MEDIÇÃO DO NÍVEL DE ENGAJAMENTO DO ALUNO USANDO INFORMAÇÕES FACIAIS

Islam Alkabbany Asem Ali Amal Farag † Ian Bennett † Mohamad Ghanoum Aly Farag

CVIP Lab Universidade de Louisville, EUA
† TSN, Inc., Palo Alto, CA, EUA

ABSTRATO

Neste artigo, propomos uma nova estrutura que mede o nível de envolvimento dos alunos em um ambiente de aula ou em um ambiente de e-learning. A estrutura proposta captura o vídeo do usuário e rastreia seus rostos através de os frames do vídeo. Diferentes características são extraídas do rosto do usuário, por exemplo, pontos de referência faciais, pose da cabeça, olhar fixo, recursos aprendidos, etc. Esses recursos são então usados para detectar o Sistema de Codificação de Ação Facial (FACS), que decompõe expressões faciais em termos de ações fundamentais de músculos individuais ou grupos de músculos (ou seja, unidades de ação). As unidades de ação decodificadas (UA's) são então usadas para medir a disposição do aluno em participar do processo de aprendizagem (ou seja, envolvimento comportamental) e sua atitude emocional em relação à aprendizagem (ou seja, envolvimento emocional). Esse estrutura permitirá que o palestrante receba um feed- de volta das características faciais, olhar e outras cinéticas corporais. O A estrutura é robusta e pode ser utilizada em inúmeras aplicações, incluindo, mas não se limitando, ao monitoramento do progresso de alunos com vários graus de dificuldades de aprendizagem, e a análise da paralisia nervosa e seus efeitos na expressão facial e interações sociais.

Termos de Indexação— Medição do nível de engajamento, Aprendizado de Máquina, Dificuldades de Aprendizagem.

1. INTRODUÇÃO

O vídeo é uma modalidade de imagem importante e útil por si só (por exemplo, na extração de sinais vitais, análise da pele, a mandíbula humana, etc.), ou em conexão com outras modalidades (por exemplo, intervenções guiadas por imagem). Este artigo trata usando imagens de vídeo para modelar o rosto e a cabeça humanos movimento para descrever o envolvimento em ambientes educacionais. O rosto é uma ferramenta importante para a comunicação social não-verbal. comunicação. Assim, a análise do movimento facial é um tópico de pesquisa ativo para cientistas comportamentais, uma vez que o trabalho de Darwin em 1872. Estudos mostraram que as dificuldades sociais, emocionais e comportamentais têm sido consideradas secundárias em relação ao problemas cognitivos da dificuldade de aprendizagem e têm sido atribuída como razão da frustração do repetido fracasso escolar comum à maioria das crianças com dificuldades de aprendizagem [1].

O vídeo é uma modalidade importante para análise de informações faciais, paralisia nervosa, monitoramento de sinais vitais, movimento corporal, etc. Em particular, os modelos de informação facial de vídeo têm tem sido utilizado com sucesso em diversas aplicações médicas, como como estimativa da frequência cardíaca para análise de depressão [2], automático avaliação do grau de paralisia do nervo facial [3], Humano reconhecimento de comportamento [4], emoções quantificando olhares durante interação social [5] etc.

O principal objetivo deste estudo é desenvolver modelos robustos de informações faciais para descrever o envolvimento humano dentro um ambiente educacional. Em particular, no nível emocional nível, onde as expressões correspondem aos movimentos musculares referente à atenção, e no nível comportamental, onde movimentos bruscos da cabeça e piscar dos olhos são usados como indicadores de atenção. Embora nosso estudo tenha sido realizado em um grupo de controle de alunos, este trabalho tem um amplo impacto. Esta pesquisa pode ser estendida ainda mais a estudantes com especial precisa. Ao detectar o desligamento, pesquisas futuras poderão utilizar esta ferramenta para desenvolver um sistema de alerta precoce para detectar alunos ansiedade e depressão.

