As Faces do Engajamento: Automático Reconhecimento do envolvimento do aluno pela Facial Expressões

Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster e Javier R. Movellan

Resumo—O envolvimento dos alunos é um conceito-chave na educação contemporânea, onde é valorizado como um objetivo por si só. Neste artigo, exploramos abordagens para o reconhecimento automático do envolvimento a partir das expressões faciais dos alunos. Estudamos se os observadores humanos podem avaliar com segurança o envolvimento a partir do rosto; analisou os sinais que os observadores usam para fazer esses julgamentos; e automatizou o processo usando aprendizado de máquina. Descobrimos que os observadores humanos concordam de forma confiável ao discriminar graus de engajamento baixo e alto (ÿ de Cohen = 0,96). Quando é necessária uma discriminação fina (4 níveis distintos) a fiabilidade diminui, mas ainda é bastante elevada (ÿ = 0,56). Além disso, descobrimos que os rótulos de engajamento de videoclipes de 10 segundos podem ser previstos com segurança a partir dos rótulos médios de seus quadros constituintes (Pearson r = 0,85), sugerindo que as expressões estáticas contêm a maior parte da informação usada pelos observadores.

Usamos aprendizado de máquina para desenvolver detectores de engajamento automáticos e descobrimos que para classificação binária (por exemplo, alto engajamento versus baixo engajamento), os detectores de engajamento automatizados funcionam com precisão comparável à dos humanos. Finalmente, mostramos que tanto os julgamentos de engajamento humano quanto os automáticos se correlacionam com o desempenho da tarefa. Em nosso experimento, o desempenho pós-teste dos alunos foi previsto com precisão comparável a partir dos rótulos de engajamento (r = 0,47) e das pontuações do pré-teste (r = 0,44).

Termos **de Indexação** —Envolvimento do aluno, reconhecimento de envolvimento, reconhecimento de expressão facial, ações faciais, sistemas de tutoria inteligentes

ÿ

1 INTRODUÇÃO

“O teste para uma educação bem-sucedida não é a quantidade de conhecimento que os alunos adquirem na escola, mas o seu apetite por saber e a sua capacidade de aprender.” Sir Richard Livingstone, 1941 [36].

O envolvimento dos alunos tem sido um tema-chave na literatura educacional desde a década de 1980. O interesse inicial no envolvimento foi impulsionado em parte por preocupações sobre as grandes taxas de abandono escolar e por estatísticas indicando que muitos alunos, estimados entre 25% e 60%, relataram estar cronicamente entediados e desinteressados na sala de aula [32], [51] . Estatísticas como essas levaram as instituições educacionais a tratar o envolvimento dos alunos não apenas como uma ferramenta para melhorar as notas, mas como uma meta independente em si mesma [16].

Hoje em dia, promover o envolvimento dos alunos é relevante não apenas nas salas de aula tradicionais, mas também em outros ambientes de aprendizagem, como jogos educativos, sistemas de tutoria inteligentes (ITS) [43], [4], [52], [30], [4], e cursos online massivamente abertos (MOOCs).

A comunidade de pesquisa em educação desenvolveu várias taxonomias para descrever o envolvimento dos alunos. Fredricks, et al. [20] analisaram 44 estudos e propuseram que existem 3 formas diferentes de engajamento: comportamental, emocional e cognitivo. Anderson, et al. [3] organizou o engajamento em dimensões comportamentais, acadêmicas, cognitivas e psicológicas. O termo envolvimento comportamental é normalmente usado para descrever a disposição do aluno em participar do processo de aprendizagem, por exemplo, assistir às aulas, permanecer concentrado na tarefa, enviar os trabalhos necessários e seguir as orientações do professor. O envolvimento emocional descreve a atitude emocional de um aluno em relação à aprendizagem – é possível, por exemplo, que os alunos executem bem o trabalho que lhes foi atribuído, mas ainda assim não gostem ou fiquem ent Esses alunos teriam alto envolvimento comportamental, mas baixo envolvimento emocional. O envolvimento cognitivo refere-se à aprendizagem de uma forma que maximiza as habilidades cognitivas de uma pessoa, incluindo atenção focada, memória e pensamento criativo [3].

O objetivo de aumentar o envolvimento dos alunos motivou o interesse em métodos para medi-lo [25]. Atualmente, as ferramentas mais populares para medir o envolvimento incluem: (1) Autorrelatos, (2) Listas de verificação observacionais e escalas de classificação e (3) Medições automatizadas.

Autorrelatos: Autorrelatos são questionários nos quais os alunos relatam seu próprio nível de atenção, distração, excitação ou tédio [14], [24], [45]. Essas pesquisas não precisam perguntar diretamente aos alunos quão “envolvidos” eles se sentem, mas podem inferir o envolvimento como uma variável explicativa latente a partir das respostas da pesquisa, por exemplo, usando análise fatorial [41]. Os auto-relatos são sem dúvida úteis. Por exemplo, é interessante saber que entre 25% e 60% dos alunos do ensino médio relatam

• J. Whitehill trabalha no Laboratório de Percepção de Máquinas (MPLab), Universidade da Califórnia, San Diego. JR Movellan trabalha no MPLab e também na Emotient, Inc. Z. Serpell trabalha no Departamento de Psicologia da Virginia Commonwealth University. Sim. Lin e A. Foster trabalham no Departamento de Psicologia da Virginia State University.
E-mail: jake@mplab.ucsd.edu, znserpell@vcu.edu, aysha.foster@gmail.com, linyichen670507@yahoo.com, movellan@emotient.com • O apoio para este trabalho foi fornecido pela NSF concede IIS 0968573 SOCS , IIS INT2-Large 0808767, CNS-0454233 e SMA 1041755 para o Centro de Dinâmica Temporal de Aprendizagem da UCSD, um Centro de Ciência de Aprendizagem da NSF.

ficar entediado e desligado [32], [51]. No entanto, os auto-relatos também têm limitações bem conhecidas. Por exemplo, alguns alunos podem achar “legal” dizer que não estão engajados; outros estudantes podem pensar que é constrangedor dizer isso. Os autorrelatos podem ser influenciados pelos efeitos da memória de primazia e atualidade. Os alunos também podem diferir dramaticamente na sua própria noção do que significa estar envolvido.

Listas de verificação observacionais e escalas de avaliação: Outra forma popular de medir o envolvimento baseia-se em questionários preenchidos por observadores externos, como professores. Esses questionários podem perguntar a opinião subjetiva do professor sobre o grau de envolvimento de seus alunos. Podem também conter listas de verificação para medidas objetivas que supostamente indicam envolvimento. Por exemplo, os alunos ficam sentados em silêncio? Eles fazem a lição de casa? Eles estão na hora certa? Eles fazem perguntas? [48]. Em alguns casos, observadores externos podem avaliar o envolvimento com base em vídeos ao vivo ou pré-gravados de atividades educacionais [46], [29]. Os observadores também podem considerar amostras do trabalho do aluno, como ensaios, projetos e notas de aula [48].

Embora tanto os auto-relatos como as listas de verificação e classificações observacionais sejam úteis, ainda são muito primitivos: carecem de resolução temporal, exigem muito tempo e esforço dos estudantes e observadores e nem sempre estão claramente relacionados com o envolvimento. Por exemplo, métricas de envolvimento como “ficar quieto”, “bom comportamento” e “sem cartões de atraso” parecem medir a conformidade e a vontade de aderir às regras e regulamentos, em vez do envolvimento em si.

Medições automatizadas: A comunidade de sistemas de tutoria inteligentes (ITS) foi pioneira no uso de medidas de engajamento automatizadas e em tempo real. Uma técnica popular para estimar o envolvimento em ITS baseia-se no tempo e na precisão das respostas dos alunos aos problemas práticos e às questões do teste. Esta técnica foi apelidada de “rastreamento de engajamento” [8] em analogia à técnica padrão de “rastreamento de conhecimento” usada em muitos ITS [30]. Por exemplo, o desempenho casual em perguntas fáceis ou tempos de resposta muito curtos podem ser usados como uma indicação de que o aluno não está envolvido e está simplesmente dando respostas aleatórias às perguntas, sem qualquer esforço.
A inferência probabilística pode ser usada para avaliar se os padrões observados de tempo/precisão são mais consistentes com um aluno engajado ou não engajado [8], [26].

Outra classe de medição automatizada de engajamento é baseada em leituras de sensores fisiológicos e neurológicos. Na literatura da neurociência, o envolvimento é normalmente equiparado ao nível de excitação ou alerta.
Medidas fisiológicas como EEG, pressão arterial, frequência cardíaca ou resposta galvânica da pele têm sido usadas para medir o envolvimento e o estado de alerta [23], [18], [39], [49], [9]. No entanto, estas medidas requerem sensores especializados e são difíceis de utilizar em estudos em larga escala.

Um terceiro tipo de reconhecimento automático de engajamento – que é o tema deste artigo – é baseado na visão computacional. A visão computacional oferece a perspectiva de estimar discretamente o envolvimento de um aluno, analisando dicas do rosto [42], [29], [10], [11], postura corporal e gestos das mãos [24], [29]. Embora os métodos baseados na visão para a medição do envolvimento tenham sido utilizados anteriormente pela comunidade ITS, ainda há muito trabalho a ser feito antes que os sistemas automáticos sejam práticos numa ampla variedade de ambientes.

Se for bem sucedido, um sistema de reconhecimento do envolvimento dos alunos em tempo real poderá ter uma ampla gama de aplicações: (1) Os sistemas de tutoria automática poderão utilizar sinais de envolvimento em tempo real para ajustar a sua estratégia de ensino da mesma forma que os bons professores fazem. Os chamados ITS sensíveis ao afeto são um tema quente na comunidade de pesquisa de ITS [13], [59], [5], [19], [26], [12], e alguns dos primeiros sistemas fechados totalmente automatizados. loop ITS que usam sensores afetivos para feedback estão começando a surgir [59], [12]. (2) Os professores humanos em ambientes de ensino à distância poderiam obter feedback em tempo real sobre o nível de envolvimento do seu público. (3) As respostas do público aos vídeos educativos podem ser usadas automaticamente para identificar as partes do vídeo quando o público se desinteressa e para alterá-las adequadamente. (4) Os investigadores educacionais poderiam adquirir grandes quantidades de dados para explorar as causas e variáveis que afetam o envolvimento dos alu
Esses dados teriam resolução temporal muito alta quando comparados ao autorrelato e aos questionários. (5) As instituições educativas poderiam monitorizar o envolvimento dos estudantes e intervir antes que seja tarde demais.

Contribuições: Neste artigo documentamos um dos estudos mais completos até o momento sobre técnicas de visão computacional para reconhecimento automático do envolvimento dos alunos. Em particular, estudamos técnicas de anotação de dados, incluindo a escala de tempo da rotulagem; comparamos algoritmos de visão computacional de última geração para detecção automática de engajamento; e investigamos correlações de engajamento com desempenho de tarefas.

Conceitualização do envolvimento: Nosso objetivo é estimar o envolvimento percebido, ou seja, o envolvimento dos alunos conforme julgado por um observador externo. A lógica subjacente é que, uma vez que os professores dependem do envolvimento percebido para adaptar o seu comportamento de ensino, então a automatização do envolvimento percebido será provavelmente útil para uma vasta gama de aplicações educativas. Nossa hipótese é que grande parte da informação usada pelos humanos para fazer julgamentos de engajamento é baseada no rosto do aluno.

Nosso artigo está organizado da seguinte forma: Primeiro, estudamos se os observadores humanos concordam entre si de maneira confiável ao estimar o envolvimento dos alunos a partir de expressões faciais. Em seguida, usamos métodos de aprendizado de máquina para desenvolver detectores automáticos de engajamento. Investigamos quais sinais são usados pelos detectores automáticos e pelos humanos ao fazer julgamentos de engajamento. Finalmente, investigamos se os julgamentos de engajamento humano e automatizado se correlacionam com o desempenho da ta

2 COLETA E ANOTAÇÃO DO CONJUNTO DE DADOS PARA UM CLASSIFICADOR AUTOMÁTICO DE ENGAJAMENTO

Os dados para este estudo foram coletados de 34 estudantes de graduação que participaram de um “Concurso Cognitivo”.

