usados para treinamento, e os dados do participante restante foram usados para teste. O processo foi repetido 15 vezes, com um teste para cada participante. A média das 15 notas dos testes foi utilizada como desempenho do modelo. O desempenho da predição foi avaliado pela raiz do erro quadrático médio (RMSE) da predição em relação aos dados e pelo coeficiente de correlação de Pearson entre os dados e a predição. (Fig. 2)

4. RESULTADOS

As apresentações do alvo sem respostas dentro de 10 segundos foram excluídas da análise do tempo de reação (TR). Essas apresentações alvo ocorreram em média em 5,5% dos ensaios em todos os participantes. O TR médio em todas as sessões de todos os participantes foi de 1,1 segundos, com desvio padrão de 2,3 segundos. Como o TR médio variou entre os participantes, normalizamos o TR como escores Z após tomar o logaritmo.

Pegamos o logaritmo do RT para minimizar os efeitos da distribuição assimétrica (geralmente uma cauda pesada por mais tempo).

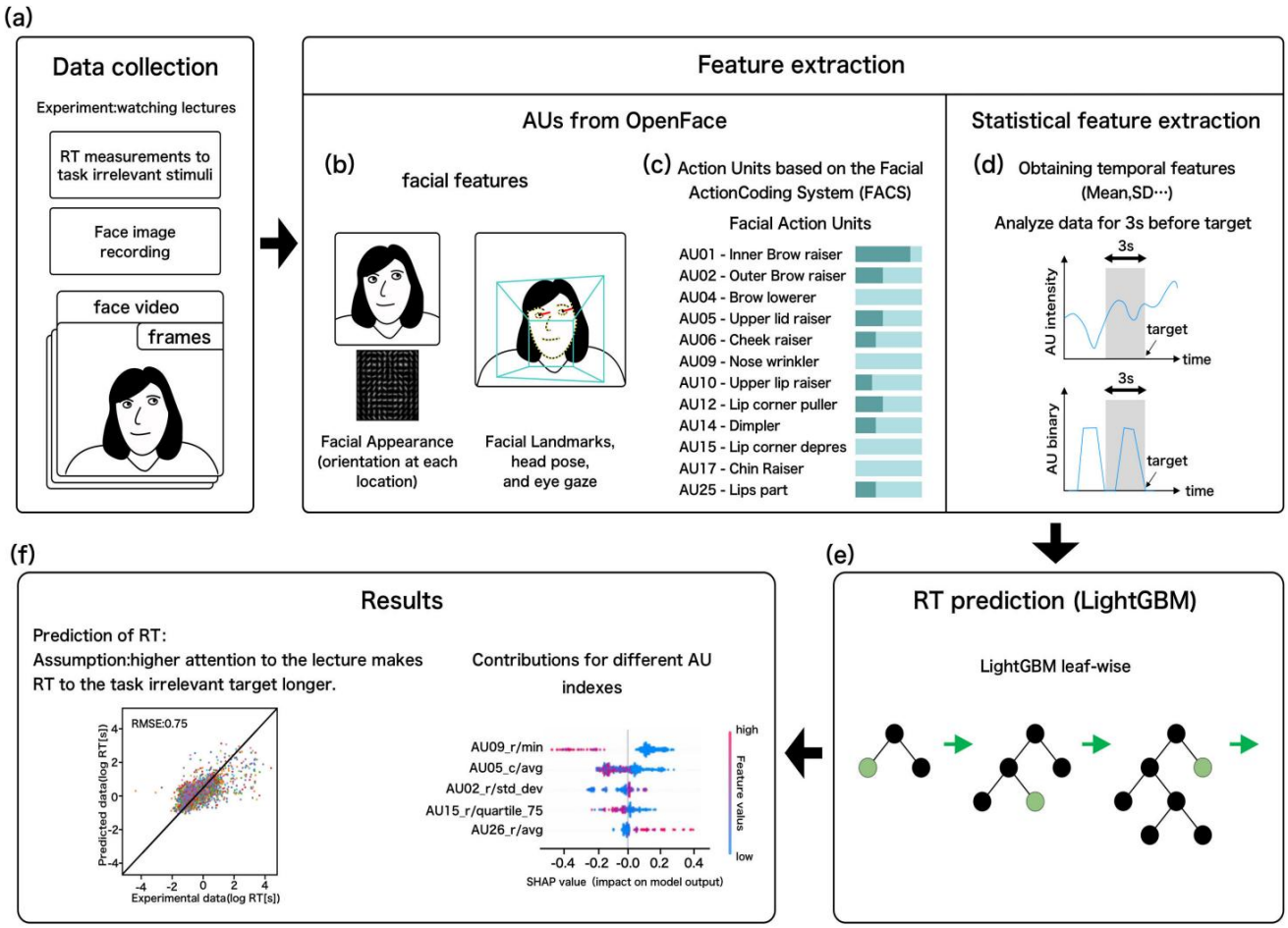

FIGURA 2. A estrutura da análise: (a) Gravação de vídeo dos rostos dos participantes enquanto assistem a palestras on-line e gravação do tempo de reação da detecção do alvo medido como o tempo desde a apresentação do alvo (desaparecimento do ruído branco e o pressionamento da tecla para a detecção). (b) Informações de orientação em cada local, como aparência facial e pontos de referência em um rosto, como olhos, nariz, boca e queixo, são detectados para todos os quadros de vídeo em cada sessão pelo OpenFace. A aparência facial e os pontos de referência são usados para obter AUs com base no Facial Action Coding System (FACS). A pose da cabeça e o olhar fixo também são detectados. A pose da cabeça é um fator importante para analisar imagens faciais de forma normalizada. (c) O OpenFace extraiu unidades de ação (AUs) de pontos de referência faciais e aparência de cada quadro. (d) Utilizamos várias medidas estatísticas de valores sequenciais de AU de uma janela de tempo (3 s para análise principal e 1s e 5s também foram usados) antes de cada apresentação do alvo. As medidas estatísticas utilizadas foram média, desvio padrão, mínimo, máximo e percentis de 25, 50 e 75 para os índices de intensidade. Apenas média e desvio padrão foram utilizados para índices binários. (e) As medidas estatísticas de todas as UAs foram usadas para prever o tempo de reação usando um método de aprendizado de máquina, LightGBM. LightGBM constrói um modelo do tipo árvore com crescimento de árvore por folha, escolhendo a folha com perda máxima de delta para crescer. (f) Comparamos os RTs previstos com os RTs medidos, mostrando sua correlação. Correlação mais alta indica que o modelo LightGBM pode prever bem os RTs para estímulos irrelevantes para a tarefa, de modo que o modelo pode prever o nível de atenção Também analisamos a força da contribuição usando um método chamado explicações aditivas de Shapley (SHAP). SHAP mostra a relação entre os valores de contribuição (força para contribuir com a previsão) e cada um dos índices de recursos.