Apesar dos avanços no reconhecimento automático das emoções humanas, houve apenas um pequeno número de estudos de características faciais. expressões relacionadas ao conhecimento cognitivo-afetivo centrado na aprendizagem estados. A metodologia de visão computacional pode estimar discretamente o envolvimento de um aluno a partir de dicas faciais, por exemplo, [6, 7, 8] Tais estudos aplicam um ou mais dos seguintes paradigmas. Observação e anotação de comportamentos afetivos, investigação de unidades de ação facial envolvidas em atividades centradas na aprendizagem. efeito; e aplicação de métodos automatizados para detecção de estados afetivos. Kapoor e Picard [6] usaram uma câmera equipada com LEDs IR para rastrear os alunos. Além disso, uma cadeira de detecção é usada para extrair informações sobre as posturas. Além disso, eles registrou a atividade que o sujeito está realizando no computador. Então uma mistura de processos gaussianos combina todos os informações e prevê o estado afetivo atual. Em seu estudo, 8 crianças (8 - 11 anos) estão matriculadas. As crianças eram solicitado a resolver quebra-cabeças em um computador. Durante 20 minutos, o a atividade da tela, a visão lateral e a visão frontal foram registradas. Dos vídeos coletados são extraídos 136 clipes Professores foram solicitados a observar e registrar o estado afetivo às 8 horas pessoas por segundo. Os estados afetivos em consideração são interesse alto, médio e baixo, entediado e “fazendo uma pausa”. As taxas de reconhecimento de uma classe SVM de juros versus desinteresse

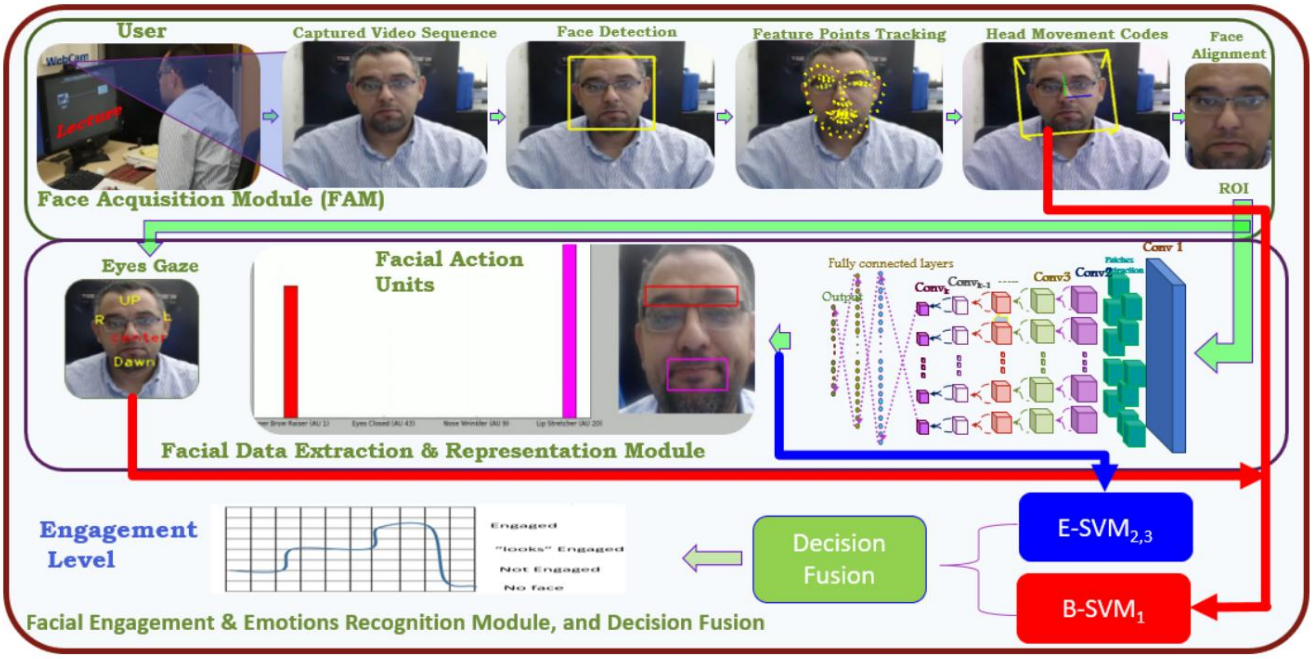

Figura 1. A estrutura proposta

significativa (para 65 amostras de interesse e 71 amostras desinteressadas 71) são 69,84% (usando informações da face superior) e são 57,06% (usando informações da face inferior). Eles alcançaram uma taxa de reconhecimento de 86,55% combinando todas as informações, não apenas as características faciais, usando uma mistura de processos gaussianos.