Treinamento de habilidades" que conduzimos em 2010-2011 [58]. O objetivo deste experimento foi medir a importância, para o ensino, de ver o rosto do aluno. No experimento, dados de desempenho de tarefas sincronizadas e de vídeo foram coletados de sujeitos interagindo com software de treinamento de habilidades cognitivas. O treinamento de habilidades cognitivas gerou um interesse substancial nos últimos anos; o objetivo é aumentar o desempenho acadêmico dos alunos, primeiro melhorando habilidades básicas como memória, velocidade de processamento e lógica e raciocínio. Alguns sistemas proeminentes incluem Brainskills (por Learning RX [1]) e FastForWord (por Scientific Learning [2]). O experimento de treinamento de habilidades cognitivas utilizou software de treinamento de habilidades cognitivas personalizado (que lembra o BrainSkills) que desenvolvemos em nosso laboratório e instalamos em um iPad da Apple. Foi utilizada uma webcam para filmar os alunos; ele foi colocado imediatamente atrás do iPad e apontado diretamente para o rosto do aluno.

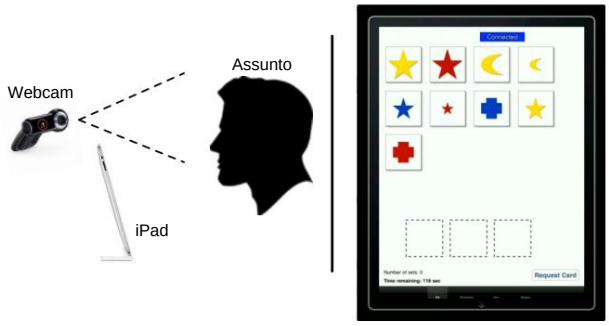

Figura 1. **Esquerda:** Configuração experimental na qual o sujeito joga software de treinamento de habilidades cognitivas em um iPad. Atrás do iPad há uma webcam que grava a sessão. **À direita:** O jogo "Set" na experiência de treino de competências cognitivas que suscitou vários níveis de envolvimento.

O software do jogo no experimento consistia em três jogos – Set, Remember e Sum – que treinavam habilidades lógicas, de raciocínio, perceptivas e de memória. Os jogos foram projetados para serem mentalmente desgastantes. Limites de tempo difíceis foram impostos em cada rodada dos jogos, e os treinadores humanos que controlavam o software do jogo (nas condições Mágico de Oz ou 1 contra 1, conforme descrito abaixo) foram instruídos a "empurrar" os alunos realizem a tarefa mais rapidamente. Neste sentido, o domínio de treino de competências cognitivas da nossa experiência pode assemelhar-se a um cenário em que um aluno está a fazer um exame stressante.

Em termos de ambiente físico, o ITS típico e o cenário de habilidades cognitivas em nosso estudo são muito semelhantes – um aluno senta-se diretamente na frente de um computador ou iPad, e uma câmera web recupera o vídeo frontal do aluno. É possível que o aparecimento de estados afetivos, como o engajamento, possa diferir entre o treinamento de habilidades cognitivas e as interações ITS. No entanto, é provável que a metodologia de rotulagem e as técnicas de visão computacional para treinar classificadores automatizados ainda possam ser generalizadas para casos de uso de ITS mais tradicionais.

As variáveis dependentes durante o experimento 2010-2011 foram o desempenho pré e pós-teste no jogo Set. O jogo "Set" em nosso estudo (ver Figura 1 **à direita)** era muito semelhante ao jogo de cartas clássico: é mostrado ao aluno um tabuleiro com 9 cartas, cada uma das quais pode variar em três dimensões: tamanho, forma e cor. O objetivo é formar o maior número possível de conjuntos válidos de 3 cartas no tempo previsto. Um conjunto é válido se e somente se as três cartas do conjunto forem todas iguais ou diferentes para cada dimensão. Depois de formar um conjunto válido, as três cartas desse conjunto são removidas do tabuleiro e três novas cartas são distribuídas. Este processo continua até que o tempo decorra.

Os dados experimentais para o estudo de engajamento neste artigo foram retirados de 34 sujeitos de dois grupos: (a) os 26 sujeitos que participaram da versão da primavera de 2011 do estudo de treinamento de habilidades cognitivas em uma faculdade/universidade historicamente negra (HBCU) em o sul dos Estados Unidos. Todos esses sujeitos eram afro-americana e 20 eram mulheres. Dados adicionais foram coletados de (b) os 8 indivíduos que participaram da versão do verão de 2011 do estudo de Treinamento de Habilidades Cognitivas em uma universidade na Califórnia (UC), todos eles asiático-americanos ou caucasianos-americanos, e 5 dos quais eram mulheres. O software de jogo usado no conjunto de dados UC era idêntico ao software usado no

HBCU, exceto por pequenas diferenças nos parâmetros do jogo (por exemplo, a rapidez com que as cartas são distribuídas). Para o presente estudo, os dados do HBCU serviram como fonte primária de dados para treinar e testar o reconhecedor de engajamento. O conjunto de dados UC permitiu-nos avaliar quão bem o sistema treinado seria generalizado para sujeitos de uma raça diferente – um problema conhecido nos modernos sistemas de visão computacional.

Na configuração experimental, cada sujeito sentou-se em uma sala privada e jogou o software de treinamento de habilidades cognitivas sozinho ou junto com o experimentador. O iPad foi colocado sobre um suporte e situado horizontalmente a aproximadamente 30 centímetros à frente do rosto do sujeito e verticalmente para que o iPad ficasse um pouco abaixo do nível dos olhos. Atrás do iPad apontando para o sujeito estava uma webcam Logitech que gravou toda a sessão.

Conforme descrito em [58], cada sujeito foi atribuído a uma condição Mágico de Oz ou a uma condição individual. Na condição Mágico de Oz, o sujeito ficava sentado sozinho na sala enquanto interagia com o software do jogo, que era controlado remotamente por um bruxo humano que podia observar o aluno em tempo real. Na condição 1 contra 1, o sujeito jogava ao lado de um treinador humano que controlaria o software abertamente. No conjunto de dados HBCU, 20 indivíduos estavam na condição Mágico de Oz e 6 indivíduos estavam na condição 1 contra 1. No conjunto de dados da UC, todos os sujeitos estavam na condição do Mágico de Oz.

Durante cada sessão, o sujeito deu consentimento informado e depois assistiu a um vídeo de 3 minutos no iPad explicando os objetivos dos três jogos e como jogá-los. O sujeito então fez um pré-teste de 3 minutos no jogo Set para medir o desempenho inicial.

O desempenho do teste foi medido como o número de

"conjuntos" de 3 cartas (de acordo com as regras do jogo) que o aluno poderia formar em 3 minutos. As cartas específicas distribuídas durante o teste foram as mesmas para todas as disciplinas.
Após o pré-teste, o sujeito passou por 35 minutos de treinamento de habilidades cognitivas utilizando o software de treinamento.
O objetivo do treinador (nas condições Mágico de Oz e 1 contra 1) era ajudar o aluno a maximizar seu desempenho no teste de Set. Durante a sessão de treinamento, o treinador pode alterar a dificuldade da tarefa, alternar tarefas e fornecer instruções motivacionais. Após o período de treinamento, o sujeito realizou um pós-teste no Set e em seguida foi feito.

2.1 Anotação de dados

Dados os vídeos gravados das sessões de treinamento cognitivo, o próximo passo foi rotulá-los para engajamento.
Organizamos uma equipe de rotuladores composta por estudantes de graduação e pós-graduação de ciência da computação, ciências cognitivas e psicologia das duas universidades onde os dados foram coletados. Esses rotuladores visualizaram e avaliaram os vídeos quanto à aparência de engajamento.
Observe que nem todos os rotuladores rotularam exatamente os mesmos conjuntos de imagens/vídeos. Em vez disso, optamos por equilibrar os objetivos de obter muitos rótulos por imagem/vídeo e anotar uma grande quantidade de dados para desenvolver um detector automatizado. Ao rotular os vídeos, o áudio foi desligado e os rotuladores foram instruídos a rotular o envolvimento com base apenas na aparência.

Em contraste com os domínios mais minuciosamente estudados de reconhecimento automático de emoções básicas (feliz, triste, zangado, enojado, com medo, surpreso ou neutro) [33], [6], [61] ou classificação de unidade de ação facial [37], [27], [7], [47] (do Facial Action Coding System [17]), estados afetivos que são relevantes para a aprendizagem, como frustração ou envolvimento, podem ser difíceis de definir claramente [50]. Portanto, chegar a uma definição suficientemente clara e elaborar um procedimento de rotulagem apropriado, incluindo o prazo em que a rotulagem ocorre, é importante para garantir a confiabilidade e a validade dos rótulos de treinamento [50]. Na experiência piloto, tentamos três abordagens diferentes para rotulagem: 1) Assistir a videoclipes (na velocidade normal de visualização) e atribuir rótulos de envolvimento contínuo pressionando as teclas de seta para cima/para baixo.

2) Assistir a videoclipes e fornecer um único número para avaliar o vídeo inteiro.

3) Visualizar imagens estáticas e fornecer um único número para avaliar cada imagem.

Achamos a abordagem (1) muito difícil de executar na prática. Um problema foi a tendência de se habituar ao nível recente de envolvimento de cada sujeito e de ajustar a classificação atual em relação ao nível médio de envolvimento desse sujeito no passado recente. Isto poderia gerar rótulos que não são diretamente comparáveis entre sujeitos ou mesmo dentro de sujeitos. Outro problema era como classificar eventos curtos, por exemplo, fechar brevemente os olhos ou olhar para o lado: deveriam esses breves momentos ser rotulados como "não envolvimento", ou deveriam ser ignorados como comportamento normal se o caso contrário, o assunto parece altamente engajado? Finalmente, era difícil fornecer rótulos contínuos que fossem sincronizados no tempo com o vídeo; a sincronização adequada exigiria primeiro a varredura do vídeo em busca de eventos interessantes e, em seguida, assisti-lo novamente e ajustar cuidadosamente o envolvimento para cima ou para baixo a cada momento.

Descobrimos que a tarefa de rotulagem era mais fácil usando as abordagens (2) e (3), desde que fossem dadas instruções claras sobre o que constitui "engajamento".

2.2 Categorias e instruções de engajamento Dada a abordagem de fornecer um único número de engajamento para um videoclipe ou imagem inteira, decidimos pela seguinte escala aproximada para avaliar o engajamento:

1: Nada envolvido – por exemplo, desviando o olhar do computador e obviamente sem pensar na tarefa, olhos completamente fechados.

2: Nominalmente engajado – por exemplo, olhos mal abertos, claramente não "empenhados" na tarefa.

3: Envolvido na tarefa – o aluno não precisa de nenhuma advertência para "permanecer na tarefa".

4: Muito engajado – o aluno poderia ser "elogiado" pelo seu nível de envolvimento na tarefa.

X: O clipe/quadro não estava muito claro ou não continha nenhuma pessoa.

Imagens de exemplo para cada nível de engajamento são mostradas na Figura 2. Observe que essas diretrizes certamente pertencem ao "engajamento comportamental" [20], mas também contêm elementos de engajamento cognitivo e emocional. Por exemplo, se um aluno está ou não "interessado" na tarefa está relacionado com a sua atitude em relação à tarefa de aprendizagem.
Além disso, nas nossas definições acima, a distinção entre os níveis de envolvimento 3 e 4 está relacionada com o estado motivacional do aluno.

Os rotuladores foram instruídos a rotular clipes/imagens de acordo com "Quão engajado o sujeito parece estar". A chave aqui é a palavra aparecer – propositalmente não queríamos que os rotuladores tentassem inferir o que estava "realmente" acontecendo dentro dos cérebros dos alunos porque isso deixava o problema da rotulagem muito aberto. Isto tem como consequência que, se um sujeito piscasse, ele seria rotulado como muito não engajado (Engajamento = 1) porque, naquele instante, ele parecia não estar engajado. Na prática, descobrimos que isso tornou a tarefa de rotulagem mais clara para os rotuladores e ainda rendeu rótulos de engajamento informativos. Se a média das pontuações de engajamento de vários quadros for calculada ao longo de um videoclipe (consulte a Seção 2.4), as piscadas momentâneas não afetarão muito a pontuação média, de qualquer maneira. Além disso, os rotuladores foram instruídos a avaliar o envolvimento com base no conhecimento de que os participantes estavam interagindo com o software de treinamento em um iPad diretamente à sua frente.