RT). Também utilizamos valores normalizados de AUs por pontuação Z para evitar os efeitos de variações individuais das características faciais.

Esperávamos que as variações das UAs após a normalização estivessem relacionadas a mudanças nos processos mentais, enquanto os valores absolutos da UA incluem diferenças faciais entre diferentes indivíduos. Em seguida, aplicamos LightGBM para modelar a relação entre RT e expressões faciais e testamos o modelo usando um método de validação cruzada de 15 vezes. A Figura 3a mostra os resultados de previsão do modelo de dados agrupados. O eixo horizontal mostra o RT medido no experimento e o eixo vertical mostra a previsão do LightGBM. Cada ponto representa cada apresentação alvo de todas as sessões de todos os participantes e cores diferentes indicam diferentes combinações de treinamento-teste (15 combinações diferentes com cores diferentes). O RMSE do desvio dos dados das previsões (ou do desvio das previsões dos dados) foi de 0,75. A média dos dados de RT é zero, com um desvio padrão unitário após a pontuação Z por definição. Assim, o RMSE da previsão do modelo (0,75, que é menor que 1) indica que o modelo pode explicar, pelo menos parcialmente, a variação dos dados (25% neste caso). O coeficiente de correlação de Pearson entre dados e predição foi de 0,66. Um teste estatístico de ausência de correlação mostrou que a correlação foi estatisticamente significativa ($p < 0,001$, $t(2412) = 11$). Usamos um teste para examinar se o coeficiente de correlação de Pearson não é significativamente diferente de zero e mostramos a suposição

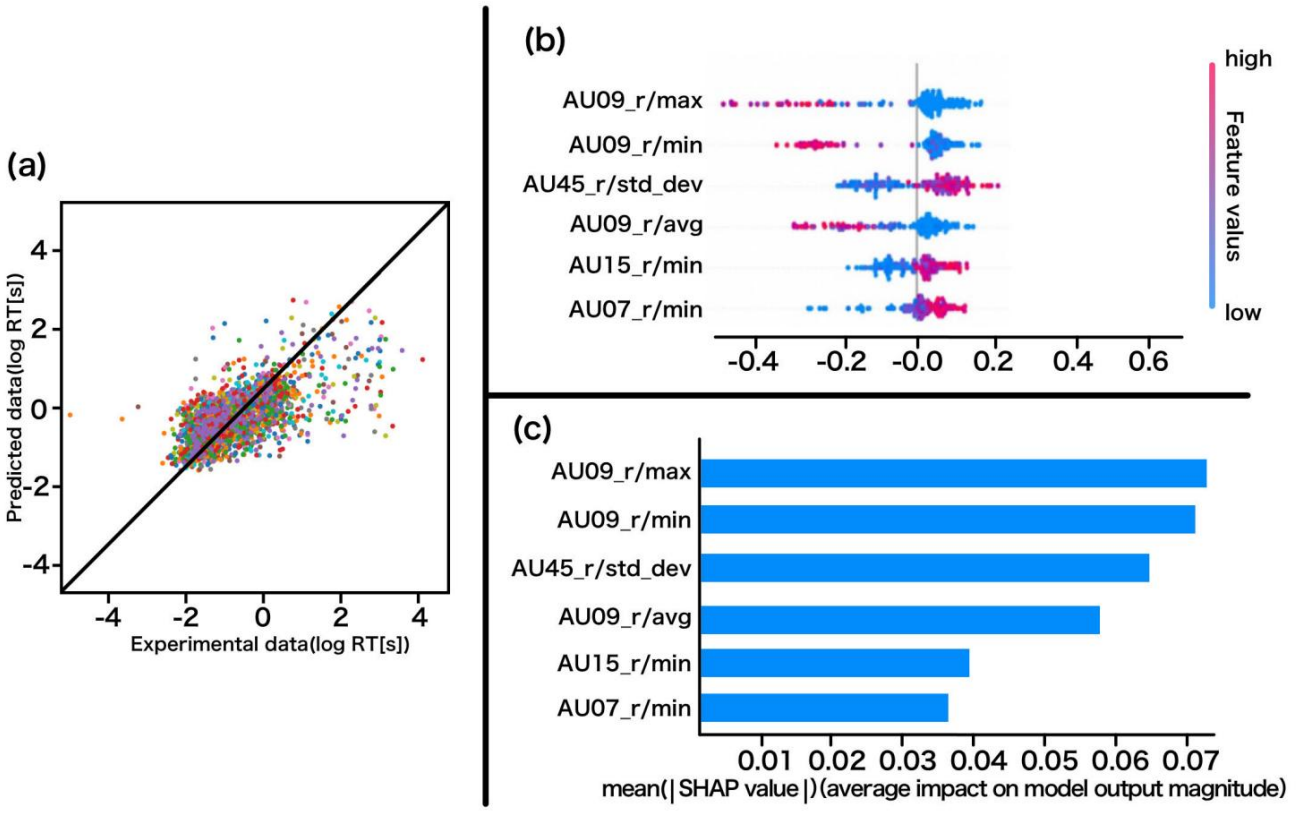

FIGURA 3. (a) Correlação entre o tempo de reação (TR) medido e o TR previsto do modelo. Cada ponto representa o RT de cada apresentação alvo de todas as sessões de todos os participantes. Cores diferentes indicam diferentes combinações de treino-teste (15 combinações). (b) Os índices são organizados de acordo com o nível de contribuição para a predição obtida pelo método de explicações aditivas de Shapley (SHAP). Cada ponto é de cada RT, como na figura de correlação da figura 3 (a), e a cor (vermelho ou azul) indica uma contribuição positiva ou negativa. O eixo horizontal indica o nível de contribuição para a predição do RT pelo modelo (c) O valor absoluto que corresponde à contribuição de cada índice para a predição estimada pelo SHAP.

de não diferente foi rejeitado com nível de 5%. Além da significância estatística do coeficiente de correlação, também utilizamos um teste estatístico de RMSE para mostrar que nossa previsão é melhor que o acaso. Comparamos o RMSE da previsão do modelo e o dos dados, que é um após a pontuação Z, usando um teste t ($p < 0,001$, $t(14) = 16,62$). A presente análise previu com sucesso o RT para alvos irrelevantes para a tarefa, que assumimos variar dependendo dos estados de atenção.

Essa previsão do RT, por sua vez, previu o estado de atenção no momento alguns segundos antes da apresentação do alvo durante o aprendizado. Concluímos que as características faciais e os movimentos da cabeça e dos olhos contêm informações sobre a atenção.

