Avaliação do Desempenho de Abordagens Computacionais para Inferência Automática de Tons de Pele

Igor Joaquim da S. Costa[1], Thais F. Silva[1], Marisa Vasconcelos[1], Julio C. S. Reis[2], Jussara M. Almeida[1], Virgílio Almeida[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Brasil
[2]Universidade Federal de Viçosa (UFV) – Brasil

{igor.joaquim, thaisferreira,jussara,virgilio}@dcc.ufmg.br
marisa.vasconcelos@gmail.com, jreis@ufv.br

Abstract. *Automatic estimates of skin tone are challenging due to racial and gender biases in machine learning approaches. In this work, we explore a labeled dataset of images to evaluate two computational approaches widely explored in the literature, ITA and CASCo, in order to investigate their robustness in performing this task. The results reveal that these approaches still have weaknesses that should be considered before their application in sensitive contexts.*

Resumo. *Estimativas automáticas do tom de pele enfrentam desafios devido a vieses raciais e de gênero em abordagens de aprendizado de máquina. Neste trabalho, exploramos um conjunto de dados rotulado e avaliamos duas abordagens computacionais amplamente exploradas na literatura, ITA e CASCo, a fim de investigar a robustez e limitações nessa tarefa. Nossos resultados mostram que essas abordagens ainda apresentam falhas significativas, comprometendo sua aplicação em contextos reais onde a precisão é essencial.*

1. Introdução

Estimar o tom de pele é um problema multidisciplinar que abrange áreas como a medicina, as ciências sociais e a computação. Na medicina, especialmente na dermatologia, a estimativa do tom de pele é crucial para prever, por exemplo, a sensibilidade à luz ultravioleta (UV), fator importante para identificar o risco de doenças de pele como o câncer de pele [Khosla et al. 2021]. Já nas ciências sociais, o tom de pele afeta interações e processos, refletindo desigualdades sociais e raciais presentes em diversas sociedades [Schumann et al. 2023]. Capturá-lo com precisão melhora a identificação de grupos sub-representados e fornece dados valiosos para políticas públicas, além de análise de desigualdades em contextos como saúde, educação e emprego.

Na computação, especialmente aprendizado de máquina, a inferência automática do tom de pele de um indivíduo enfrenta desafios consideráveis, muitos dos quais estão relacionados aos vieses raciais e de gênero, presentes em modelos, como os de reconhecimento facial [Buolamwini and Gebru 2018]. Uma abordagem comum para identificar esses vieses é analisar a diferença de desempenho dos modelos entre grupos majoritários e minoritários. No entanto, essa análise é complexa, em parte devido à interseção entre atributos de grupos minoritários e "atributos protegidos" – características individuais que, por questões legais, não devem ser consideradas em nenhuma tomada de decisão. A falta de dados rotulados representativos e o uso de escalas pouco precisas agravam o problema, tornando a auditoria algorítmica e a mitigação de vieses ainda mais desafiadoras.

Tipo	*i*	*ii*	*iii*	*iv*	*v*	*vi*
Cor						
Proporção	0.02	0.15	0.30	0.35	0.16	0.02

Tabela 1. Escala *Fitzpatrick*, cores correspondentes e proporções na base.

Para melhorar a inferência do tom de pele, uma abordagem comum é rotular bases de dados existentes usando escalas desenvolvidas para esse fim, como a *Fitzpatrick Skin Type*, que classifica o tom de pele com base na melanina e na resposta à exposição UV [Fitzpatrick 1988]. Neste contexto, métodos computacionais estão sendo propostos para inferir automaticamente o tom de pele a partir de imagens [Chardon et al. 1991], mas há poucos estudos que avaliam a aplicação dessas abordagens em contextos reais, como sistemas de reconhecimento facial, que têm impacto direto na vida das pessoas. Desta forma, avaliar os algoritmos considerando as diferentes escalas e potenciais vieses é crucial. Neste artigo comparamos abordagens amplamente exploradas em esforços anteriores para inferência automática de tom de pele a partir de uma imagem de entrada e, a partir de um conjunto de dados rotulado e padronizado, avaliamos o desempenho delas considerando diferentes métricas. Os resultados evidenciam que estas abordagens ainda não são robustas o suficiente para aplicação em contextos onde a alta precisão é indispensável.