Qualquer olhar ao redor da sala ou para outra pessoa (ou seja, o experimentador) deveria ser considerado sem envolvimento (classificação 1) porque implicava que o sujeito não estava envolvido com o iPad. (Esses momentos ocorreram bem no início ou no final de cada sessão, quando o

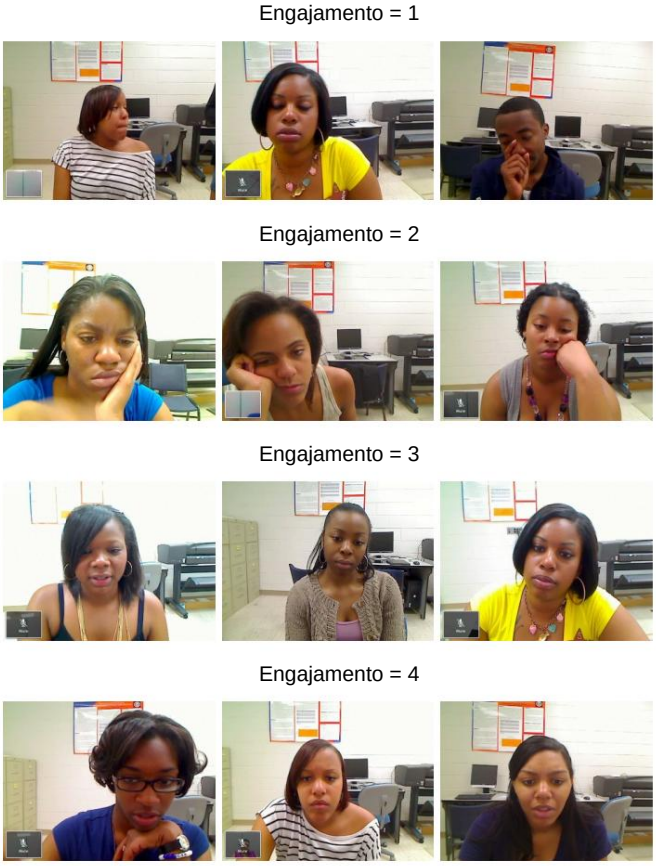

Figura 2. Exemplos de rostos para cada nível de envolvimento dos sujeitos da HBCU. Todos os sujeitos deram consentimento por escrito para a publicação de suas imagens faciais.

O experimentador estava montando ou desmontando o experimento.) O objetivo aqui era ajudar o sistema a generalizar para uma variedade de ambientes onde os alunos deveriam olhar diretamente para frente deles.

2.3 Prazo

Uma variável importante na anotação de vídeo é a escala de tempo em que a rotulagem ocorre. Para a abordagem (2) (descrita na Seção 2.1), experimentamos duas escalas de tempo diferentes: clipes de 60 segundos e clipes de 10 segundos. A abordagem (3) (imagens únicas) pode ser vista como o limite inferior da duração de um videoclipe. Em um experimento piloto, comparamos essas três escalas de tempo quanto à confiabilidade entre codificadores. Como métrica de desempenho utilizamos o ÿ de Cohen (ver Apêndice para mais detalhes). Como os rótulos de engajamento pertencem a uma escala ordinal ({1, 2, 3, 4}) e não são simplesmente categorias, utilizamos um ÿ ponderado com pesos quadráticos para penalizar a discordância dos rótulos.

Para a tarefa de rotulagem de 60 segundos, todas as sessões de vídeo (ÿ 45 minutos/sujeito) dos participantes da HBCU foram assistidas do início ao fim em clipes de 60 segundos, e 2 rotuladores inseriram uma única pontuação de envolvimento após visualizar cada clipe. Para a tarefa de rotulagem de 10 segundos, 505 videoclipes de 10 segundos cada foram extraídos em momentos aleatórios dos vídeos da sessão e mostrados a 7 rotuladores em ordem aleatória (em termos de tempo e assunto). Entre os clipes de 60 segundos e as tarefas de rotulagem de 10 segundos, achamos a tarefa de rotulagem de 10 segundos mais intuitiva. Ao visualizar os clipes mais longos, era difícil saber que rótulo dar se o sujeito parecesse não engajado no início, mas parecesse altamente engajado no final. A confiabilidade entre codificadores da tarefa de rotulagem de clipes de 60 segundos foi ÿ = 0,39 (em 2 rotuladores); para a tarefa de rotulagem de clipe de 10 segundos ÿ = 0,68 (em 7 rotuladores).

Para a abordagem (3), criamos um software de etiquetagem personalizado no qual 7 etiquetadores anotaram lotes de 100 imagens cada. As imagens de cada lote eram quadros de vídeo extraídos em momentos aleatórios dos vídeos da sessão. Cada lote continha um conjunto aleatório de imagens abrangendo vários pontos no tempo de vários assuntos. Os rotuladores avaliaram cada imagem individualmente, mas puderam visualizar muitas imagens e seus rótulos atribuídos simultaneamente na tela. O software de etiquetagem também forneceu um botão Classificar para classificar as imagens em ordem crescente por rótulo de envolvimento.
Na prática, descobrimos que este é um método intuitivo e eficiente de rotular imagens para a aparência de engajamento. A confiabilidade intercodificador para rotulagem baseada em imagem foi ÿ = 0,56. Essa confiabilidade também pode ser aumentada calculando-se a média dos rótulos baseados em quadros em vários quadros consecutivos no tempo (ver Seção 2.4).

2.4 Informações estáticas versus informações de movimento

Uma questão interessante é quanta informação sobre o envolvimento dos alunos é capturada nos pixels estáticos dos quadros de vídeo individuais em comparação com a dinâmica do movimento. Realizamos um estudo piloto para examinar esta questão. Em particular, selecionamos aleatoriamente 120 videoclipes (10 segundos cada) do conjunto de todos os vídeos da HBCU.
A amostra aleatória continha clipes de 24 sujeitos.
Cada clipe foi então dividido em 40 quadros espaçados de 0,25 segundos. Esses quadros foram então embaralhados no tempo e entre os assuntos. Um rotulador humano rotulou esses quadros de imagem para a aparência de engajamento, conforme descrito na “abordagem (3)” da Seção 2.1. Finalmente, os valores de engajamento atribuídos a todos os quadros de um clipe específico foram remontados e calculada a média; essa média serviu como uma estimativa da pontuação de engajamento “verdadeira” dada pelo mesmo rotulador ao visualizar aquele videoclipe conforme descrito na “abordagem (2)” acima.
Descobrimos que, com relação às pontuações de engajamento verdadeiro, as pontuações estimadas deram ÿ = 0,78 e uma correlação de Pearson r = 0,85.
Essa precisão é bastante alta e sugere que a maior parte da informação sobre a aparência do engajamento está contida nos pixels estáticos, e não no movimento em si.

Também examinamos os videoclipes nos quais as pontuações de engajamento reconstruídas diferiam mais das pontuações verdadeiras. Em particular, classificamos os 120 videoclipes rotulados em ordem decrescente de desvio absoluto do rótulo estimado (pela média dos rótulos baseados em quadros) do rótulo “verdadeiro” dado ao videoclipe visto como um todo. Em seguida, examinamos esses clipes e tentamos explicar a discrepância: No primeiro clipe (maior desvio absoluto), a sujeito balançava a cabeça

de um lado para o outro como se estivesse ouvindo música (embora ela não estivesse). É provável que o codificador tenha tratado isso como um comportamento não engajado. Esse comportamento pode ser difícil de capturar a partir de julgamentos de quadros estáticos. No entanto, também foi um caso anômalo.

No segundo clipe, o sujeito virou a cabeça para o lado para olhar para o experimentador, que conversava com ele por vários segundos. Nos julgamentos ao nível do quadro, isto foi percebido como fora da tarefa e, portanto, como um comportamento não engajado; isso corresponde às instruções dadas aos codificadores de que eles avaliam o engajamento sob a suposição de que o sujeito deve estar sempre olhando para o iPad. Para a gravadora de videoclipe, entretanto, o codificador julgou que o aluno estava altamente engajado porque estava ouvindo atentamente o experimentador. Este é um exemplo de inconsistência por parte do codificador quanto ao que constitui envolvimento e não indica necessariamente um problema com a divisão dos clipes em quadros.

Finalmente, em vários clipes, os sujeitos às vezes desviavam o olhar para baixo, para olhar a parte inferior da tela do iPad. No nível do quadro, era difícil distinguir o sujeito olhando para a parte inferior do iPad do sujeito olhando para o próprio colo ou mesmo fechando os olhos, ambos os quais seriam considerados falta de envolvimento. A partir do vídeo foi mais fácil distinguir esses comportamentos do contexto. No entanto, esses eventos de olhar para baixo eram raros e podem ser efetivamente filtrados por meio de uma média simples.

Apesar desses problemas, a precisão relativamente alta da estimativa de rótulos baseados em vídeo a partir de rótulos baseados em quadros sugere uma abordagem para construir um classificador automático de engajamento: em vez de analisar videoclipes como vídeos, divida-os em seus quadros de vídeo, e, em seguida, combine as estimativas de engajamento para cada quadro. Usamos essa abordagem para rotular os dados da HBCU e da UC para engajamento. Na próxima seção, descrevemos nossa arquitetura proposta para reconhecimento automático de engajamento com base neste design quadro a quadro.

3 ARQUITETURAS DE RECONHECIMENTO AUTOMÁTICO

Com base na descoberta da Seção 2.4 de que os rótulos baseados em videoclipes podem ser estimados com alta fidelidade simplesmente calculando a média dos rótulos baseados em quadros, concentramos nosso estudo no reconhecimento **quadro a quadro** do envolvimento dos alunos. Isso significa que muitas técnicas desenvolvidas para classificação de emoções e unidades de ação facial podem ser aplicadas ao problema de reconhecimento de engajamento. Neste artigo propusemos um pipeline de 3 estágios.

1) Registro facial: as posições do rosto e dos marcos faciais (olhos, nariz e boca) são localizadas automaticamente na imagem; as coordenadas da caixa facial são calculadas; e o patch do rosto é cortado da imagem [35]. Experimentamos resoluções faciais de 36 × 36 e 48 × 48 pixels.

2) O patch facial recortado é classificado por quatro classificadores binários, um para cada categoria de engajamento l ÿ {1, 2, 3, 4}.

3) As saídas dos classificadores binários são alimentadas em um regressor para estimar o nível de engajamento da imagem.

O estágio (1) é padrão para análise facial automática e nossa abordagem específica é descrita em [35]. A etapa (2) é discutida na próxima subseção, e a etapa (3) é discutida na Seção 3.11. Essa arquitetura é uma reminiscência de um sistema automatizado de estimativa de pose de cabeça que desenvolvemos anteriormente [57], que combina as saídas de vários classificadores binários para formar um julgamento de valor real.

3.1 Classificação binária

Treinamos 4 classificadores binários de engajamento – um para cada um dos 4 níveis descritos na Seção 2.1. A tarefa de cada um desses classificadores é discriminar uma imagem (ou quadro de vídeo) que pertence ao nível de engajamento l de uma imagem que pertence a algum outro nível de engajamento l = l. Chamamos esses detectores de 1-v-outro, 2-v-outro, etc.

Comparamos três combinações de tipo de recurso + classificador comumente usadas e comprovadamente eficazes da literatura de reconhecimento automático de expressão facial:

- GentleBoost com recursos de Box Filter **(Boost(BF)):** esta é a abordagem popularizada por Viola e Jones em [53] para detecção de rostos. • Máquinas de vetores de suporte com recursos Gabor **(SVM(Gabor)):** esta abordagem alcançou algumas das mais altas precisões na literatura para ação facial e classificação básica de emoções [35]. • Regressão logística multinomial com resultados de expressão do Computer Expression Recognition Toolbox [35] **(MLR(CERT)):** aqui, tentamos aproveitar um sistema automatizado existente para análise de expressão facial para treinar classificadores de engajamento.