Para detectar as emoções que acompanham a aprendizagem de nível profundo, McDaniel et al. [7] investigaram características faciais. Os estados afetivos em consideração são tédio, confusão, deleite, fluxo, frustração e surpresa. Para realizar o estudo, 28 alunos de graduação foram convidados a interagir com o Auto-Tutor. Em seguida, a anotação dos estados afetivos foi feita pelo aprendiz, um colega e dois juízes treinados. A verdade dos dados é obtida dos juízes treinados, com confiabilidade interjuízes Kappa de Cohen (0,49). Em seguida, os dados foram amostrados em 212 videoclipes emocionais (3-4 segundos) com estados afetivos: tédio, confusão, deleite, frustração e neutro. Finalmente, dois codificadores treinados codificaram as expressões faciais dos participantes usando o Sistema de Codificação de Ação Facial de Ekman.

Em seguida, foram realizadas correlações computadas para determinar até que ponto cada uma das unidades de ação (UA's) era diagnóstica dos estados afetivos de tédio, confusão, deleite, frustração e neutralidade. Suas análises indicaram que UAs específicas poderiam classificar a confusão, o deleite e a frustração como neutros, mas o tédio era indistinguível do neutro.

Recentemente, Whitehill et al. [8] introduziram uma abordagem para reconhecimento automático do envolvimento a partir das expressões faciais dos alunos. Eles alegaram que os observadores humanos concordam de forma confiável ao discriminar 4 níveis distintos de envolvimento (k de Cohen = 0,56). Eles coletaram um conjunto de dados de 34 estudantes de pós-graduação, que treinaram usando software de treinamento de habilidades cognitivas. O rosto do participante foi gravado durante o treinamento. Para anotar os dados, 7 rotuladores codificaram os quadros de vídeo usando uma escala para avaliar o engajamento: 1: Não engajado, 2: Nominalmente engajado, 3: Envolvido na tarefa, 4: Muito engajado e X: quadro pouco claro. Em seguida, 24.285 quadros foram selecionados de forma que a diferença entre quaisquer dois rotuladores não exceda um e nenhum rotulador atribuiu X ao quadro.
O rótulo de “verdade básica” de um quadro é a média inteira de todos os rótulos. As características Gabor foram extraídas da face detectada para gerar um vetor de características 40x48x48. Em seguida, quatro classificadores SVM binários foram usados para detectar um nível fora dos quatro níveis de envolvimento. Finalmente, um regressor logístico multinomial foi utilizado para combinar a saída dos quatro classificadores binários.

A principal contribuição deste artigo pode ser resumida da seguinte forma:

- Introduzir um conjunto de dados de engajamento com mais de 73.500 imagens faciais anotadas de 14 indivíduos.
- Propor uma estrutura para medir o nível de envolvimento usando as informações faciais do stream de vídeo usando apenas webcams disponíveis na prateleira.

2. O QUADRO PROPOSTO

O sistema de medição de nível do sistema de medição de nível de engajamento baseado em facial proposto é mostrado na Fig. 1. O sistema consiste em quatro módulos principais; Módulo de aquisição facial, Módulo de extração e representação de dados faciais,

Módulo de reconhecimento de nível de engajamento e fusão de decisão
Módulo.

O módulo de aquisição facial extrai uma região facial alinhada (ROI) de cada quadro de uma determinada sequência de vídeo, aplicando um conjunto de algoritmos em cascata. Primeiro, um rosto é detectado usando o algoritmo de detecção de rosto, que é baseado no detector de rosto Viola-Jones [9]. Para reduzir as faces falso-positivas, um
detector de pele é usado para medir a proporção de pele no detectado rosto do candidato.