Uma análise mais aprofundada revelou o nível de contribuição de cada índice para a predição (ou seja, a importância de cada índice para a predição) usando um método chamado explicações aditivas de Shapley (SHAP) [28]. SHAP fornece o valor que corresponde à contribuição de cada recurso de entrada para a previsão (Fig. 3 c). AU9 (enrugador de nariz), AU45 (piscar), AU15 (depressor de canto labial) e AU7 (apertador de pálpebras) foram os cinco melhores contribuintes entre todas as UAs. A análise também fornece o grau de contribuição de cada recurso de entrada para prever cada evento de detecção de alvo, conforme mostrado pelos pontos na Figura 3 (b). Os pontos vermelhos indicam valores altos de índices de características faciais e os pontos azuis indicam valores baixos. Os padrões de distribuição de pontos de dados em vermelho e azul mostram, por exemplo, que AU9 contribui negativamente para o TR. Valores mais altos (pontos vermelhos) foram distribuídos na direção negativa do eixo horizontal, indicando que um TR mais curto estava associado a mais enrugamento nasal, o que, por sua vez, sugere que menos atenção foi dada à palestra quando mais enrugamento nasal foi exibido. Discutiremos o efeito dessas UAs com mais detalhes na seção Discussão.

Realizamos sessões de controle para confirmar que assistir a um vídeo de palestra influencia o TR para o alvo auditivo. Na condição controle, os participantes foram solicitados a detectar o alvo sem prestar atenção ao vídeo-aula. Considera-se que os TR nesta condição refletem atenção total ao alvo auditivo. O TR médio para as duas sessões de controle em todos os participantes foi de 0,7 segundos. Esta duração do RT é claramente inferior à média do RT nas sessões teóricas, que foi de 1,1 seg. e o coeficiente de correlação de Pearson entre o

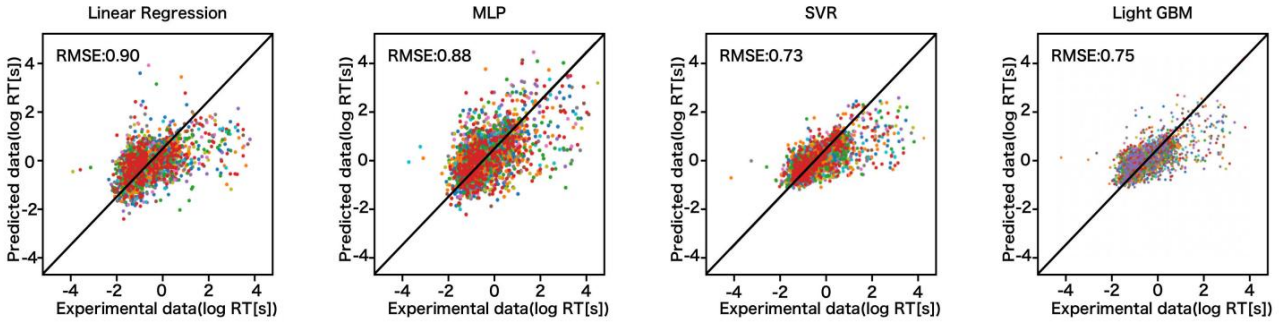

FIGURA 4. Comparação de quatro modelos diferentes: Support Vector Regression (SVR), Multilayer perceptron (MLP), Linear Regression e LightGBM.

sessões experimentais e de controle foi estatisticamente significativa ($p < 0,05$, $t(14) = 2,74$), indicando que o TR ao alvo foi uma medida apropriada de atenção às palestras.

Tentamos prever os RTs das condições de controle com o mesmo procedimento utilizado para a sessão de aula. Os resultados revelaram que o RMSE das predições foi de 0,89, e o coeficiente de correlação de Pearson entre dados e predição foi de 0,45, o que não foi estatisticamente significativo ($p = 0,092$, $t(517) = 3,1$).

Há três questões a serem examinadas antes de aceitar os resultados. A primeira é se os resultados dependem da escolha dos métodos de aprendizado de máquina, a segunda é se dependem da seleção das janelas de tempo e a terceira é se dependem de variações individuais. Primeiro, usamos três modelos diferentes além do LightGBM como comparação: Support Vector Regression (SVR), Multilayer perceptron (MLP) e Linear Regression. Os resultados mostraram que a precisão do lightGBM é semelhante à do SVR, que é melhor que o MLP e a Regressão Linear (Fig.4), e que o tempo necessário para análise foi o mais curto para o lightGBM entre os quatro métodos.

Em segundo lugar, não há razão teórica para selecionar um determinado período de tempo para a extração de características faciais para estimar os RTs. Usamos janelas de 1 e 5 segundos, além de janelas de 3 segundos, para ver o efeito do tempo na análise. Os resultados são semelhantes para os três casos (fig. 5). Um teste t de RMSE da predição do modelo com um mostrou significância estatística tanto para janelas de 1 quanto de 5 segundos ($p < 0,001$, $t(14) = 15,32$ para 1s e $p < 0,001$, $t(14) = 18,08$ para 5s).

Terceiro, testamos se um modelo construído com dados de outros indivíduos (através de modelos individuais) pode prever os dados de outro indivíduo. A Figura 6 mostra os resultados das previsões. Surpreendentemente, os resultados não revelaram nenhuma previsão bem-sucedida entre os participantes. Assim, um modelo baseado em um grupo de indivíduos não poderia ser usado para prever o nível de atenção de um indivíduo do grupo. As informações faciais relacionadas à atenção pareciam variar de participante para participante.

V. DISCUSSÃO

No presente estudo, medimos o RT para alvos irrelevantes para tarefas como um índice de nível de atenção. Com os RTs, desenvolvemos um método para prever o envolvimento em videoaulas usando uma técnica de aprendizado de máquina. Nossa abordagem foi prever o tempo de resposta sob a suposição de que o tempo de resposta se tornaria mais longo quando mais atenção fosse dada à palestra, reduzindo a atenção a um alvo que era irrelevante para a palestra. O modelo construído para a predição forneceu informações sobre as características faciais que mais contribuíram para a predição, que foram: AU9 (enrugador de nariz), AU45 (piscar), AU15 (abaixador de canto labial) e AU7 (apertador de pálpebras). Aqui, discutimos possíveis explicações para a importância desses fatores na previsão de RTs. AU9 esteve negativamente relacionado ao TR. Um TR mais longo foi associado a mais atenção à palestra, sugerindo que AU9 estava negativamente relacionado à quantidade de atenção dada à palestra.

O aumento do enrugamento do nariz foi associado ao desvio de atenção da palestra. Por outro lado, os resultados sugerem que AU45 (piscar), AU15 (abaixador de canto labial) e AU7 (apertador de tampa) foram positivamente relacionados ao nível de atenção dada à palestra. Assim, maior depressão do canto dos lábios, piscadas mais frequentes e maior aperto das pálpebras são esperados quando a pessoa presta mais atenção às palestras.