Nosso objetivo não é julgar a eficácia de cada tipo de recurso (ou de cada método de aprendizagem) isoladamente, mas sim avaliar a eficácia dessas arquiteturas de visão computacional de última geração para uma nova tarefa de visão. Como relativamente pouca investigação ainda examinou a forma de reconhecer os estados emocionais específicos dos alunos em ambientes reais de aprendizagem, é uma questão em aberto até que ponto estes métodos funcionariam bem para o reconhecimento do envolvimento. Descrevemos cada abordagem com mais detalhes abaixo.

3.1.1 Impulso(GC)

Os recursos de filtro de caixa (BF) medem diferenças na intensidade média de pixels entre regiões retangulares vizinhas de uma imagem. Eles demonstraram ser altamente eficazes para detecção automática de rosto [53], bem como detecção de sorriso [56]. Por exemplo, para detectar rostos, um filtro de caixa de 2 retângulos pode capturar o fato de que a região dos olhos do rosto é normalmente mais escura do que a parte superior das bochechas. Em tempo de execução, os recursos BF são rápidos de extrair usando a técnica de “imagem integral” [53]. No momento do treinamento, entretanto, o número de características BF relativas à resolução da imagem é muito alto comparado a outras representações de imagem (por exemplo, uma decomposição de Gabor), o que pode levar ao sobreajuste.

Os recursos BF são normalmente combinados com um classificador reforçado, como Adaboost [21] ou GentleBoost (Boost) [22],

Machine Translated by Google

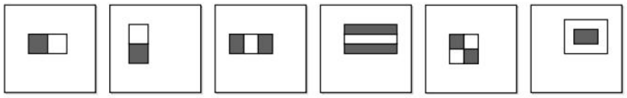

Figura 2. Cada filtro é calculado tomando a diferença das somas dos pixels nas caixas brancas e nas caixas cinza. Os tipos de filtro

3. Recursos do Box Filter (BF), também conhecidos como filtros wavelet do tipo Haar, que foram usados no estudo.

Se o desempenho for apenas ligeiramente superior ao acaso, o sistema combinado (ou seja, o classificador forte) pode atingir níveis muito elevados de precisão.

Na Ref. 12, foi demonstrado que os métodos boosting podem ser reinterpretados do ponto de vista da estimativa de máxima verossimilhança sequencial, uma interpretação que

que realiza a seleção de recursos durante o treinamento

A implementação no GentleBoost é realizada escolhendo sequencialmente classificadores fracos e classificação real em tempo de execução. Em nosso GentleBoost combina-los para minimizar uma função de erro qui-quadrado. Em nossa aplicação, cada

implementação, cada aluno fraco consiste em um não-regressor paramétrico suavizado com um kernel gaussiano de largura de banda ÿ, para estimar a razão log-verossimilhança do rótulo da classe dado o valor do recurso. Cada gentil

de filtros, seguido por uma função de transferência não linear.

O conjunto de filtros que usamos é o mesmo usado na Ref. 24, com o

O classificador de reforço foi treinado para 100 rodadas de reforço.

de uso geral, sem a necessidade de hardware especializado (ver Refs. 23 e 24 para explicação detalhada). Na Ref. 24, a função de transferência não linear era uma função de limite simples cuja

Para os recursos, incluímos 6 tipos de filtros de caixa em

use uma função constante por partes cujos parâmetros são escolhidos pelo algoritmo GentleBoost.

total, compreendendo recursos de dois, três e quatro retângulos semelhantes aos usados em [53], e um recurso adicional de dois retângulos "centro-surround" (ver Figura 3). Com uma resolução de imagem facial

Isso permite que cada classificador fraco produza valores reais arbitrários no intervalo [y1, +1] em vez de simplesmente dos valores

de 48×48 pixels, havia 5397601 características BF; com uma resolução facial de 36 × 36 pixels, aqui

3. Descrição do banco de dados

descrevemos brevemente alguns dos bancos de dados mais comumente usados para treinar e testar

detectores de características faciais. FERET (imagens frontais) 22 e gratuito, publicado, com 1681109 recursos.

disponíveis comercialmente. XM2VTS

fundo e pouca variação na iluminação. XM2VTS29 e BANCA-C/WorldModel1 são bancos de dados contém 1180 3.1.2 SVM (Gabor) imagens de rosto frontal de alta qualidade. BANCA-C/WorldModel contém 2380 filtros de energia Gabor de face frontal [44] são filtros passa-banda com um

imagens sem confusão de fundo e alguma variação na iluminação. Imagens em os bancos de dados acima mencionados foram obtidos em ambientes controlados com uniformidade fundo.

orientação espacial e frequência ajustáveis. Eles modelam as células complexas do córtex visual dos primatas. Quando aplicados a imagens, eles respondem às bordas em determinados orientações, por exemplo, bordas horizontais devido ao enrugamento de na testa, ou bordas diagonais devido aos "pés de galinha" ao redor dos olhos. Os filtros de energia Gabor têm um histórico comprovado em uma ampla variedade de aplicações de processamento facial, incluindo reconhecimento facial [31] e reconhecimento de expressão facial [35]. Em aplicações de aprendizado de máquina, os recursos Gabor são frequentemente classificados por uma máquina de vetores de suporte linear de margem suave (SVM) com parâmetro C especificando quanto exemplos de treinamento mal classificados devem penalizar a função objetivo. Em nossa implementação, aplicamos um "banco" de 40 Filtros de Energia Gabor composto por 8 orientações (espaçadas em intervalos de 22,5 graus) e 5 frequências espaciais variando de 2 a 32 ciclos por face. O número total de recursos Gabor é N × N × 8 × 5, onde N é a largura da imagem facial em pixels.

As bases de dados BANCA-D/A e BioID tentam simular condições do mundo real. BANCA-D/A contém 4.160 imagens de rosto frontal com fundos desordenados, iluminações variáveis e variação da pose da cabeça. O gratuito e publicamente disponível O banco de dados BioID contém 1.521 imagens de rosto frontal que variam em relação à iluminação, tamanho, origem, escala e pose da cabeça. Com base nos estudos da literatura 9,15,32

3.1.3 MLR(CERT)

O Sistema de Codificação de Ação Facial [17] é uma estrutura abrangente para descrever objetivamente a expressão facial em termos de Unidades de Ação, que medem a intensidade de mais de 40 músculos faciais distintos. A codificação manual FACS foi usada anteriormente para estudar o envolvimento do aluno e outras emoções relevantes para o ensino automatizado [28], [42].

Em nosso estudo, por estarmos interessados no reconhecimento automático de engajamento, empregamos o Computer Expression Recognition Toolbox (CERT), que é uma ferramenta de software desenvolvida por nosso laboratório para estimar automaticamente as intensidades de ação facial [35]. Embora as precisões dos classificadores de ação facial individuais variem, descobrimos que o CERT é útil para uma variedade de análises faciais tarefas, incluindo a discriminação entre dor real e dor falsa [34], detecção de fadiga do motorista [54] e estimativa da percepção dos alunos sobre dificuldade curricular [55]. O CERT gera estimativas de intensidade de 20 ações faciais, bem como a pose 3-D da cabeça (guinada, inclinação e rotação). Para reconhecimento de engajamento, classificamos os resultados do CERT usando regressão logística multinomial (MLR), treinada com um regularizador L2 no vetor de peso de força ÿ. Usamos o valor absoluto de guinada, inclinação e rotação para fornecer invariância à direção da mudança de pose.

Como estamos interessados em sistemas em tempo real que possam operar sem basear o detector em um assunto específico, usamos as saídas CERT brutas (ou seja, não fazemos pontuação z das saídas) em nossos experimentos.

Internamente, o CERT utiliza a abordagem SVM(Gabor) descrita acima. Como o CERT foi treinado em centenas a milhares de indivíduos (dependendo do canal de saída específico), o que é substancialmente maior do que o número de indivíduos coletados para este estudo, é possível que os resultados do CERT forneçam uma representação independente de identidade do rostos dos alunos, o que pode aumentar o desempenho de generalização.

3.2 Seleção de dados

Começamos com um conjunto de 13.584 quadros do conjunto de dados HBCU. Em seguida, aplicamos o seguinte procedimento para selecionar dados de treinamento e teste para cada classificador binário para distinguir lv-other:

1) Se o rótulo mínimo e máximo dado a uma imagem diferisse em mais de 1 (por exemplo, um rotulador atribui um rótulo de 1 e outro atribui um rótulo de 3), então a imagem foi descartada. Isso reduziu o pool de 13.584 para 9.796 imagens.

2) Se o detector automático de rosto (do CERT [35]) não conseguisse detectar um rosto, ou se o maior rosto detectado tivesse menos de 36 pixels de largura (geralmente indicativo de uma detecção de rosto errônea), a imagem era descartada. Isso reduziu o pool de 9.796 para 7.785 imagens.

3) Para cada uma das imagens rotuladas, consideramos o conjunto de todos os rótulos dados a essa imagem por todos os rotuladores. Se algum rotulador marcasse o quadro como X (sem face ou muito pouco claro), a imagem era descartada. Isso reduziu o pool de 7.785 para 7.574 imagens.

4) Caso contrário, o rótulo de "verdade básica" para cada imagem foi calculado arredondando o rótulo médio daquela imagem para o número inteiro mais próximo (por exemplo, 2,4 arredondamentos para 2; 2,5 arredondamentos para 3). Se o rótulo arredondado fosse igual a l, então aquela imagem era considerada um exemplo positivo para o conjunto de treinamento do classificador lv-other; caso contrário, foi considerado um exemplo negativo.

No total, houve 7.574 quadros do conjunto de dados HBCU e 16.711 do conjunto de dados UC selecionados usando esta abordagem. As distribuições de engajamento na HBCU foram de 6,03%, 9,72%, 46,28% e 37,97% para os níveis de engajamento 1 a 4, respectivamente. Para UC, foram 5,37%, 8,46%, 42,31% e 43,85%, respectivamente.

3.3 Validação cruzada

Usamos validação cruzada independente de sujeito de 4 vezes para medir a precisão de cada classificador binário treinado.
O conjunto de todos os quadros rotulados foi particionado em 4 dobras de modo que nenhum sujeito aparecesse em mais de uma dobra; portanto, a estimativa de desempenho da validação cruzada dá uma ideia de quão bem o classificador funcionaria em um assunto novo no qual o classificador não foi treinado.

3.4 Métricas de precisão

Usamos a métrica 2AFC [60], [40] para medir a precisão, que expressa a probabilidade de discriminar corretamente um exemplo positivo de um exemplo negativo em uma tarefa de classificação de escolha forçada de 2 alternativas. O 2AFC é uma estimativa imparcial da área sob a curva Receiver Operating Characteristics, que é comumente usada na literatura de reconhecimento de expressões faciais (por exemplo, [37]). Um valor 2AFC de 1 indica discriminação perfeita, enquanto 0,5 indica que o classificador está “ao acaso”.

Além disso, também calculamos o ÿ de Cohen com ponderação quadrática. Para comparar a precisão da máquina com a precisão inter-humana, calculamos também o 2AFC e o ÿ para rotuladores humanos, usando os mesmos critérios de seleção de imagens descritos na Seção 3.2. Ao calcular ÿ para os classificadores binários automáticos, otimizamos ÿ em todos os limites possíveis das saídas do detector.