Depois que um rosto é detectado, 49 pontos de características faciais são extraídos usando nossa abordagem [10]. Para rastrear o ROI por meio os quadros, os primeiros dois segmentos de linha horizontalmente paralelos são alinhados nas pontas das sobrancelhas e no centro da mariposa.
Em seguida, um algoritmo de fluxo óptico baseado em Lucas-Kanade é usado para rastreie esses segmentos de linha através de quadros, então esses segmentos são usados para estimar o retângulo da face.

A estimativa da pose do rosto é um problema de perspectiva de n pontos (PnP). No método PnP, dado um conjunto de n pares entre Pontos 3D e suas projeções 2D, bem como parâmetros intrínsecos da câmera, a pose é estimada. Portanto, dados 49 3D
pontos faciais esparsos de [11] e os pontos de recurso 2D correspondentes extraídos no módulo anterior, a rotação e
a translação da cabeça pode ser estimada.

O módulo de extração e representação de dados faciais extrai o estado emocional interno de um aluno usando Ekman e
Friesen Facial Action Coding System (FACS)[12], que codifica o movimento dos músculos faciais em unidades de ação
(AU). Usamos nossa nova abordagem baseada em CNN [13] que treinado no conjunto de dados BP4D [14]

O segundo Módulo (Módulo de Extração e Representação de Dados Faciais) extrai as unidades de ação facial (FACS) [12]
usando uma nova abordagem baseada em CNN [13]. usamos o modelo proposto em [13], que treinou no conjunto de dados BP4D [14],
como modelo pré-treinado para nosso processo de treinamento. Em seguida, ajustamos este modelo para detectar 33 unidades de ação facial usando mais conjuntos de dados, por exemplo, conjuntos de dados CK+ [15], MPI [16]. Este módulo também estimar os códigos de movimento ocular. Primeiro, a região do olho é extraída do ROI e, em seguida, um banco de 40 filtros Gabor (8 orientações, 5 frequências espaciais) é aplicado para obter uma característica
vetor que costumava treinar o classificador SVM.

Por fim, o nível de engajamento do aluno é estimado por meio do último módulo. Dadas as características faciais de alto nível (por exemplo,
FACS, códigos de movimento da cabeça e códigos de movimento dos olhos), o módulo de reconhecimento de nível de engajamento é usado para estimar o nível de envolvimento como um dos quatro níveis: nível 0 (não há
rosto detectado), nível 1, o aluno não está engajado comportamentalmente ou seja, olhar para fora da tela), nível 2 (o aluno parece engajado
ou seja, olhar dentro da tela, mas sem nenhum sinal de engajamento),
e nível 3 (o aluno está emocionalmente engajado). Esses níveis
2. Este módulo é composto por dois classificadores SVM comportamental e SVM emocional.

é processado usando os módulos anteriores, se um rosto for detectado,
ângulos de pose e olhar são alimentados no classificador Behavioral-SVM para verificar o envolvimento comportamental; caso contrário, é um
quadro ruim. Se o classificador identificar o sujeito como comportamentalmente engajado, os 33 FACS serão alimentados no SVM Emocional.
classificador para identificar o nível de envolvimento emocional.

Figura 2. Diferentes tipos de categorias de engajamento

3. EXPERIMENTOS

Neste estudo, investigamos como os observadores humanos julgam o envolvimento a partir do rosto. Identificar essas observações ajuda
na automatização do processo usando aprendizado de máquina. Alcançar esse objetivo, coletamos um conjunto de dados de engajamento que tem mais de 73.500 imagens faciais anotadas para 14 indivíduos.

3.1. Coleção de dados

Um aplicativo baseado na web foi implementado para capturar vídeo facial dos sujeitos durante a visualização de uma palestra de 10 a 15 minutos seguida de um questionário. Cada sujeito usou seu próprio dispositivo que
está equipado com uma webcam e um navegador web. Então, facial vídeos foram gravados durante a palestra. O conjunto de dados coletado é composto por 14 alunos de graduação e pós-graduação, que são aprendidos dois cursos de engenharia diferentes.

3.2. Anotação de conjunto de dados

Para anotar o conjunto de dados, 4 rotuladores codificaram os quadros de vídeo usando 4 níveis de engajamento. Depois de excluir a categoria “0”, coletamos 109.325 quadros. Em seguida, aplicamos o seguinte processo de votação: um quadro é incluído no conjunto de dados se pelo menos 3 anotadores usaram a mesma etiqueta para o quadro. Portanto,

Tabela 1. As taxas de reconhecimento do Behavioral-SVM e os classificadores Emotional-SVM para processo de validação cruzada.