A depressão do canto labial pode estar relacionada a situações em que o aluno tem dificuldade em compreender a palestra. Isto pode levar o aluno a tentar assistir mais às aulas e a exibir uma expressão facial séria. Apertar as pálpebras e piscar são ações faciais semelhantes, e ambas podem estar relacionadas ao esforço para compreender o conteúdo das palestras, abrindo mais os olhos. No entanto, piscar mais e apertar as pálpebras também pode estar relacionado à sonolência. Quando uma pessoa está com sono, é provável que ela não compareça às palestras ou a qualquer estímulo irrelevante para a tarefa, o que resultaria em um TR mais longo para o alvo, mesmo sem um alto nível de atenção sendo prestado à palestra. Embora o presente experimento tenha sido projetado assumindo apenas dois estados de atenção, atendendo à palestra ou ao alvo irrelevante da tarefa, o nível de atenção poderia ser potencialmente reduzido pela sonolência, resultando em RTs mais longos para o alvo com diminuição da atenção à palestra.

Tentamos estimar o efeito da sonolência durante as aulas e reanalisamos os dados.

Para excluir a possível influência da sonolência nos resultados, reanalisamos os dados após a remoção dos dados com faces sonolentas. Para identificar os horários em que um participante apareceu

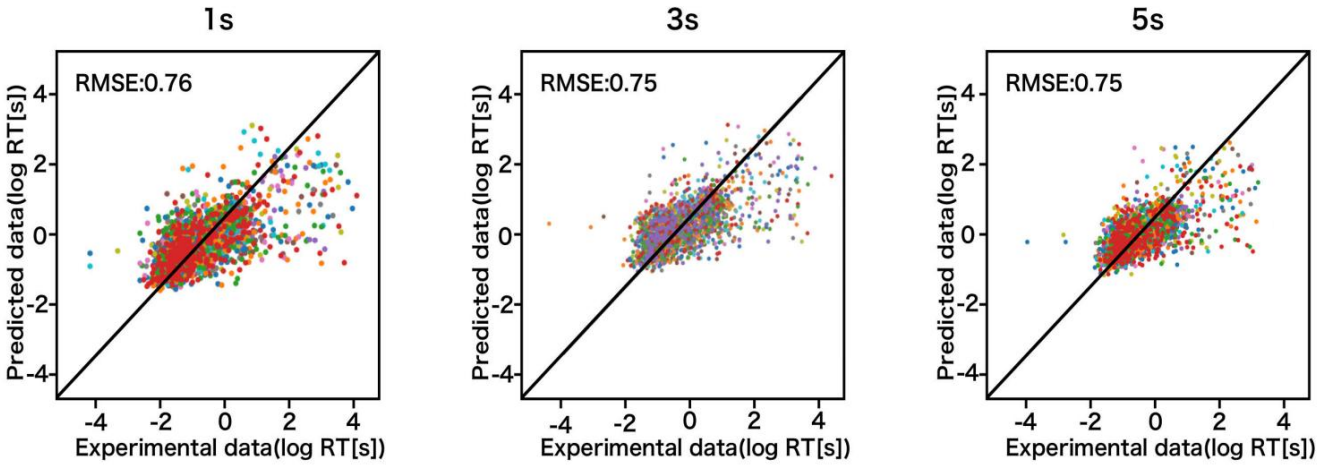

FIGURA 5. Aplicamos janelas de tempo de 1, 3 e 5 segundos para ver o efeito do tempo para analisar as características faciais.

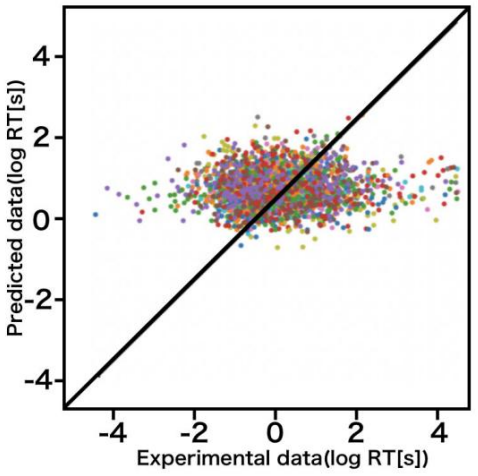

FIGURA 6. Correlação entre RTs medidos e previstos, utilizando o modelo de teste entre indivíduos. As configurações são as mesmas da Fig. 3 (a) .

para ter sono, utilizamos dados de movimento ocular e avaliação subjetiva de sonolência em vídeos. Um estudo anterior relatou que os olhos ficam parados quando estão com sono [28]. Tentamos detectar quando os alunos estavam com sono usando os dados do olhar. Calculamos o desvio padrão das posições do olhar, obtido através da análise OpenFace, durante 3 segundos antes de cada apresentação do alvo. O histograma na Fig. 7 (a) mostra a distribuição do desvio padrão das localizações do olhar, que reflete a atividade do movimento ocular. O eixo horizontal mostra a escala logarítmica do ângulo visual em radianos, mostrando claramente os dados com valores pequenos. Os resultados da distribuição podem ser descritos como valores de desvio padrão seguindo uma distribuição de pico único com um pico de aproximadamente -0,75 em log graus. No entanto, parecia haver um pico em valores muito pequenos de aproximadamente -1,34 em log graus. Os movimentos oculares para as imagens de vídeo que foram julgadas subjetivamente como sonolentas exibiram um desvio padrão inferior a -1,13 log deg. Assim, definimos as faces do vídeo com desvio padrão do olhar localização menor que -1,13 log deg como rostos que refletiam sonolência. Observe que esta análise não se baseia em medições precisas do movimento ocular, mas em uma estimativa aproximada por processamento de imagem usando OpenFace, pela qual estimamos que a resolução espacial era superior a 2 radianos. Apesar da baixa precisão deste método, a estabilidade do olhar pôde ser avaliada com base na distribuição mostrada na Fig .

Reanalisamos os dados após remover os dados associados à sonolência usando um limite do desvio padrão da localização do olhar inferior a -1,13 log deg. Os resultados sem rostos sonolentos revelaram que o RMSE das previsões foi de 0,77 (ver Fig. 7 b), que foi menor que o RMSE da linha de base de 1,0. O coeficiente de correlação de Pearson entre dados e predição foi de 0,67. e a correlação foi estatisticamente significativa ($p < 0,001$, $t(2298) = 11$), também utilizamos um teste estatístico de RMSE para mostrar que nossa previsão é melhor que o acaso. Comparamos o RMSE da previsão do modelo e o dos dados, que é um após a pontuação Z, usando um teste t ($p < 0,001$, $t(14) = 15,07$). Estes resultados confirmam que as expressões faciais podem ser usadas para prever estados de atenção enquanto assiste a uma palestra. A Figura 7 (c) mostra o nível de contribuição de cada UA para a previsão do SHAP. Semelhante à análise original (Fig. 3), AU9 (enrugador de nariz) foi o maior contribuinte, e AU45 (piscar) e AU15 (depressor de canto labial) foram o segundo e terceiro maiores contribuintes, respectivamente. No entanto, AU7 (apertador de tampa), que estava entre os cinco principais contribuintes na análise original, não estava mais incluído entre os cinco primeiros. Esses resultados indicam que enrugar o nariz, piscar e pressionar os cantos dos lábios são fatores importantes na previsão da atenção às palestras.