2. Configuração Experimental

Nesta seção, descrevemos as abordagens computacionais para a inferência automática do tom de pele a partir de imagens, priorizando aquelas que suportam a escala de *Fitzpatrick*. Selecionamos duas abordagens: (i) *ITA* e (ii) *CASCo*. O (i) ITA (*Individual Topology Angle*) classifica o tom de pele em seis categorias com base na luz refletida, utilizando o espaço de cores L*a*b*. Esse método gera uma classificação objetiva que considera a quantidade de luz refletida e a cromaticidade das cores da pele [Chardon et al. 1991]. O ITA apresenta forte correlação com o índice de melanina e permite a extração fácil de informações a partir das imagens. Os grupos classificados, são: "*very light*", "*light*", "*intermediate*", "*tan*", "*brown*" e "*dark*", adaptáveis à escala de *Fitzpatrick* [Krishnapriya et al. 2022]. Já o (ii) CASCo (*Classification Algorithm for Skin Color*) [René Alejandro Rejón Piña 2023] utiliza detecção facial, segmentação de pele e agrupamento K-Means para classificar tons de pele em retratos. O algoritmo processa automaticamente as imagens, filtrando áreas não relacionadas à pele, como olhos e cabelos, e identificando os tons de pele dominantes, com base no espaço de cores HSV. Nesse estudo adotamos uma escala visual de *Fitzpatrick*, conforme [Leeb et al. 2024].

Para garantir a comparabilidade das abordagens, padronizamos a saída de cada método utilizando a escala *Fitzpatrick*, reconhecida e aceita na comunidade dermatológica e de pesquisa [Buolamwini and Gebru 2018]. Esta escala classifica o tom de pele em seis tipos, com base na resposta da pele à luz solar e na quantidade de melanina. A Tabela 1 apresenta a escala visual, suas classes denominadas Tipo (*Type*) de *i* a *vi*, com valores retirados de [Leeb et al. 2024] (linha "Cor"). Além disso, para avaliar o desempenho das abordagens na identificação automática de tons de pele, utilizamos a base de dados *Casual Conversations*[1] versão 2, que contém 26.467 vídeos de 5.567 voluntários com informações autodeclaradas, como idade, gênero e atributos físicos, além de rótulos de tom de pele definidos por anotadores treinados com base em escalas distintas, incluindo

[1] `https://ai.meta.com/datasets/casual-conversations-v2-downloads/`

Classe	ITA			CASCo		
	Precisão	**Revocação**	**F1-escore**	**Precisão**	**Revocação**	**F1-escore**
Tipo i	0.19	0.16	0.17	0.18	0.34	0.24
Tipo ii	0.26	0.26	0.26	0.12	0.01	0.01
Tipo iii	0.40	0.29	0.34	0.36	0.16	0.22
Tipo iv	0.06	0.20	0.10	0.01	0.00	0.00
Tipo v	0.35	0.17	0.23	0.31	0.35	0.33
Tipo vi	0.03	0.45	0.06	0.02	0.14	0.03
Micro avg			0.22			0.23

Tabela 2. Métricas obtidas a partir da execução das abordagens ITA and CASCo.

a *Fitzpatrick*, explorada neste estudo. Essa escolha foi motivada pela necessidade de imagens rotuladas em formato de retrato, fundamentais para a detecção facial eficaz. Extraímos uma imagem de cada participante, totalizando 5.567 imagens, predominando os tons intermediários (Tipos *iii*, *iv* e *v*) na escala *Fitzpatrick*. A proporção de imagens por classe (i.e., tipo) também é apresentada Tabela 1 (linha "Proporção"). Por fim, tratamos o problema como uma tarefa de classificação e utilizamos métricas como precisão, revocação, e F1-escore para avaliar o desempenho das abordagens de inferência do tom de pele.

3. Resultados

Realizamos uma análise comparativa das abordagens ITA e CASCo a partir da base de dados rotulada dos tons de pele, cujos resultados são apresentados na Tabela 2. Os modelos mostraram desempenhos discrepantes nas métricas de precisão, revocação e F1-escore. O ITA apresentou maior estabilidade na precisão, embora com valores mais baixos, enquanto o CASCo variou, apresentando resultados inferiores a 1% para a classe "Tipo *iv*". Ambos os modelos enfrentaram dificuldades na identificação das classes, com o ITA se destacando na detecção de tons mais escuros e o CASCo exibindo melhor desempenho para tons intermediários. Os valores de F1-escore indicam um equilíbrio insatisfatório entre precisão e revocação para ambos os modelos, exceto em alguns casos específicos.