	Comportamental-SVM	Emocional-SVM
A	66%	74%
Precisão	80%	70%

o número total de frames selecionados é de 73.530 com a seguinte distribuição: 8.783 frames na categoria “1” (11,9%), 41.440 quadros na categoria “2” (56,4%) e 23.307 quadros na categoria categoria “3” (31,7%). O acordo entre codificadores emparelhados para este conjunto é calculado usando correlações de Pearson. Os seis os valores pareados são 0,66, 0,71, 0,62, 0,72, 0,58 e 0,61.

3.3. Classificação automática de engajamento

Conforme mostrado na seção anterior, este conjunto de dados está desequilibrado, Usar esse conjunto de dados desequilibrado no treinamento pode levar a resultados tendenciosos. modelos. Para resolver esse problema, optamos por reduzir a amostragem do conjunto de dados, que envolve a remoção de alguns dos principais dados das classes (no nosso caso, classes 2 e 3). Esta redução da amostragem a operação depende da manutenção de um quadro de cada 10 quadros sucessivos para cada categoria.

O conjunto de dados com amostragem reduzida é usado para treinar dois SVM classificadores para os compromissos comportamentais e emocionais. O O classificador comportamental-SVM é treinado usando os ângulos de pose e dados do olhar para identificar se um quadro pertence à categoria 1 ou não. Se o quadro estiver comportamentalmente engajado, ele é usado para treinar o classificador Emotional-SVM usando os AU's para identificar se o quadro pertencer à categoria 2 ou 3. Usamos as intensidades estimadas de AU (ou seja, a distância da amostra do recurso ao hiperplano de separação do SVM para aquela UA) para treinar o Classificador Emocional-SVM.

O experimento de validação cruzada leave-one-out (LOOCV) é conduzido usando 14 conjuntos de treinamento e teste. Em seguida, são obtidos os resultados do reconhecimento. O desempenho é medido usando a área sob as Características Operacionais do Receptor (ROC) estatísticas da curva A e a precisão mostrada na Tabela. 1, isso experimento mede o desempenho do sistema ponta a ponta. A estrutura proposta extrai um vetor de características 41-D de cada quadro. Esses recursos incluem código de 33 UAs, 3 ângulos de pose e 5 códigos de olhar. Este vetor de recursos 41-D é extraído em menos de 200 ms. Isto faz com que o sistema seja aplicável para processamento on-line (ou seja, 5 fps).

Realizamos outro experimento para melhorar o desempenho do engajamento emocional off-line (vídeo gravado) processamento em detrimento do tempo de execução, usando um alto vetor de recurso de dimensão Redimensionamos o ROI para 32x32 e aplique 40 filtros Gabor (4 orientações, 5 frequências espaciais e 2 escalas) para obter um vetor de recursos 40960-D. Este vetor é usado para treinar o classificador Emotional-SVM usando tusing LOOCV. O desempenho é medido usando a área sob Estatísticas da curva ROC A e precisão conforme mostrado na Tabela. 2

Figura 3. Resultados da amostra.

Tabela 2. Taxas de reconhecimento dos classificadores Emotional-SVM para processo de validação cruzada treinados com diferentes recursos atividades .

Recurso	Unidade de ação facial	Filtros Banco de Gabor
A	74%	82%
Precisão	70%	80%
Dimensões	33	40960

A Figura 3 mostra uma amostra das curvas geradas pela nossa plataforma introduzida.

4. CONCLUSÃO

Este artigo propôs um conjunto de dados de engajamento com mais de 73.500 imagens faciais anotadas anotadas para três níveis diferentes de envolvimento. Usando um processo de votação que envolveu pelo menos pelo menos três anotadores humanos, uma taxa de concordância razoável foi alcançado. Também foi introduzida uma nova técnica para medir o nível de engajamento, esta técnica utiliza apenas o fluxo de vídeo fornecido por webcams disponíveis para extrair diferentes características faciais que são usadas para classificar automaticamente o envolvimento comportamental e emocional do usuário com quase desempenho em tempo real.