Diferenças individuais na relação entre estado de concentração interna e expressão facial, que não foram capturadas em estudos anteriores que utilizaram classificações subjetivas [14], [15], [16], foram encontradas no presente estudo.
Consideramos várias razões possíveis para estes resultados. Uma possibilidade é que a identificação individual tenha afetado a

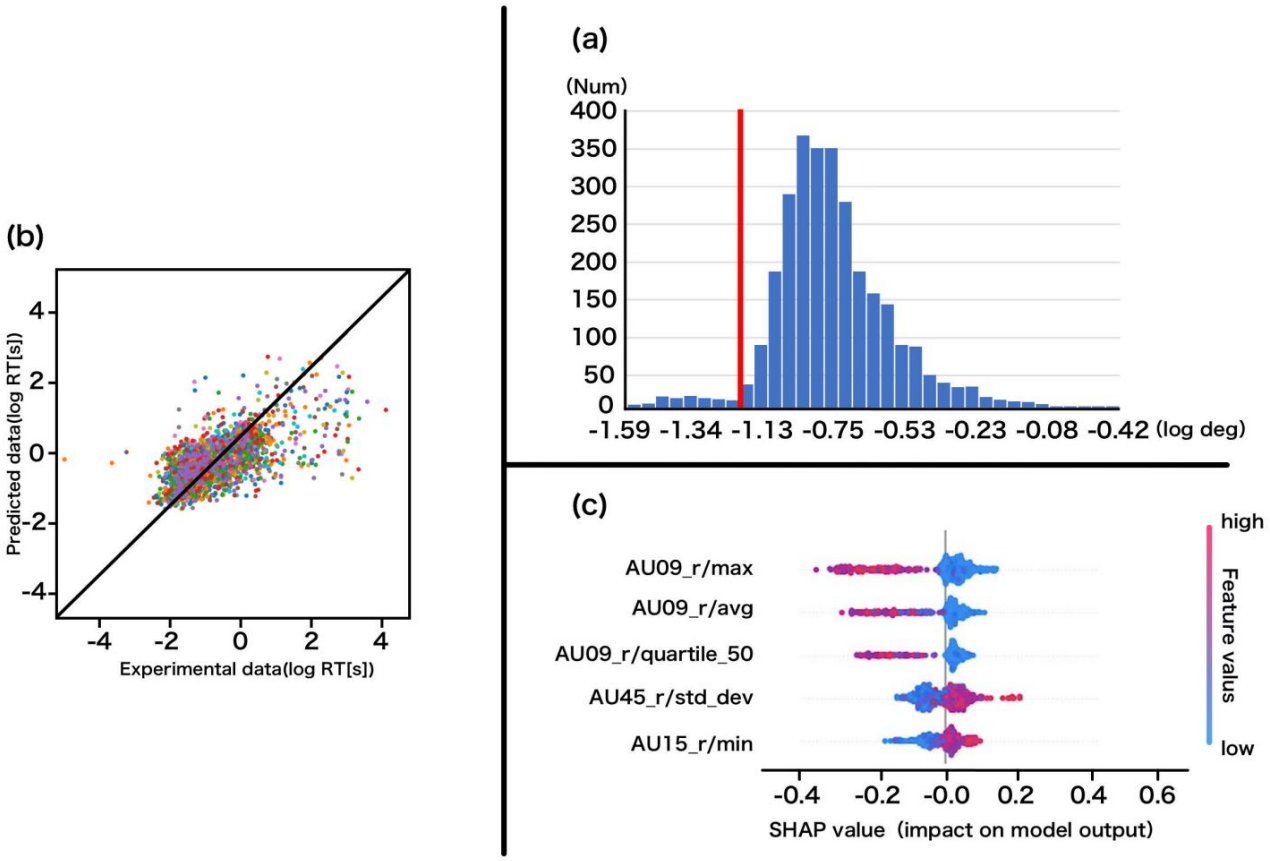

FIGURA 7. (a) Histograma do desvio padrão dos movimentos do olhar (DP do olhar). Os SDs do olhar antes das apresentações alvo com rostos sonolentos estimados subjetivamente foram menores que a linha vermelha, e assumimos que os RTs com SDs do olhar maiores que a linha vermelha não foram influenciados pela sonolência. (b) Correlação entre TR medidos e previstos para dados sem influência da sonolência. As configurações são as mesmas da Fig. 3 (a). (c) Os índices são organizados de acordo com o nível de contribuição para a previsão obtida através do SHAP. As configura

descobertas. Como as próprias UA podem conter informações sobre as características faciais de participantes individuais, a análise da UA pode identificar indivíduos. Se houver variação individual substancial nos RTs no presente experimento, a identificação de indivíduos pelas características faciais poderia potencialmente prever os resultados do RT com algum nível de precisão, porque há uma correlação entre as características faciais e o RT dos indivíduos.

Porém, como utilizamos valores normalizados de RTs e AUs para cada participante, as médias de cada parâmetro não apresentaram correlações entre os parâmetros.

Em outras palavras, as diferenças individuais que encontramos poderiam ser explicadas por diferenças individuais nas contribuições das características faciais para a estimativa do RT.

Para investigar os efeitos da variabilidade individual, primeiro conduzimos as mesmas análises para os dados de cada participante.

Como a quantidade de dados de cada participante é relativamente pequena, realizamos uma análise de validação cruzada de 5 vezes (em vez de 15 vezes) nos próprios dados de cada participante. O RMSE médio da previsão em relação aos dados para todos os participantes foi muito próximo da linha de base 1,01 (Fig. 8 c). O RMSE é tão ruim quanto o dos modelos individuais (mostrados na Fig. 6), provavelmente devido à pequena quantidade de dados usados para cada modelo, mesmo com validação cruzada de 5 vezes. Nós então,

examinamos o efeito do tamanho do conjunto de dados na precisão da previsão e descobrimos que aproximadamente 20% de todos os dados foram necessários para obter um efeito de treinamento com RMSE de cerca de 0,8 (Fig. 8 f). Para manter a proporção do conjunto de dados superior a 20%, comparamos as previsões entre participantes e entre participantes usando conjuntos de dados de três ou cinco grupos de participantes, em vez de conjuntos de dados de participantes individuais.

Melhores previsões na análise dentro do grupo em comparação com aquelas na análise entre grupos eram esperadas se houvesse grandes variações individuais nas expressões de características relacionadas ao envolvimento com a palestra. No caso da divisão em três grupos, dois dos três grupos foram usados para treinamento e o terceiro grupo foi usado para um teste de análise entre grupos, enquanto quatro grupos foram usados para treinamento com o quinto como teste no caso de cinco grupos. -divisão de grupo. Para análise intragrupo, os dados foram divididos em três ou cinco conjuntos, selecionando igual número de dados de cada grupo (cada conjunto de dados tinha um terço do primeiro grupo, um terço do segundo grupo e um terço do terceiro grupo no caso de três grupos). Esses três ou cinco conjuntos de dados foram usados para testes de validação tripla ou quíntupla.