Na Figura 1 apresentamos matrizes de confusão das abordagens. Observamos que ambas possuem vieses distintos. O CASCo tende a prever melhor os tons médios (i.e., *iv* e *v*), como evidenciado pelos altos valores fora da diagonal principal. Isso pode ser atribuído à pouca distância entre valores consecutivos na paleta e à iluminação das imagens testadas, o que dificulta a classificação. Em contraste, o ITA concentra mais erros na diagonal principal, agrupando tons próximos, mas gerando uma escala menos precisa. Para investigar o impacto do desbalanceamento dos dados no desempenho das abordagens, conduzimos um experimento adicional com amostras balanceadas. Utilizamos o método *bootstrap* para gerar 100 amostras balanceadas, com 849 instâncias de cada classe e calculamos intervalos de confiança para todas as métricas. Embora o desempenho tenha melhorado, as diferenças em relação à base desbalanceada não foram substanciais. O ITA obteve uma micro-F1 de $25\% \pm 0.01$, representando um aumento de $\approx 3\%$ em relação à base desbalanceada, enquanto o CASCo apresentou uma queda de $\approx 7\%$, alcançando $16\% \pm 0.01$. Em suma, essa queda reflete o viés do CASCo em relação aos tipos *iii* e *iv* o desbalanceamento original, tornando-o ainda mais evidente no conjunto balanceado.

4. Discussões e Conclusão

Este trabalho avaliou métodos computacionais de inferência automática de tons de pele, destacando suas limitações. Em resumo, constatou-se que tanto a abordagem ITA quanto a CASCo apresentam limitações significativas em termos de acurácia e consistência,

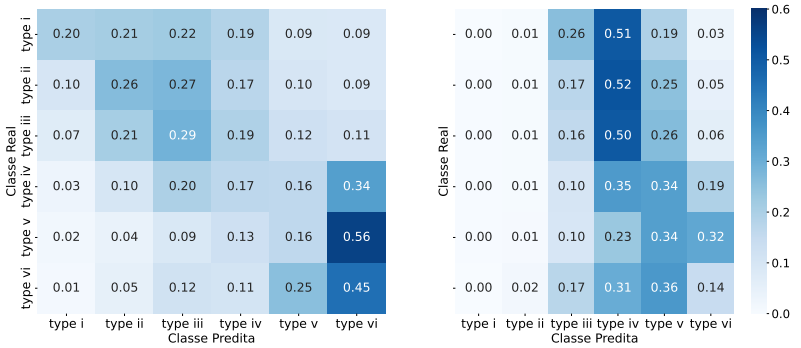

Figura 1. Matrizes de confusão: ITA à esquerda e CASCo à direita.

sendo o ITA mais eficaz para tons mais escuros e o CASCo para tons intermediários. Ademais, ambos apresentaram baixa revocação, sugerindo limitações para aplicações que requerem alta precisão. Embora esses métodos ofereçam uma solução prática para classificação rápida, a baixa acurácia e inconsistências levantam preocupações sobre sua eficácia, especialmente em países como o Brasil, onde a diversidade populacional é vasta. O uso inadequado pode, assim, resultar em classificações imprecisas ou generalizações inadequadas, tornando essencial uma análise cuidadosa por profissionais especializados. Futuras pesquisas devem explorar cenários que representem melhor a diversidade brasileira, aprimorando a precisão e aplicabilidade dos métodos.

Agradecimentos. MPMG, CNPq, CAPES, FAPEMIG, FAPESP e CIIA-Saúde.

Referências

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of FAT*.

Chardon, A., Cretois, I., and Hourseau, C. (1991). Skin colour typology and suntanning pathways. *Int. J. of Cosmet. Sci.*, 13(4):191–208.

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871.

Khosla, N., Grullon, K., and Rosenblatt, A. (2021). Prevention of racialized medicine in pediatric dermatology: A call to re-examine skin tone typing. *Pediatric Dermatology*, 38:167–169.

Krishnapriya, K., Pangelinan, G., King, M. C., and Bowyer, K. W. (2022). Analysis of manual and automated skin tone assignments. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 429–438.

Leeb, G., Auchus, I., Law, T., Bickler, P. E., Feiner, J., Hashi, S., Monk, E. P., Igaga, E., Bernstein, M., Chou, Y., Hughes, C., Schornack, D., Lester, J., Moore, K., Okunlola, O., Fernandez, J., Shmuylovich, L., and Lipnick, M. (2024). The performance of 11 fingertip pulse oximeters during hypoxemia in healthy human participants with varied, quantified skin pigment. *EBioMedicine*.

René Alejandro Rejón Piña, C. M. (2023). Classification algorithm for skin color (casco): A new tool to measure skin color in social science research. *Social Science Quarterly*.

Schumann, C., Olanubi, G., Wright, A., Monk, J., Heldreth, C., and Ricco, S. (2023). Consensus and subjectivity of skin tone annotation for ml fairness. In *NeurIPS*.