3.5 Seleção de hiperparâmetros

Cada um dos classificadores listados acima possui um hiperparâmetro associado a ele (seja ÿ, C ou ÿ). A escolha do hiperparâmetro pode impactar substancialmente a precisão do teste, e é uma armadilha comum fornecer uma estimativa excessivamente otimista da precisão de um classificador ajustando manualmente o hiperparâmetro com base no desempenho do conjunto de teste. Para evitar essa armadilha, em vez disso, otimizamos os hiperparâmetros usando apenas o conjunto de treinamento, dividindo ainda mais cada conjunto de treinamento em 4 dobras internas de validação cruzada independentes do sujeito em um paradigma de validação cruzada dupla. Selecionamos hiperparâmetros dos seguintes conjuntos de valores: ÿ ÿ {10ÿ2 , 10ÿ1,5 ,. . . , 100}, C ÿ {0,1, 0,5, 2,5, 12,5, 62,5, 312,5} e ÿ ÿ 10ÿ4 {10ÿ5 , , . . . , 10+5}.

3.6 Resultados: Classificação Binária

Os resultados da classificação são mostrados na Tabela 1 para resolução de face recortada de 48 × 48 pixels. Cada célula relata a precisão (2AFC) calculada em média em 4 dobras de validação cruzada, juntamente com o desvio padrão entre parênteses. As precisões na resolução de 36 × 36 pixels foram ligeiramente inferiores. Todos os resultados são para classificação independente de assunto.

Na parte superior da Tabela 1, vemos que a precisão da classificação binária dada pelos classificadores de máquina é muito semelhante à precisão inter-humana. Todas as três arquiteturas testadas forneceram desempenho semelhante em média nas quatro tarefas (1-v-outro, 2-v-outro, etc.). No entanto, MLR(CERT) teve pior desempenho em 1-v-diferente dos outros classificadores. Como discutimos em Na seção 4, muitas imagens rotuladas como Engajamento = 1 exibem fechamento dos olhos. É possível que o detector de fechamento ocular do CERT seja relativamente impreciso e, em comparação, as abordagens Boost(BF) e SVM(Gabor) são capazes de aprender um detector preciso de fechamento ocular a partir dos próprios dados de treinamento. Por outro lado, o CERT tem um desempenho melhor do que as outras abordagens para 4-v-other. Conforme descrito na Seção 4, Engajamento = 4 pode ser discriminado usando informações de pose. Aqui, o CERT pode ter uma vantagem porque o detector de pose do CERT foi treinado em dezenas de milhares de indivíduos.

Precisão 2AFC (e desenvolvimento padrão) – treinar em HBCU, testar em HBCU

	Classificador			
Tarefa	**MLR(CERT)**	**Boost(BF)**	**SVM(Gabor)**	**Humano**
1-v-outro	0,862 (0,061)	0,965 (0,012)	0,914 (0,031)	0,909 (0,021)
2-v-outro	0,721 (0,130)	0,709 (0,130)	0,711 (0,038)	0,620 (0,143)
3-v-outro	0,574 (0,045)	0,607 (0,065)	0,630 (0,075)	0,606 (0,070)
4-v-outros	0,697 (0,076)	0,632 (0,111)	0,660 (0,127)	0,650 (0,068)
Média	0,714	0,728	0,729	0,696

Precisão 2AFC (e desenvolvimento padrão) - treinar em HBCU, testar em UC

	Classifier		
Task	**MLR (CERT)**	**Boost (BF)**	**SVM (Gabor)**
1-v-other	0,782 (0,120)	0,845 (0,111)	0,831 (0,091)
2- v-outro	0,682 (0,095)	0,597 (0,102)	0,668 (0,067)
3-v-outro	0,507 (0,058)	0,464 (0,063)	0,570 (0,055)
4-v-outro	0,613 (0,108)	0,469 (0,153)	97 (0,041)
Média	0,646	0,594	0,691

TABELA 1

Parte superior: Precisão de reconhecimento de engajamento dentro do conjunto de dados (HBCU), independente do assunto (métrica 2AFC) para cada nível de engajamento l ÿ {1, 2, 3, 4} usando cada uma das três arquiteturas de classificação, juntamente com classificação inter-humana precisão. **Parte inferior:** Precisão do reconhecimento de engajamento em um conjunto de dados diferente (UC) não usado para

ÿ de Cohen (e desenvolvimento padrão) - treinar em HBCU, testar em

	Classificador			
Tarefa	HBCU	**MLR (CERT)**	**Boost (BF)**	**SVM (Gabor)**	**Humano**
1-v-outro	0,393 (0,094)	0,662 (0,060)	0,528 (0,167)	0,629 (0,245)
2-v-outro	0,254 (0,209)	0,246 (0,164)	0,222 (0,132)	0,272 (0,260)
3-v-outro	0,154 (0,063)	0,193 (0,090)	0,213 (0,109)	0,209 (0,154)
4-v-outros	0,301 (0,098)	0,214 (0,119)	0,261 (0,135)	0,256 (0,109)
Média	0,275	0,329	0,306	0,341

ÿ de Cohen (e desenvolvimento padrão) - treinar em HBCU, testar em

	UC Classifier		
Tarefa	**MLR (CERT)**	**Boost (BF)**	**SVM (Gabor)**
1-v-outro	0,329 (0,236)	0,400 (0,258)	0,414 (0,261)
2-v-outro	0,123 (0,100)	0,068 (0,040)	0,154 (0,134)
3-v -outros	0,078 (0,049)	0,027 (0,022)	0,096 (0,054)
4-v-outros	0,137 (0,123)	0,063 (0,058)	0,260 (0,125)
Média	0,167	0,140	0,231

TABELA 2

Semelhante à Tabela 1, mas mostra ÿ de Cohen em vez de 2AFC.

Acima: conjunto de testes HBCU. **Abaixo:** conjunto de generalização UC.

Como a confiabilidade entre codificadores é comumente relatada usando ÿ de Cohen, também relatamos esses valores, tanto para conjuntos de dados HBCU quanto UC, na Tabela 2. O ÿ “Média” é a média dos valores ÿ para os 4 classificadores binários. Como na Tabela 1, ambos os classificadores SVM(Gabor) e BF(Boost)

demonstrar desempenho próximo da precisão inter-humana.

No geral, achamos que os resultados encorajam que a classificação de envolvimento da máquina pode atingir níveis de precisão inter-humanos.

3.7 Generalização para um conjunto de dados diferente

Um problema bem conhecido dos classificadores de faces contemporâneos é generalizar para pessoas de uma raça diferente das pessoas no conjunto de treinamento; em particular, os detectores faciais modernos muitas vezes têm dificuldade em detectar pessoas com pele escura [56]. Para o nosso estudo, coletamos dados tanto na HBCU, onde todos os sujeitos eram afro-americanos, quanto na UC, onde todos os sujeitos eram asiático-americanos ou caucasianos-americanos. Isso nos dá a oportunidade de avaliar até que ponto um classificador treinado em um conjunto de dados generaliza para o outro. Aqui, medimos o desempenho dos classificadores binários descritos acima que foram treinados no HBCU ao classificar indivíduos da UC.

Os resultados são mostrados nas Tabelas 1 e 2 **(parte inferior)** para cada método de classificação. O classificador mais robusto foi o SVM (Gabor): o 2AFC médio caiu apenas ligeiramente de 0,729 para 0,691, e o ÿ médio caiu de 0,306 para 0,231. Curiosamente, a arquitetura MLR(CERT) não era particularmente robusta à mudança na população, apesar de ter sido treinada num número muito maior de assuntos. É possível que os recursos de pose da cabeça medidos pelo CERT e úteis para o conjunto de dados HBCU não sejam generalizados para o conjunto de dados UC. Entre as abordagens Boost(BF) e SVM(Gabor), é possível que o maior número de características BF em comparação com características Gabor tenha levado ao sobreajuste – os classificadores Boost(BF) generalizaram bem para sujeitos dentro da mesma população, mas não para sujeitos de uma população diferente.

3.8 Discriminação de estados emocionais extremos

Além dos resultados da classificação lv-repouso descritos acima, também avaliamos a precisão do classificador SVM(Gabor) na tarefa de discriminar entre um aluno que está muito engajado (ou seja, Engajamento = 4) de um aluno que é muito pouco engajado. -engajado (ou seja, Engajamento = 1).

Nesta tarefa binária, a precisão (2AFC) no conjunto de teste HBCU foi de 0,9280. No conjunto de generalização UC, foi 0,7979.

3.9 Efeito dos procedimentos de seleção de dados Conforme descrito na Seção 3.2, excluímos imagens nas quais há grande discordância de rótulos (etapa 1). É concebível que isto possa distorcer os resultados, tornando-os demasiado optimistas, porque as imagens “mais duras” podem ser aquelas sobre as quais os rotuladores tendem a discordar. Em uma análise complementar usando apenas a primeira dobra de treinamento/teste para avaliação, comparamos classificadores SVM(Gabor) em conjuntos de imagens criados sem excluir imagens com alta discordância (o que resultou em 10.409 imagens nas quais o rosto foi detectado, em vez de 7.574), para SVM(Gabor)

Matriz de confusão para classificadores binários (HBCU)

	E = 1	**E = 2**	**E = 3**	**E = 4**
1-v-outro	0,8434	0,3780	0,1014	0,1281
2-v-outro	0,4096	0,6525	0,2852	0,2277
3-v-outro	0,3849	0,4802	0,7038	0,5028
4-v-outro	0,2857	0,3658	0,4961	0,7365

Matriz de confusão para classificadores binários (UC)

	E = 1	**E = 2**	**E = 3**	**E = 4**
1-v-outro	0,7780	0,4729	0,2386	0,2643
2-v-outro	0,3603	0,6054	0,3650	0,2646
3-v-outro	0,3878	0,4303	0,5482	0,3829
4-v-outro	0,2786	0,3674	0,4448	0,7323

TABELA 3

Matrizes de confusão para os classificadores binários SVM(Gabor). Cada célula é a probabilidade de o classificador lv-other classificar uma imagem, cujo envolvimento “verdadeiro” é dado por E = l conjunto de testes , como noivado l. **Topo:** resultados para HBCU. **Abaixo:** resultados para o conjunto de generalização UC.

classificadores treinados para excluir essas imagens. Os resultados foram muito semelhantes: a precisão média (2AFC) em todos os 4 classificadores de engajamento binário foi de 0,7632 após a exclusão das imagens com alta discordância, e apenas um pouco menor em 0,7570 sem essas imagens. Isto sugere que o maior número de imagens disponíveis para treinamento pode compensar os rótulos mais ruidosos.

3.10 Matrizes de confusão

Uma questão importante ao desenvolver classificadores automatizados é que tipos de erros o sistema comete.
Por exemplo, o classificador 1-v-rest alguma vez acredita que uma imagem, cujo verdadeiro rótulo de engajamento é 4, é na verdade 1? Para responder a esta questão, devemos primeiro selecionar um limite na saída com valor real de cada classificador para que ele possa tomar uma decisão binária. Aqui, escolhemos o limite ÿ para maximizar a taxa de erro balanceada em cada dobra de teste, que definimos como a média da taxa de falsos positivos e da taxa de falsos negativos. Se a saída de valor real do classificador lv-rest em uma imagem for maior que ÿ, então o classificador decide que a imagem tem nível de engajamento l; caso contrário, ele decide que a imagem tem algum nível de engajamentodiferente de l. Usando esse procedimento de seleção de limite e calculando a média dos resultados entre as dobras, calculamos as matrizes de confusão nos conjuntos de dados HBCU e UC dos classificadores de engajamento SVM (Gabor); as matrizes são mostradas na Tabela 3. Cada célula fornece a probabilidade de que o classificador binário lv-other classifique uma imagem, cujo envolvimento “verdadeiro” (o rótulo médio arredondado sobre todos os rótulos dados a uma imagem específica) é dado por E = l como engajamento l. Observe que nem as linhas nem as colunas somam 1 – isso é natural porque os classificadores são binários, e não de 4 vias.