A Figura 8 mostra os resultados das análises dentro e entre grupos para os três e cinco grupos, além das análises

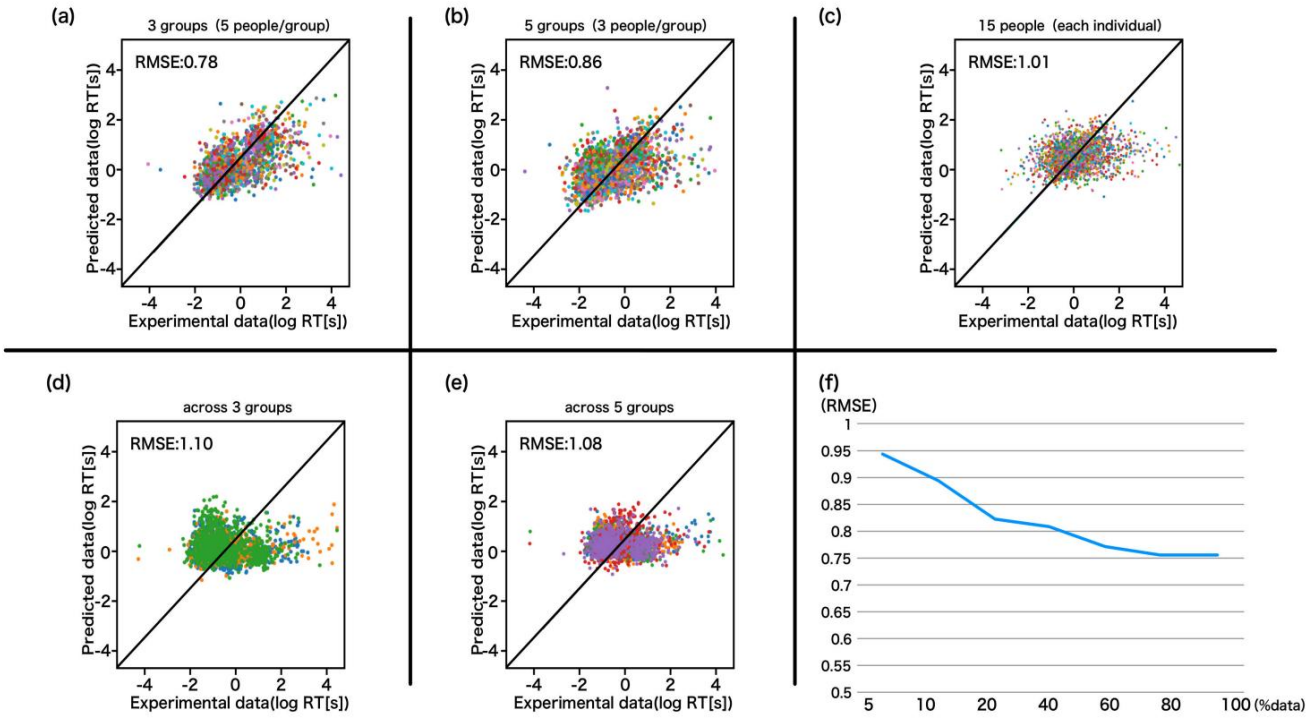

FIGURA 8. (a) os resultados dos três dentro dos grupos, (b) cinco dentro dos grupos, (c) cada indivíduo, (d) três entre grupos e (e) cinco entre grupos. (f) o desempenho da previsão em função do tamanho dos dados.

previsões individuais médias. A precisão da previsão foi melhor para análises dentro do grupo em comparação com análises entre grupos. Os valores de RMSE foram 0,86 e 0,78 para três e cinco análises intragrupo, respectivamente, e 1,10 e 1,08 para três e cinco análises intergrupos, respectivamente. Um teste t de RMSE da predição do modelo com um mostrou significância estatística tanto para 3 quanto para 5 grupos ($p < 0,001$, $t(14) = 7,11$ para 3 grupos e $p < 0,001$, $t(14) = 11,19$ para 5 grupos). Esses resultados indicam variações individuais não triviais na relação entre expressões faciais e engajamento. Essas variações não significam que não haja um fator comum compartilhado por alguns indivíduos, porque foi demonstrado que o agrupamento de dados de muitos participantes melhora a previsão (compare as Figuras 7 e 8). Os valores SHAP para as análises de três e cinco grupos mostraram que AU9 e AU2 estavam entre as cinco melhores características em ambos os grupos. AU9 também foi incluído na análise original com todos os dados. Este resultado sugere que estas características são importantes para todos os indivíduos, enquanto outras características que diferem substancialmente entre os indivíduos podem prejudicar as previsões entre os participantes. Embora a variação individual limite a utilização do modelo sem dúvida, é possível construir um modelo para um grupo de indivíduos com propriedades semelhantes.

Os resultados sugerem que a variação individual é substancial e parece ser uma desvantagem em geral quando a presente técnica é aplicada a um sistema de suporte, utilizando um modelo treinado com diferentes indivíduos. No entanto, o modelo pode ser personalizado para cada indivíduo e os modelos construídos para indivíduos específicos podem ser mais precisos. Embora a variação individual deva ser investigada mais detalhadamente para compreender os fatores essenciais, a técnica aqui desenvolvida pode ser usada para aplicações em condições reais de educação.

Embora estudos psicofísicos tenham utilizado estímulos sonoros como sonda para medir o nível de atenção [30], [31], [32], tal abordagem não é prática nas situações reais de aprendizagem.
Portanto, investigamos se as imagens faciais são suficientes para fornecer índices de nível de atenção. O desempenho do modelo depende do desempenho do OpenFace. Embora Baltrusaitis et al. [33] relataram que a precisão do OpenFace é melhor que outros métodos, é óbvio que seu desempenho não é perfeito e depende das condições de gravação das faces. Nossa estimativa de RT de AUs, portanto, inclui erros de estimativa de características faciais em uma determinada quantidade. Acreditamos que esta análise é útil para obter informações sobre as condições (estados mentais) de um aluno em cada momento para fazer feedbacks apropriados. Por exemplo, 70% de detecção correta de menos atenção a uma palestra deve ser útil para fornecer um sinal de alerta aos alunos e/ou ao palestrante. Três vezes em cada dez avisos errados não devem ser problema se o sinal de alerta usado não perturbar muito a aula. Além disso, a taxa de detecção torna-se superior a 99% se houver mais de cinco alunos que perdem a atenção à aula, mesmo que seja 70% para um.

aluno.

VI. CONCLUSÃO

Concluindo, revelamos que as expressões faciais podem ser usadas para prever o nível de atenção dos alunos às videoaulas, o que serve como um índice de envolvimento dos alunos. As características faciais capturadas por uma câmera de vídeo podem prever tempos de reação (RTs), que são considerados indicativos de estados de atenção.