5. REFERÊNCIAS

[1] Lilly Dimitrovsky, Hedva Spector, Rachel Levy-Shiff, e Eli Vakil, “Interpretação de expressões faciais de afeta crianças com dificuldades de aprendizagem com dificuldades verbais ou déficits não-verbais”, Journal of Learning Disabilities, vol. 31, não. 3, pp. 286–292, 1998.

[2] Aamir Mustafa, Shalini Bhatia, Munawar Hayat e Roland Goecke, “Estimativa da frequência cardíaca por tratamento facial vídeos para análise de depressão”, em 2017, Sétima Conferência Internacional sobre Computação Afetiva e Interação Inteligente (ACII). IEEE, 2017, pp.

[3] Ting Wang, Shu Zhang, Junyu Dong, Lian Liu e Hui Yu, “Avaliação automática do grau do nervo facial paralisia”, Ferramentas e Aplicativos Multimídia, vol. 75, não. 19, pp.

[4] Remco Veltkamp e Nico van der Aa, “simpósios de reconhecimento do comportamento humano a partir de vídeo”, 9ª conferência internacional sobre métodos e técnicas em pesquisa comportamental, 2014.

[5] Shengyao Guo, Eric Ho, Yalun Zheng, Qiming Chen, Vivian Meng, John Cao, Si Wu, Leanne Chukoskie e Pamela Cosman, “Usando detecção de rosto e objeto para quantificar olhares durante interações sociais”, em Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Simpósio de Imagem Biomédica. IEEE, 2018, pp. 1144–1148.

[6] Ashish Kapoor e Rosalind W. Picard, “Multimodal afetam o reconhecimento em ambientes de aprendizagem”, em Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005, pp.

[7] B. King P. Chipman K. Tapp B. McDaniel, S. DMello e A. Graesser, “. características faciais para estado afetivo detecção em ambientes de aprendizagem”, em Proceedings da 29ª Sociedade Anual de Ciência Cognitiva, 2007, p. 467472.

[8] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster e Javier R. Movellan, “Os rostos de envolvimento: Reconhecimento automático do envolvimento do aluno a partir de expressões faciais”, IEEE Trans. Afetivo Computação, vol. 5, não. 1, pp. 86–98, 2014.

[9] Paul Viola e Michael J. Jones, “Robusto rosto em tempo real detecção”, Int. J. Computação. Visão, vol. 57, não. 2 de maio 2004.

[10] Eslam Mostafa, Asem A. Ali, Ahmed Shalaby e Aly Farag, “Um detector de características faciais que integra holística informação facial e modelo baseado em peças”, em CVPR-Workshops, 2015.

[11] Tal Hassner, Shai Harel, Eran Paz e Roee Enbar, “Frontalização facial eficaz em imagens irrestritas” Conferência IEEE 2015 sobre Visão Computacional e Padrões Reconhecimento (CVPR), 2015.

[12] P. Ekman, WV Friesen e JC Hager, Ação Facial Sistema de Codificação (FACS): Manual, A Human Face, 2002.

[13] Asem M. Ali, Islam Alkabbany, Amal Farag, Ian Ben-nett e Aly Farag, “Detecção de unidades de ação facial sob propor variações usando aprendizagem de regiões profundas”, na Sétima Conferência Internacional sobre Computação Afetiva e Interação Inteligente (ACII), 2017.

[14] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Cana-van, Michael Reale, Andy Horowitz, Peng Liu e Jeffrey M. Girard, “BP4D-espontâneo: uma alta resolução banco de dados de expressões faciais dinâmicas em 3D espontâneas.” Imagem Vision Comput., vol. 32, não. 10, pp. 2014.

[15] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar e Iain Matthews, “O conjunto de dados cohn-kanade estendido (CK +): um conjunto de dados completo para unidade de ação e expressão específica de emoção”, em CVPR-Workshops, 2010.

[16] Kathrin Kaulard, Douglas W. Cunningham, Heinrich H. Blthoff e Christian Wallraven, “O banco de dados de expressão facial mpi é um banco de dados validado de emoções e expressões faciais de conversação”, PLoS ONE, vol. 7, não. 3, 2012.