,

Como esperado, as diagonais da matriz dominam todos os outros valores nas linhas, o que significa que cada classificador lv-rest tem maior probabilidade de responder a uma imagem cujo verdadeiro nível de engajamento é l. No entanto, também observamos que os classificadores binários às vezes cometem “erros flagrantes”. Por exemplo, o classificador 3-v-other no

Matriz de confusão do regressor MLR (HBCU)

	D = **1**	**D** = **2**	**D** = **3**	**D** = **4**
E = **1**	0,5961	0,1735	0,1258	0,1047
E = **2**	0,2039	0,3634	0,3521	0,0807
E = **3**	0,0313	0,1511	0,4669	0,3507
E = **4**	0,0621	0,3971	0,5029	

Matriz de confusão do regressor MLR (UC)

	D = **1**	**D** = **2**	**D** = **3**	**D** = **4**
E = **1**	0,5351	0,1225	0,1963	0,1461
E = **2**	0,2588	0,2218	0,2552	0,2642
E = **3**	0,1525	0,1382	0,3950	0,3143
E = **4**	0,1254	0,0603	0,3898	.4244

Matriz de confusão de rotuladores humanos (HBCU)

	D = **1**	**D** = **2**	**D** = **3**	**D** = **4**
E = **1**	0,5943	0,2509	0,1445	0,0103
E = **2**	0,0274	0,4230	0,4910	0,0586
E = **3**	0,0029	0,1108	0,6924	0,1939
E = **0,0011**	0,0299	0,5456	0,4234	

TABELA 4

Matrizes de confusão especificando a probabilidade condicional P(D = l | E = l) da saída D do regressor de engajamento automático baseado em MLR dado o rótulo de engajamento “verdadeiro” E. **Topo:** resultados no conjunto de testes HBCU. **Meio:** resultados no conjunto de generalização UC. **Abaixo:** resultados para rotuladores humanos no conjunto de testes HBCU.

O conjunto de dados HBCU respondeu positivamente em 38,49% das imagens cujo verdadeiro nível de envolvimento era 1.

3.11 Regressão Após

realizar a classificação binária da imagem de entrada para cada nível de engajamento l ÿ {1, 2, 3, 4}, o estágio final do pipeline é combinar as saídas dos classificadores binários em uma estimativa de engajamento final. Para os classificadores binários, escolhemos a arquitetura SVM (Gabor) e usamos duas estratégias alternativas: (1) regressão linear para regressão de engajamento com valor real e (2) regressão logística multinomial (MLR) para classificação de nível de engajamento discreto de 4 vias.

3.12 Resultados

3.12.1 Regressão linear A precisão

da validação cruzada de 4 vezes, independente do sujeito, medida usando a correlação de Pearson r, foi de 0,50 no conjunto de testes HBCU. Para efeito de comparação, a precisão inter-humana na mesma tarefa foi de 0,71. No conjunto de generalização UC, a correlação média de Pearson (mais de 4 vezes) do regressor foi de 0,36.

3.12.2 Regressão logística multinomial (MLR)

Como alternativa à regressão linear, usamos regressão logística multinomial (MLR) para obter resultados de engajamento de valor discreto em {1, 2, 3, 4}. O ÿ médio de Cohen (em todas as 4 dobras) do MLR, quando comparado com rótulos humanos no conjunto de dados HBCU, foi de 0,42 (desenvolvimento padrão = 0,13); no conjunto de generalização UC, foi de 0,23 (desv. padrão = 0,13).

A Tabela 4 mostra matrizes de confusão tanto no conjunto de testes HBCU quanto no conjunto de generalização UC usando MLR como regressor. A tabela mostra a probabilidade condicional (média de 4 vezes) de que a saída D do detector seja igual a l, dado que o envolvimento “verdadeiro” E (definido como o rótulo de envolvimento médio arredondado sobre todos os rotuladores humanos que rotularam aquela imagem) é igual a l. Por exemplo, no conjunto de dados HBCU, a probabilidade de o regressor de engajamento produzir D = 1, dado que o verdadeiro engajamento foi E = 1, é 0,5961. A parte inferior da tabela mostra a matriz de confusão análoga para rotuladores humanos. No geral, a matriz de confusão do regressor MLR automatizado e a matriz para humanos são semelhantes.

Observe, no entanto, que o regressor automatizado às vezes comete erros “flagrantes”, por exemplo, classificando erroneamente imagens cujo envolvimento verdadeiro é 1 como pertencentes à categoria de envolvimento 4 (P(D = 4 | E = 1) = 0,1047 para HBCU).

Finalmente, na tarefa de discriminar E = 1 de E = 4 (semelhante à Seção 3.8), a precisão da MLR foi ÿ = 0,72; para rotuladores humanos nesta tarefa, ÿ = 0,96.

4 ENGENHARIA REVERSA DAS ETIQUETAS

Dado que o nosso objetivo neste projeto é reconhecer o envolvimento dos alunos conforme percebido por um observador externo, é instrutivo analisar como os rotuladores humanos formaram os seus julgamentos. Podemos usar os pesos atribuídos aos recursos CERT que foram aprendidos pelos classificadores MLR (CERT) para avaliar quantitativamente como os rotuladores humanos julgaram o envolvimento - se o peso MLR atribuído a AU 45 (fechamento dos olhos) tivesse uma grande magnitude, por Por exemplo, isso sugeriria que o fechamento dos olhos foi um fator importante na forma como os humanos rotularam o conjunto de dados no qual o classificador MLR foi treinado. Em particular, examinamos os pesos MLR do classificador 4-v-other MLR (CERT).

Antes do treinamento, o conjunto de dados de treinamento foi primeiro normalizado para ter variação unitária para cada recurso, de modo que todos os recursos tivessem a mesma escala. Após treinar o MLR, selecionamos os 5 pesos do MLR com maior magnitude; os resultados são mostrados na Figura 4.

A característica mais discriminante foi o valor absoluto do roll (rotação da face no plano), com o qual Engajamento = 4 foi associado negativamente (peso de -0,5659). É possível que o apoio da mão que é proeminente para Engajamento = 2 também induza o giro da cabeça, e que o classificador MLR (CERT) tenha aprendido essa tendência. A segunda ação facial mais discriminativa foi a Unidade de Ação 10 (levantar o lábio superior), que foi positivamente correlacionada com Engajamento = 4. No entanto, esta correlação pode ser potencialmente espúria, pois houve muitos momentos em que os alunos exibiram gestos de mão na boca que pode ter corrompido as estimativas da UA 10. Tais gestos foram reconhecidos como uma oclusão importante em ambientes de ensino automatizados [38].

AU 1 (elevação interna da sobrancelha), AU 45 (fechamento dos olhos) e o valor absoluto do pitch (inclinação da cabeça para cima e para baixo) também foram negativamente correlacionados com Engajamento = 4. Foi relatado anteriormente que AU 1 se correlaciona com

UA 1 AU 10 AU 10 (da HBCU)

Peso do recurso	CERT
abdômen (Rolagem) ÿ0,5659	
AU 10	+0,5089
UA 1	ÿ0,4430
AU 45 (fechamento dos olhos)	ÿ0,2851
abs (inclinação)	ÿ0,2644

Figura 4. Pesos associados às diferentes Unidades de Ação (AUs) e coordenadas de pose da cabeça para discriminar Engajamento = 4 de Engajamento = 4, junto com exemplos de AUs 1 e 10. Fotos cortesia de Carnegie Grupo de análise automática de rosto da Mellon University, http: //www.cs.cmu.edu/ÿface/facs.htm.

Correlações de envolvimento com pontuações de testes		
	Pré-teste	Pós-teste
Rotuladores humanos		
Rótulo médio de engajamento 0,52ÿ		0,37
P(Engajamento = 1) ÿ0,39 P(Engajamento = 2) ÿ0,32 P(Engajamento = 3) ÿ0,34		ÿ0,22
		ÿ0,18
P(Engajamento = 4) 0,57ÿ		ÿ0,40
Classificador automático		0,47ÿ
P(Engajamento = 4)	0,64ÿ	0,27

TABELA 5

Estatísticas de engajamento correlacionadas com qualquer pré-teste ou desempenho pós-teste. P(Engajamento = l) denota a fração de quadros de vídeo em que um assunto o nível de engajamento foi estimado em l. Correlações com a * são estatisticamente significativos (p < 0,05, bicaudal).

autorrelato de frustração dos alunos [10], mas não de envolvimento. As correlações negativas com AU 45 e pitch são intuitivos – eles sugerem que o aluno tem desligado (ou mesmo adormecido), ou está olhando para baixo longe da tela.

5 CORRELAÇÃO COM PONTUAÇÕES DE TESTE

Nesta seção investigamos a correlação entre percepções humanas e automáticas de envolvimento com desempenho e aprendizagem do teste do aluno. Mostramos resultados para Correlação de Pearson. Os resultados da correlação de postos de Spearman foram geralmente mais baixos e não são relatados.

5.1 Desempenho de teste

5.1.1 Rótulos humanos

Primeiro comparamos os julgamentos humanos de engajamento com teste o desempenho calculando o rótulo de engajamento médio em todos os quadros rotulados para cada sujeito no HBCU conjunto de dados e, em seguida, correlacionar esses engajamentos médios rótulos com pontuações de pré-teste e pós-teste (ver Tabela 5). A correlação de Pearson entre engajamento e pré-teste foi r = 0,52 (p = 0,0167, bicaudal) e entre engajamento e pós-teste foi r = 0,37 (p = 0,1027, dof = 19, bicaudal).

Também examinamos qual dos 4 níveis de engajamento foi mais preditivo do desempenho da tarefa, correlacionando a fração de quadros rotulados como Engajamento = 1, Engajamento = 2, etc., com desempenho do aluno no teste. Apenas Engajamento = 4 foi positivamente correlacionado com desempenho no pré-teste (r = 0,57, p = 0,0066, dof=19, bicaudal) e pós-teste (r = 0,47, p = 0,0324, dof=19, bicaudal). Na verdade, a fração de quadros para os quais um aluno parecia estar no nível de envolvimento 4 (que denotamos como P (Engajamento = 4)) foi um preditor melhor do que o preditor de engajamento médio descrito acima. Todos outros níveis de engajamento l < 4 foram negativos (embora não significativamente) correlacionado com o desempenho do teste, sugerindo que Engajamento = 4 é o único valor “positivo” estado de engajamento.

Para efeito de comparação, a correlação entre as notas dos alunos no pré-teste e pós-teste foi de r = 0,44 (p = 0,0471, dof = 19, bicaudal), que é ligeiramente (embora não estatisticamente significativo) menor do que a correlação entre P(Engajamento = 4) e pós-teste. Em outras palavras, as percepções humanas sobre o envolvimento dos alunos eram apenas um preditor tão bom do desempenho pós-teste quanto o pontuação do pré-teste do aluno. Uma correlação parcial entre P(Engajamento = 4) e pós-teste, dada a nota do pré-teste, deu r = 0,29 (p = 0,2073, dof = 19, bicaudal).

Por fim, vale a pena notar que outra interpretação do a correlação entre engajamento e pré-teste é que a pontuação do pré-teste de um aluno é preditiva do seu nível de envolvimento durante a sessão de aprendizagem subsequente.

5.1.2 Estimativas automáticas

Também calculamos a correlação entre julgamentos de envolvimento e pré e pós-teste do aluno desempenho. Como o melhor preditor do desempenho do teste dos julgamentos humanos foi da fração de quadros rotulado como Engajamento = 4, focamos no resultado do classificador 4-v-outro. Em particular, correlacionamos a fração de quadros em todo o vídeo de cada assunto sessão que o detector 4-v-outro previu ser um quadro “positivo” por limiarização com ÿ onde ÿ é , a saída mediana do detector em todos os quadros dos sujeitos. Em outras palavras, quadros nos quais a saída do detector excedido ÿ foi considerado um quadro “positivo” para nível de envolvimento 4. A correlação com este automático P(Engajamento = 4) preditor e desempenho pré-teste foi de 0,64 (p = 0,0023, dof=19, bicaudal); para pós-teste desempenho, foi r = 0,27 (p = 0,2436, dof=19, 2-cauda). Este é o mesmo padrão de correlações que em Seção 5.1.1 – o engajamento foi mais preditivo no pré-teste do que no pós-teste.

5.2 Aprendizagem

Além do desempenho bruto do teste, também examinamos correlações entre engajamento e aprendizagem. A diferença média entre as pontuações do pós-teste e do pré-teste

(em 21 indivíduos) foi de 2,81 séries, o que foi estatisticamente significativo ($t(20) = 4,3746$, $p = 0,0002$, bicaudal), e o que sugere que os alunos estavam aprendendo. No entanto, não encontramos correlações significativas entre envolvimento e aprendizagem, seja usando rótulos humanos ou rótulos de engajamento estimados automaticamente.