Características faciais específicas, como enrugamento do nariz, piscar e depressão nos cantos dos lábios, parecem estar associadas à atenção durante as videoaulas. A aplicação da tecnologia de expressão facial tem o potencial de melhorar a qualidade do ensino. Contudo, antes de implementá-lo em condições reais de ensino, algumas considerações devem ser levadas em conta. Em primeiro lugar, os mecanismos subjacentes às contribuições destas características ainda não são compreendidos, o que é essencial para a generalização. Em segundo lugar, foram observadas diferenças individuais significativas. Personalizar o modelo pode ser uma solução possível. Em pesquisas futuras, nos concentraremos em explorar as diferenças individuais e a relação fisiológica entre o envolvimento e as expressões faciais durante a aprendizagem.

REFERÊNCIAS

[1] S. Shioiri, Y. Sato, Y. Horaguchi, H. Muraoka e M. Nihei, "Quali-informática na sociedade com dados em escala Yotta", em *Proc. Internacional IEEE. Simp. Sistema de Circuitos (ISCAS),* maio de 2021, pp.

[2] Y. Sato, Y. Horaguchi, L. Vanel e S. Shioiri, "Predição de preferências de imagem a partir de expressões faciais espontâneas", *Interdiscipl. Inf. Ciência,* vol. 28, não. 1, pp. 45–53, 2022.

[3] Y. Horaguchi, Y. Sato e S. Shioiri, "Estimativa de preferências para imagens por análise de expressão facial", *IEICE Tech. Rep.,* vol. 120, não. 306, pp. 71–76, 2020.

[4] C. Thomas e DB Jayagopi, "Prevendo o envolvimento dos alunos nas salas de aula usando dicas comportamentais faciais", em *Proc. 1º Int. ACM SIGCHI. Workshop Interacção Multimodal. Educ.,* Glasgow, Reino Unido, novembro de 2017, pp.

[5] NK Mehta, SS Prasad, S. Saurav, R. Saini e S. Singh, "Rede neural de autoatenção DenseNet tridimensional para detecção automática do envolvimento do aluno", *Appl. Intel.,* vol. 52, não. 12 , pp .

[6] DK Darnell e PA Krieg, "O envolvimento dos alunos, avaliado através da frequência cardíaca, não mostra nenhuma redefinição após sessões de aprendizagem ativa em palestras," *PLoS ONE,* vol. 14, não. 12, dez. 2019, art. não. e0225709, doi: 10.1371/journal.pone.0225709.

[7] DM Bunce, EA Flens e KY Neiles, "Por quanto tempo os alunos conseguem prestar atenção nas aulas? Um estudo sobre o declínio da atenção dos alunos usando clickers", *J. Chem. Educ.,* vol. 87, não. 12, pp. 1438–1443, dezembro de 2010, doi: 10.1021/ed100409p.

[8] H. Kato, K. Takahashi, Y. Hatori, Y. Sato e S. Shioiri, "Predição de envolvimento a partir de mudanças temporais na expressão facial", em *Proc. Conferência Mundial Computação. Educ.,* Hiroshima, Japão, agosto de 2022.

[9] HL O'Brien e EG Toms, "O desenvolvimento e avaliação de uma pesquisa para medir o envolvimento do usuário", *J. Amer. Soc. Inf. Ciência. Tecnologia,* vol. 61, não. 1, pp. 50–69, janeiro de 2010.

[10] AM Leiker, AT Bruzi, MW Miller, M. Nelson, R. Wegman e KR Lohse, "Os efeitos da seleção autônoma de dificuldade no envolvimento, motivação e aprendizagem em uma tarefa de videogame controlada por movimento" *Zumbir. Movimento Sci.,* vol. 49, pp. 326–335, outubro de 2016, doi: 10.1016/j.humov.2016.08.005.

[11] LS Pagani, C. Fitzpatrick e S. Parent, "Relacionando a atenção do jardim de infância aos caminhos de desenvolvimento subsequentes do envolvimento em sala de aula na escola primária", *J. Abnormal Child Psychol.,* vol. 40, não. 5, pp. 715–725, julho de 2012, doi: 10.1007/s10802-011-9605-4 .

[12] MN Nguyen, S. Watanabe-Galloway, JL Hill, M. Siahpush, MK Tibbits e C. Wichman, "Modelo ecológico de envolvimento escolar e transtorno de déficit de atenção/hiperatividade em crianças em idade escolar," *Eur . Psiquiatria Infantil e Adolescente,* vol. 28, não. 795–805, junho de 2019, doi: 10.1007/s00787-018-1248-3.

[13] M. Kinnealey, B. Pfeiffer, J. Miller, C. Roan, R. Shoener e ML Ellner, "Efeito da modificação da sala de aula na atenção e envolvimento de alunos com autismo ou dispraxia", *Amer. J. Terapia Ocupacional,* vol. 66, não. 5, pp. 511–519, 2012, doi: 10.5014/ajot.2012.004010.

[14] Ó. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein e E. Kasneci, "Análise de engajamento multimodal de vídeos faciais na sala de aula", *IEEE Trans. Afeto. Computação,* vol. 14, não. 2, pp. 1012–1027, abril/junho. 2023, doi: 10.1109/TAFFC.2021.3127692.

[15] H. Monkaresi, N. Bosch, RA Calvo e SK D'Mello, " Detecção automatizada de envolvimento usando estimativa baseada em vídeo de expressões faciais e frequência cardíaca, " *IEEE Trans. Afeto. Computação,* vol. 8, não. 15–28, janeiro de 2017, doi: 10.1109/TAFFC.2016.2515084.

[16] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster e JR Movellan, "As faces do engajamento: Reconhecimento automático do envolvimento do aluno a partir de expressões faciais", *IEEE Trans. Afeto. Computação,* vol. 5, não. 1, pp. 86–98, janeiro de 2014, doi: 10.1109/TAFFC.2014.2316163.

[17] R. Miao, H. Kato, Y. Hatori, Y. Sato e S. Shioiri, "Análise de expressões faciais para a estimativa de concentração em palestras online," em *Proc. Conferência Mundial Computação. Educ.,* Hiroshima, Japão, agosto de 2022.

[18] AF Kramer, LJ Trejo e D. Humphrey, "Avaliação da carga de trabalho mental com sondas auditivas irrelevantes para tarefas", *Biol. Psicol.,* vol. 40, nºs. 1–2, pp. 83–100, maio de 1995.

[19] A. Pfefferbaum, JM Ford, WT Roth e BS Kopell, "Diferenças de idade nas associações de tempo de reação P3", *Electroencephalogr. Neurofisiologia Clínica,* vol. 257–265, agosto de 1980, doi: 10.1016/ 0013-4694(80)90220-5.

[20] MI Posner, "Orientação da atenção", *Quart. J. Exp. Psicol.,* vol. 32, pp. 3–25, fevereiro de 1980.

[21] SA Hillyard, RF Hink, VL Schwent e TW Picton, "Sinais elétricos de atenção seletiva no cérebro humano", *Science,* vol. 182, não. 177–180, outubro de 1973, doi: 10.1126/science.182.4108.177.

[22] S. Shioiri, M. Ogawa, H. Yaguchi e P. Cavanagh, "Facilitação atencional de detecção de cintilação em objetos em movimento", *J. Vis.,* vol. 15, não. 14, pág. 3, out. 2015, doi: 10.1167/15.14.3.

[23] S. Shioiri, H. Honjyo, Y. Kashiwase, K. Matsumiya e I. Kuriki, "A atenção visual se espalha amplamente, mas seleciona informações localmente", *Sci. Rep.,* vol. 6, não. 1, pág. 35513, outubro de 2016, doi: 10.1038/srep35513.

[24] (2023). *@Fuku-Programação.* Acesso em: 14 fev. 2023. [Online]. Disponível: https://www.youtube.com/watch?v=uVaOzQLxXt0 [25] T. Baltrušaitis, P. Robinson e L.-P. Morency, "OpenFace: Um kit de ferramentas de análise de comportamento facial de código aberto", em *Proc. Conferência de Inverno IEEE. Apl. Computação. Vis. (WACV),* março de 2016, pp.

[26] P. Ekman e WV Friesen, *Sistema de codificação de ação facial: uma técnica para medir o movimento facial.* São Francisco, CA, EUA: Consulting Psychologists Press, 1978.

[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye e TY Liu, " LightGBM: Uma árvore de decisão de aumento de gradiente altamente eficiente, " em *Processo. Av. Inf. Neural. Processo. Sistema,* 2017, pp.

[28] SM Lundberg e SI Lee, "Uma abordagem unificada para interpretar previsões de modelos", em *Proc. Av. Inf. Neural. Processo. Sistema,* 2017, pp.

[29] T. Abe, T. Nonomura, Y. Komada, S. Asaoka, T. Sasai, A. Ueno e Y. Inoue, "Detectando vigilância deteriorada usando porcentagem de tempo de fechamento das pálpebras durante testes de manutenção comportamental de vigília, " *Int. J. Psicofisiol.,* vol. 82, não. 269–274, dezembro de 2011, doi: 10.1016/j.ijpsycho.2011.09.012.

[30] MA Bedard, F. El Massioui, B. Pillon e JL Nandrino, "Tempo para reorientar a atenção: uma hipótese pré-motora do mecanismo subjacente", *Neuropsicologia,* vol. 31, não. 3, pp. 241–249, março de 1993, doi: 10.1016/0028-3932(93)90088-h.

[31] G. Rhodes, "Atenção auditiva e a representação da informação espacial", *Perception Psychophys.,* vol. 42, não. 1, pp. 1–14, janeiro de 1987, doi: 10.3758/bf03211508.

[32] JR Simon, E. Acosta e SP Mewaldt, "Efeito do locus do tom de alerta no tempo de reação da escolha auditiva", *Memory Cognition,* vol. 3, não. 2, pp. 70–167, março de 1975, doi: 10.3758/BF03212893.

[33] T. Baltrusaitis, A. Zadeh, YC Lim e L.-P. Morency, "OpenFace 2.0: Kit de ferramentas de análise de comportamento facial", em *Proc. 13º Int. IEEE. Conf. Automático. Reconhecimento de gestos faciais. (FG),* maio de 2018, pp. 59–66, doi: 10.1109/FG.2018.00019.

RENJUN MIAO nasceu em Wenzhou, Zhejiang, em 1986. Possui bacharelado em engenharia de automação mecânica pela Zhe-jiang University City College, em 2008, e mestrado em engenharia da informação pela Universidade de Tohoku, Japão, em 2012, onde ele está atualmente cursando o doutorado. graduação, com foco em computação afetiva, principalmente na detecção da qualidade da educação online dos alunos por meio da mudança de expressões faciais.

De 2010 a 2012, seu principal foco de pesquisa foi o processamento de sinais de cor e forma na neurologia visual do cérebro. É Engenheiro desde a graduação e Supervisor de Desenvolvimento Educacional SAAS, em 2017.

HARUKA KATO recebeu o bacharelado em engenharia e o mestrado em engenharia da informação pela Universidade de Tohoku, em 2021 e 2023, respectivamente.

De 2021 a 2023, sua principal pesquisa foi computação afetiva, principalmente na detecção do envolvimento do aluno durante o estudo por meio da mudança de eletroencefalografia e expressões faciais.
Seus interesses de pesquisa incluem engajamento, atenção durante o estudo e expressões faciais.

YASUHIRO HATORI recebeu o diploma de bacharelado em engenharia da informação e o mestrado e doutorado. diplomas em engenharia pela Universidade de Tsukuba, em 2007, 2009 e 2014, respectivamente.

De 2014 a 2016, foi bolsista de pós-doutorado no Instituto de Pesquisa de Comunicação Elétrica da Universidade de Tohoku.
De 2016 a 2018, foi pós-doutorado no Instituto Nacional de Ciência e Tecnologia Avançada. Desde 2018, ele é professor assistente na Universidade de Tohoku. Seus interesses de pesquisa incluem movimento ocular, atenção visual e integração multissensorial.

YOSHIYUKI SATO recebeu o diploma de bacharelado pela Universidade de Kyoto, em 2004, e o mestrado e doutorado. diplomas da Universidade de Tóquio, Japão, em 2006 e 2009, respectivamente.

De 2010 a 2016, foi Professor Assistente na Universidade de Eletrocomunicações, Japão. De 2012 a 2013, foi Professor Visitante na Northwestern University, EUA. De 2016 a 2018, foi pesquisador de projetos na Universidade de Tóquio.

Desde 2018, ele é professor assistente especialmente nomeado na Universidade de Tohoku, no Japão. Seus interesses de pesquisa incluem modelagem matemática e de aprendizado de máquina de comportamentos humanos, incluindo percepção, cognição, atenção, funções motoras e comunicações.

SATOSHI SHIOIRI recebeu o diploma de bacharelado em engenharia mecânica e o mestrado e doutorado. diplomas em engenharia de informação física pelo Instituto de Tecnologia de Tóquio, em 1981, 1983 e 1986, respectivamente.

De 1986 a 1989, foi pós-doutorado na Universidade de Montreal.
De 1989 a 1990, foi pós-doutorado no Advanced Telecommunications Research Institute International, Kyoto. Foi professor assistente, professor associado e professor na Universidade de Chiba, de 1990 a 2004. É professor na Universidade de Tohoku, desde 2004. Seus interesses de pesquisa incluem percepção de movimento, percepção de profundidade, percepção de cores, atenção visual, movimento dos olhos e visão para a ação.