5.3 Discussão

A correlação entre engajamento e pontuações pré e pós-teste é interessante. Particularmente revelador é que o desempenho pós-teste pode ser previsto com a mesma precisão olhando para os rostos dos alunos durante a aprendizagem ($r = 0,47$) como observando suas pontuações no pré-teste (0,44). Esses resultados são consistente com [15], que encontrou uma correlação positiva entre "energia do aluno" (valência) e pré-teste de matemática pontuação, bem como uma correlação positiva entre um aluno estar "na tarefa" e pontuações de pós-teste de matemática,

A falta de correlação entre envolvimento e aprendizagem foi um tanto decepcionante, mas acreditamos que seja um uma pista importante para o planejamento de pesquisas futuras. O mais alunos engajados têm pontuações mais altas no pré-teste, o que sugere que pode haver efeitos de teto. É possível, por exemplo, que melhorar a pontuação de um teste de 10 para 11 é mais difícil do que melhorar de 1 para 2. Nós explorou esta hipótese otimizando a correlação entre envolvimento e ganhos de aprendizagem em diferentes transformações monótonas tanto de engajamento quanto de resultados dos testes. Em particular, pesquisando em todos os mapeamentos de $\{1, \ldots, 4\}$ em $\{0, \ldots, 4\}$ para engajamento e de $\{0, \ldots, 12\}$ (o intervalo de pontuações dos testes observado em nosso experimento) em $\{0, \ldots, 20\}$ para teste pontuações, identificamos uma transformação que deu correlações moderadas ($r = 0,44$, $p = 0,0458$, dof = 19, bicaudal), mas estatisticamente significativas entre aprendizagem e noivado. Esta transformação monotônica não linear estava efetivamente "desfazendo" o efeito teto, ponderando a aprendizagem ganha mais fortemente que começou com maiores linhas de base do pré-teste. No entanto, tentamos um grande número de transformações monotônicas e, portanto, a estatística importância desta análise deve ser tomada com uma grão de sal. Notamos também que a correlação entre o envolvimento e a aprendizagem podem tornar-se significativos se o o número de assuntos aumentou.

Finalmente, e mais importante, em análises laboratoriais de curto prazo estudos como o nosso, a maioria dos alunos está bastante motivada e engajado. Na verdade, ao examinar os vídeos, raramente encontrei períodos de não envolvimento prolongado. Esse é obviamente diferente das situações de sala de aula em que alguns alunos estão consistentemente engajados e alguns alunos consistentemente desligados ao longo de dias, meses e anos. O trabalho futuro beneficiaria se se concentrasse em situações de aprendizagem a longo prazo, onde a variação no envolvimento é mais provável de ser observado e o efeito do envolvimento na aprendizagem é mais provável que se torne aparente.

6. CONCLUSÕES

O aumento do envolvimento dos alunos emergiu como uma chave desafio para professores, pesquisadores e profissionais da educação instituições. Muitas das ferramentas atuais usadas para medir envolvimento - como autorrelatos, introspecção do professor avaliações e listas de verificação – são complicados, não têm a resolução temporal necessária para entender a interação entre envolvimento e aprendizagem e, em alguns casos, capturar a conformidade dos alunos em vez do envolvimento.

Neste artigo, exploramos o desenvolvimento de reconhecimento automatizado em tempo real do envolvimento a partir das expressões faciais dos alunos. A intuição motivadora foi que os professores avaliam constantemente o nível de envolvimento dos seus alunos e que as expressões faciais desempenham um papel fundamental nessas avaliações. Assim, compreender e automatizar o processo de como as pessoas julgam o envolvimento dos alunos do rosto pode ter aplicações importantes.

Nosso trabalho amplia pesquisas anteriores sobre engajamento reconhecimento usando visão computacional [42], [29], [10], [11] e é sem dúvida o estudo mais completo sobre este tópico até o momento: coletamos um conjunto de dados de expressões faciais de alunos durante a execução de uma tarefa de treinamento cognitivo. Experimentamos múltiplas abordagens para humanos observadores para avaliar o envolvimento dos alunos. Nós achamos isso a confiabilidade interobservador é maximizada quando o comprimento do os clipes observados são de aproximadamente 10 segundos. Mais curta os clipes não fornecem contexto suficiente e a confiabilidade é prejudicada. Clipes mais longos tendem a ser mais difíceis de avaliar porque muitas vezes misturam diferentes níveis de envolvimento. Quando discriminando níveis baixos versus altos de engajamento, a confiabilidade interobservador foi alta (ÿ de Cohen = 0,96). Nós também descobriram que os julgamentos de engajamento de clipes de 10 segundos poderia ser aproximado de forma confiável (Pearson $r = 0,85$) por calculando a média de julgamentos de quadro único ao longo dos 10 segundos. Isso indica que as expressões estáticas contêm a maior parte do as informações que os observadores usam para avaliar o envolvimento dos alunos. Descobrimos que os observadores confiam na postura da cabeça e ações faciais elementares, como levantar a sobrancelha, fechar os olhos e lábio superior levantado para fazer seus julgamentos.

Nossos resultados sugerem que os métodos de aprendizado de máquina poderia ser usado para desenvolver um detector de envolvimento automático em tempo real com precisão comparável à de observadores humanos. Mostramos que os julgamentos de engajamento humano e automático se correlacionam com o desempenho da tarefa. Em particular, o desempenho pós-teste dos alunos foi previsto com a mesma precisão (e estatisticamente significativa) pela observação do rosto do aluno durante a aprendizagem ($r = 0,47$) e pelas pontuações do pré-teste ($r = 0,44$). Nós não conseguiu encontrar correlações significativas entre a percepção engajamento e aprendizado. No entanto, a análise estatística a posteriori sugere que isto pode ser devido a efeitos de teto e uma limitação fundamental do laboratório de curto prazo estudos como o nosso. Nesses estudos, a maioria dos estudantes tende geralmente bastante engajado, o que é bem diferente de o envolvimento ou desligamento de longo prazo encontrado em salas de aula. Isto aponta para a importância de longo prazo estudos que se aproximam da ecologia da sala de aula em que alguns alunos estão engajados e outros estão cronicamente desligado por dias, meses e anos.

Embora o progresso alcançado aqui seja modesto, reforça a ideia de que o reconhecimento automático do envolvimento dos alunos

O desenvolvimento é possível e pode potencialmente revolucionar a educação tal como a conhecemos. Por exemplo, utilizando sistemas de visão computacional, um conjunto de câmeras de baixo custo e alta resolução poderia monitorar os níveis de envolvimento de salas de aula inteiras, sem a necessidade de autorrelatos ou questionários.
A resolução temporal da tecnologia poderia ajudar a compreender quando e por que os alunos se desinteressam e talvez a agir antes que seja tarde demais. Os professores baseados na Web poderiam obter estatísticas em tempo real sobre o nível de envolvimento dos seus alunos em todo o mundo.
Os vídeos educativos podem ser melhorados com base nos sinais agregados de envolvimento fornecidos pelos espectadores.
Tais sinais indicariam não apenas se um vídeo induz alto ou baixo envolvimento, mas, o mais importante, quais partes dos vídeos o fazem.
Nosso trabalho destaca a importância de focar em estudos de campo de longo prazo em ambientes de sala de aula da vida real. A coleta de dados nesses ambientes é fundamental para treinar sistemas de reconhecimento de engajamento mais confiáveis e ecologicamente válidos. Mais importante ainda, são necessários estudos sustentados e de longo prazo em salas de aula reais para obter uma melhor compreensão da interação entre envolvimento e aprendizagem na vida real.

APÊNDICE: PRECISÃO INTERHUMANA

Os classificadores neste artigo foram treinados e avaliados no rótulo médio, em todos os rotuladores humanos, atribuído a cada imagem. Para permitir uma comparação justa entre a precisão inter-humana e a precisão máquina-humana, avaliamos a precisão (usando ÿ de Cohen, r de Pearson ou 2AFC) de cada rotulador humano comparando seus rótulos com o rótulo médio, sobre todos os outros rotuladores, dada a cada imagem. Em seguida, calculamos a média das pontuações de precisão individuais de todos os rotuladores e relatamos isso como a confiabilidade inter-humana. Observe que esse acordo de “deixar um rotulador de fora” é normalmente maior do que o acordo médio entre pares.

REFERÊNCIAS

[1] LearningRx, 2012. www.learningrx.com.
[2] Aprendizagem Científica, 2012. www.scilearn.com.
[3] A. Anderson, S. Christenson, M. Sinclair e C. Lehr. Confira e conecte: A importância do relacionamento para promover o engajamento com a escola. Jornal de Psicologia Escolar, 42:95–113, 2004.
[4] JR Anderson. Aquisição de habilidade cognitiva. Revisão Psicológica, 89(4):369–406, 1982.
[5] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan e B. Woolf. Reparando o desengajamento com intervenções não invasivas. Anais da Conferência de 2007 sobre Inteligência Artificial na Educação, páginas 195–202, 2007.

[6] M. Bartlett, G. Littlewort, I. Fasel e J. Movellan. Detecção facial em tempo real e reconhecimento de expressões faciais: desenvolvimento e aplicações para interação humano-computador. Em Anais do Workshop CVPR sobre Interação Humano-Computador, 2003.
[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel e J. Movellan. Reconhecimento automático de ações faciais em expressões espontâneas. Revista de Multimídia, 2006.
[8]J. Beck. Rastreamento de engajamento: usando tempos de resposta para modelar o desligamento dos alunos. Em Anais da Conferência de 2005 sobre Inteligência Artificial na Educação, páginas 88–95, 2005.
[9] M. Chaouachi, P. Chalfoun, I. Jraidi e C. Frasson. Afeto e engajamento mental: rumo à adaptabilidade para sistemas inteligentes. Em FLAIRS, 2010.
[10] S. D'Mello, S. Craig e A. Graesser. Avaliação multimétodo da experiência e expressão afetiva durante a aprendizagem profunda. Jornal Internacional de Tecnologia de Aprendizagem, 4(3):165–187, 2009.
[11] S. D'Mello e A. Graesser. Detecção multimodal semiautomática de afetos a partir de sinais de conversação, linguagem corporal grosseira e características faciais. Modelagem de Usuário e Interação Adaptada ao Usuário, 20(2):147–187, 2010.
[12] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins e A. Graesser. Um momento para emoções: quando a sensibilidade ao afeto é ou não eficaz na promoção do aprendizado profundo. Nos Anais da 10ª Conferência Internacional sobre Sistemas Tutores Inteligentes, páginas 245–254, 2010.
[13] S. D'Mello, R. Picard e A. Graesser. Rumo a um AutoTutor sensível ao afeto. Sistemas Inteligentes IEEE, Edição Especial sobre Sistemas Educacionais Inteligentes, 22(4):53–61, 2007.
[14] SK D'Mello, S. Craig, J. Sullins e A. Graesser. Prever estados afetivos expressos por meio de um procedimento de emoção em voz alta a partir do diálogo de iniciativa mista do AutoTutor. Jornal Internacional de Inteligência Artificial em Educação, 16(1):3–28, 2006.
[15] T. Dragon, I. Arroyo, B. Woolf, W. Burleson, R. el Kaliouby e H. Eydgahi. Visualizar o afeto e a aprendizagem dos alunos por meio da observação da sala de aula e de sensores físicos. Nos Anais da 9ª Conferência Internacional sobre Sistemas Tutores Inteligentes, páginas 29–39, 2008.

[16] J. Dunleavy e P. Milton. O que você fez na escola hoje? explorando o conceito de envolvimento dos alunos e suas implicações para o ensino e a aprendizagem no Canadá. Associação Canadense de Educação (CEA), páginas 1–22, 2009.

[17] P. Ekman e W. Friesen. O sistema de codificação de ação facial: uma técnica para medir o movimento facial. Consulting Psychologists Press, Inc., São Francisco, CA, 1978.
[18] S. Fairclough e L. Venables. Predição de estados subjetivos a partir da psicofisiologia: uma abordagem multivariada. Psicologia biológica, 71:100–110, 2006.
[19] K. Forbes-Riley e D. Litman. Adaptação a múltiplos estados afetivos no diálogo falado. Em Anais do Grupo de Interesse Especial sobre Discurso e Diálogo, páginas 217–226, 2012.
[20] JA Fredricks, PC Blumenfeld e AH Paris. Engajamento escolar: Potencial do conceito, estado das evidências. Revisão de Pesquisa Educacional, 74(1):59–109, 2004.
[21] Y. Freund e RE Schapire. Uma generalização teórica da decisão da aprendizagem on-line e uma aplicação para impulsionar. Na Conferência Europeia sobre Teoria da Aprendizagem Computacional, páginas 23–37, 1995.
[22] J. Friedman, T. Hastie e R. Tibshirani. Regressão logística aditiva: uma visão estatística do boosting. Anais de Estatística, 28(2):337–407, 2000.

[23] B. Goldberg, R. Sottilare, K. Brawner e H. Holden. Prever o envolvimento do aluno durante interações interculturais bem definidas e mal definidas baseadas em computador. Nos Anais da 4ª Conferência Internacional sobre Computação Afetiva e Interação Inteligente, páginas 538–547, 2011.

[24] J. Grafsgaard, R. Fulton, K. Boyer, E. Wiebe e J. Lester. Análise multimodal do canal afetivo implícito na comunicação textual mediada por computador. Em Anais da Conferência Internacional sobre Interação Multimodal, páginas 145–152, 2012.
[25] L. Harris. Uma investigação fenomenográfica das concepções dos professores sobre o envolvimento dos alunos na aprendizagem. O Pesquisador Educacional Australiano, 5(1):57–79, 2008.
[26] J. Johns e B. Woolf. Um modelo de mistura dinâmica para detectar a motivação e a proficiência dos alunos. Dentro Anais da Vigésima Primeira Conferência Nacional sobre Inteligência Artificial (AAAI06), páginas 2–8, 2006.

[27] T. Kanade, J. Cohn e Y.-L. Tian. Banco de dados abrangente para análise de expressões faciais. Em Anais da 4ª Conferência Internacional IEEE sobre Reconhecimento Automático de Rosto e Gestos (FG'00), páginas 46 – 53, março de 2000.
[28] A. Kapoor, S. Mota e R. Picard. Rumo a um companheiro de aprendizagem que reconhece o afeto. No Simpósio de Outono da AAAI, 2001.
[29] A. Kapoor e R. Picard. Reconhecimento de efeitos multimodais em ambientes de aprendizagem. Em Anais da 13ª conferência internacional anual da ACM sobre Multimídia, páginas 677–682, 2005.
[30] KR Koedinger e JR Anderson. Aulas inteligentes vão para a escola na cidade grande. Jornal Internacional de Inteligência Artificial em Educação, 8:30–43, 1997.
[31] M. Lades, JC Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, RP Wurtz e W. Konen. Reconhecimento de objetos invariantes à distorção na arquitetura de link dinâmico. Transações IEEE em Computadores, 42:300–311, 1993.
[32] R. Larson e M. Richards. Tédio nos anos do ensino médio: culpar as escolas versus culpar os alunos. Jornal americano de educação, 99:418–443, 1991.

[33] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind e J. Movellan.

Dinâmica de expressão facial extraída automaticamente do vídeo.
Computação de Imagem e Visão, 24(6):615–625, 2006.
[34] G. Littlewort, M. Bartlett e K. Lee. Codificação automática de expressões faciais exibidas durante dor simulada e genuína. Computação de Imagem e Visão, 27(12):1797–1803, 2009.
[35] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan e M. Bartlett. Caixa de ferramentas de reconhecimento de expressão de computador. Dentro Anais da Conferência Internacional sobre Reconhecimento Automático de Rosto e Gestos (FG'11), páginas 298–305, 2011.
[36] R. Livingstone. O futuro na educação. Imprensa da Universidade de Cambridge, 1941.

[37] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar e I. Matthews. O conjunto de dados Cohn-Kanade estendido: um conjunto de dados completo para unidade de ação e expressão específica de emoção. No Workshop CVPR sobre Comportamento Humano-Comunicativo, páginas 94–101, 2010.
[38] M. Mahmoud e P. Robinson. Interpretando gestos manuais. Em Anais da Conferência Internacional sobre Computação Afetiva e Interação Inteligente, páginas 248–255, 2011.
[39] S. Makeig, M. Westerfield, J. Townsend, T.-P. Jung, E. Courchesne e TJ Sejnowski. Componentes funcionalmente independentes de potenciais relacionados a eventos iniciais em uma tarefa de atenção visual espacial.
Transações Filosóficas da Royal Society: Biological Science, 354:1135–44, 1999.

[40] S. Mason e A. Weigel. Uma estrutura genérica de verificação de previsões para fins administrativos. Revisão Mensal do Tempo, 137:331–349, 2009.

[41] G. Matthews, S. Campbell, S. Falconer, L. Joyner, J. Huggins e K. Gilliland. Dimensões fundamentais do estado subjetivo em ambientes de desempenho: envolvimento na tarefa, angústia e preocupação.
Emoção, 2(4):315–340, 2002.
[42] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp e A. Graesser. Características faciais para detecção de estados afetivos em ambientes de aprendizagem. Em Proceedings of the 29th Annual Cognitive Science Society, páginas 467–472, 2007.

[43] J. Mostow, A. Hauptmann, L. Chase e S. Roth. Rumo a um treinador de leitura que escuta: Detecção automatizada de erros de leitura oral. Em Anais da Décima Primeira Conferência Nacional sobre Inteligência Artificial (AAAI93), páginas 392–397, 1993.
[44] JR Movellan. Tutorial sobre filtros Gabor. Relatório técnico, Tutoriais MPLab, UCSD MPLab, 2005.
[45] H. O'Brien e E. Toms. O desenvolvimento e avaliação de uma pesquisa para medir o envolvimento do usuário. Jornal da Sociedade Americana de Ciência e Tecnologia da Informação, 61(1):50–69, 2010.
[46] J. Ocumpaugh, RS Baker e MMT Rodrigo. Manual de treinamento do protocolo do método de observação Baker-Rodrigo 1.0. Relatório técnico, EdLab, Manila, Filipinas, 2012.
[47] M. Pantic e I. Patras. Dinâmica da expressão facial: Reconhecimento de ações faciais e seus segmentos temporais a partir de sequências de imagens de perfil facial. Sistemas, Homem e Cibernética – Parte B: Cibernética, 36(2), 2006.

[48] J. Parsons e L. Taylor. Envolvimento dos alunos: o que sabemos e o que devemos fazer. Relatório técnico, Universidade de Alberta, 2011.

[49] A. Pope, E. Bogart e D. Bartolome. Sistema biocibernético avalia índices de envolvimento do operador em tarefas automatizadas.
Psicologia Biológica, 40:187–195, 1995.
[50] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati e RS
Padeiro. Métodos de elicitação de conhecimento para modelagem de afetos na educação. Revista Internacional sobre Inteligência Artificial na Educação, 2013.

[51] D. Shernof, M. Csikszentmihalyi, B. Schneider e E. Shernoff.
O envolvimento dos alunos nas salas de aula do ensino médio sob a perspectiva da teoria do fluxo. Psicologia Escolar Trimestral, 18(2):158–176, 2003.

[52] K. VanLehn, C. Lynch, K. Schultz, J. Shapiro, R. Shelby e L. Taylor. O sistema de ensino de física dos Andes: Lições aprendidas. Jornal Internacional de Inteligência Artificial e Educação, 15(3):147–204, 2005.

[53] P. Viola e M. Jones. Detecção robusta de objetos em tempo real. International Journal of Computer Vision, 57(2):137–154, 2004.
[54] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett e J. Movellan. Detecção de motorista sonolento por meio de análise de movimentos faciais. Dentro Anais da Conferência Internacional IEEE sobre Interação Humano-Computador, páginas 6–18, 2007.
[55] J. Whitehill, M. Bartlett e JR Movellan. Reconhecimento automático de expressões faciais para sistemas tutores inteligentes. Em Anais do Workshop CVPR 2008 sobre Análise do Comportamento Comunicativo Humano, páginas 1–6, 2008.

[56] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett e J. Movellan.
Rumo à detecção prática de sorriso. Transações em Análise de Padrões e Inteligência de Máquina, 31(11):2106–2111, 2009.
[57] J. Whitehill e J. Movellan. Uma abordagem discriminativa para o rastreamento da pose da cabeça quadro a quadro. Nos Anais da Conferência Internacional IEEE sobre Reconhecimento Automático de Rosto e Gestos, 2008.

[58] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett e J. Movellan. Rumo a um sistema instrucional de habilidades cognitivas sensível ao afeto ideal. Em Workshop de Visão Computacional e Reconhecimento de Padrões sobre Comportamento Comunicativo Humano, páginas 20–25,
2011.
[59] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper e R. Picard. Tutores conscientes do afeto: reconhecendo e respondendo ao afeto do aluno. Jornal Internacional de Tecnologia de Aprendizagem, 4(3):129–164, 2009.

[60] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett e J. Movel-lan. Arquiteturas multicamadas para reconhecimento de unidades de ação facial.
Transações IEEE sobre Sistemas, Homem e Cibernética B: Cibernética, 42(4):1027–1038, 2012.
[61] G. Zhao e M. Pietikainen. Reconhecimento dinâmico de texturas usando padrões binários locais com aplicação a expressões faciais.
Transações em Análise de Padrões e Inteligência de Máquina, 29(6):915–928, 2007.

Jacob Whitehill é pesquisador em aprendizado de máquina, visão computacional e suas aplicações na educação. Ele possui doutorado pela Universidade da Califórnia, San Diego, mestrado pela Universidade de Western Cape e bacharelado por Stanford. Desde 2012 ele é cofundador e cientista pesquisador da Emotient.

Zewelanji Serpell é professor associado do Departamento de Psicologia da Virginia Commonwealth University. A sua investigação centra-se no desenvolvimento e avaliação de intervenções para melhorar o funcionamento executivo dos alunos e otimizar a aprendizagem em contextos escolares. Ela possui bacharelado em psicologia pela Clark University em Worcester MA e doutorado. em Psicologia do Desenvolvimento pela Howard University em Washington, DC.

Yi-Ching Lin recebeu uma bolsa de Modelagem e Simulação na Old Dominion University, onde está cursando seu doutorado em Estudos Ocupacionais e Técnicos no Departamento de Educação e Estudos Profissionais STEM. Sua pesquisa atual centra-se no uso de modelagem dinâmica de sistemas para avaliar a influência de vários fatores na escolha dos alunos por cursos STEM. Ela possui mestrado em psicologia pela Virginia State, bacharelado em psicologia pela University of Missouri Columbia e bacharelado em engenharia química pela National Taipei University of Technology.

Aysha Foster é uma pesquisadora de ciências sociais cujos interesses incluem estratégias eficazes de aprendizagem e saúde mental para jovens de minorias. Ela recebeu seu mestrado e doutorado em psicologia pela Virginia State University e bacharelado em Biologia pela Prairie View A&M University. Atualmente ela é coordenadora de pesquisa na Virginia Commonwealth University em um projeto que examina a maleabilidade do controle executivo em alunos do ensino fundamental.

Javier Movellan fundou o Laboratório de Percepção de Máquinas na UCSD, onde é professor pesquisador. Ele também é fundador e pesquisador principal da Emotient. Os interesses de pesquisa de Javier incluem aprendizado de máquina, percepção de máquina, análise automática do comportamento humano e robôs sociais. Sua equipe ajudou a desenvolver o primeiro sistema comercial de detecção de sorriso, incorporado em câmeras digitais de consumo. Ele também foi pioneiro no desenvolvimento de robôs sociais e seu uso na educação infantil. Antes de ocupar seu cargo na UCSD, ele foi bolsista Fulbright na UC Berkeley, onde recebeu seu doutorado.