PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

USO DE MACHINE LEARNING PARA ESTIMAR O CONFORTO PERCEBIDO EM HUMANOS

THALES AUGUSTO SANTIN
PEDRO DALMAZO VAZ

Proposta de Trabalho de Conclusão apresentada como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Rafael Scopel Silva

Porto Alegre
2024

SUMÁRIO

1. INTRODUÇÃO

A computação gráfica revolucionou a forma como visualizamos e interagimos com os mundos digitais. No centro desta revolução estão os personagens 3D virtuais, que têm a capacidade de criar experiências visuais imersivas para o público. Sua aplicação tem se expandido continuamente, desde o desenvolvimento de grandes animações cinematográficas até a criação de mascotes e avatares personalizados para ambientes virtuais.

No entanto, é essencial abordar a teoria da *Uncanny Valley* (UV) [Mor70] ao projetar esses personagens, para garantir que sua aparência não cause desconforto por parecer quase humana, mas não totalmente convincente. A teoria supõe que quando as réplicas humanas se comportam de forma muito parecida, mas não idêntica, a seres humanos reais, podem provocar repulsa entre os observadores humanos. Neste contexto, é crucial que a percepção visual de personagens virtuais não cause desconforto para seres humanos, pois qualquer sensação de estranheza pode prejudicar a aceitação de sua mídia.

O Machine Learning (ML) surge como uma solução para abordar o desafio da UV na modelagem 3D de personagens virtuais. Por meio do treinamento com vastos conjuntos de dados, algoritmos de ML podem aprender e identificar as características que tornam um personagem virtual mais aceitável ou repulsivo para observadores humanos.

Recentemente, um trabalho conduzido por Dal Molin, Araújo e Musse [DMdAAM22] explorou o fenômeno do UV e sua relação com humanos virtuais, introduzindo uma nova métrica de conforto e desconforto usando computação gráfica (CG), denominada pelos autores de *Computed Comfort Score* (CCS). Além disso, os autores propuseram um modelo que se utiliza dessa métrica para estimar o nível de conforto que uma face CG causaria na percepção humana.

Este trabalho associa-se aos esforços dos autores e busca compreender como o CCS aplicado a um modelo de ML pode contribuir para estimar o nível de conforto de personagens virtuais. Buscamos explorar mais a fundo os conceitos de UV, juntamente com a importância das características que levam à percepção positiva ou negativa de um personagem. Reutilizamos o dataset do trabalho de Dal Molin, Araújo e Musse [DMdAAM22] e reimplementaremos o melhor modelo desse estudo, que utiliza Support Vector Regression (SVR) como algoritmo de predição. Com o mesmo dataset, implementaremos nove variações de modelos com características variadas, utilizando Convolutional Neural Networks (CNN). Compararemos o desempenho dos modelos CNN pelas métricas de Precisão de Classificação Binária e Erro Absoluto Médio, e selecionaremos o(s) melhor(es) para comparação com o modelo de SVR. Com isso, buscamos melhorar a performance de modelos de predição com base em CCS e esperamos contribuir para uma melhor compreensão das capacidades e diferenças entre os dois tipos de modelos estudados, SVR e CNN.

Este trabalho está estruturado de modo a fornecer uma análise abrangente da aplicação de Machine Learning na percepção de conforto de personagens virtuais 3D. Iniciando com uma revisão de literatura no Capítulo 2, exploramos estudos anteriores e teorias relevantes, especialmente focando na UV e sua relação com a CG. No Capítulo 3, definimos os objetivos específicos do projeto, delineando as questões de pesquisa e as metas a serem alcançadas. A metodologia empregada, incluindo a descrição do dataset, pré-processamento de dados, e as abordagens de modelagem usando SVR e CNN são detalhadas no Capítulo 4. No Capítulo 5 discutimos os resultados adiquirados durante o trabalho e realizamos comparações entre os modelos desenvolvidos. Finalmente, no Capítulo 6, ponderamos sobre o trabalho, considerando o decorrer e os resultados do mesmo.

2. REVISÃO DE LITERATURA

Este trabalho está diretamente relacionado ao trabalho realizado por Dal Molin, Araújo e Musse [DMdAAM22] na criação de um modelo de ML para avaliação de personagens em uma nova métrica CCS, introduzida no mesmo trabalho, com o objetivo de estimar o conforto percebido de personagens virtuais. O modelo desse trabalho se baseia em SVR, uma generalização do algoritmo de Support Vector Machine (SVM) para aplicação em problemas de regressão. Segundo Awad e Khanna [AK15], SVRs são algoritmos supervisionados baseados em *frameworks* de aprendizado estatísticos treinados a partir de uma função de perda simétrica. Ainda segundo os autores, SVRs aparentemente se mostraram bem posicionados para lidar com problemas de classificação em situações reais.

A importância do conforto percebido é algo já estudado e, no âmbito desse trabalho, o conceito de UV, proposto por Mori [Mor70], é um ponto central. Podemos explicar o UV da seguinte forma: à medida que a semelhança humana de um robô aumenta, a nossa afinidade pelo robô também aumenta. Porém, isso ocorre somente até certo ponto, quando essa afinidade despenca em um 'vale'. Ou seja, há uma percepção positiva do robô quando a semelhança deste robô 'atravessa' o vale. Esse conceito permaneceu um tempo em desuso, mas foi retomado devido à sua relevância em trabalhos como o de MacDorman [MI06], que tratou dos mecanismos sociais de androides, e Hanson [Han06], que abordou a exploração da extensão estética de robôs humanoides.

Tumblin e Ferwerda [TF01] propuseram que a percepção é composta por um conjunto de processos que, além de registrar passivamente os estímulos físicos, também constrói representações mentais da realidade e, assim, é responsável por fornecer estimativas do momento presente, essenciais para a compreensão e interação com o ambiente. Zell, Zibrek e McDonnell [ZZM19] descrevem como avaliações de dados perceptivos são essenciais para entender a percepção humana de CG. Prakash e Rogers [PR14] explicaram que o recente aumento de realismo gráfico também aumenta as expectativas para humanos virtuais. MacDorman, Coram, Ho e Patel [MCHP10] afirmam que novas modelagens e simulações em ambientes 3D podem ser utilizadas para avaliações psicológicas e de entretenimento.

Considerando esse contexto, Tinwell, Grimshaw, Nabi e Williams [TGANW11] indicam que a preocupação com a avaliação da aparência e comportamento de personagens virtuais devido ao UV é relevante para um grande escopo de aplicações. Bailenson, Swinth, Hoyt, Persky, Dimov e Blascovich [BSH+05], por sua vez, demonstraram em seu estudo algumas características de personagens virtuais que são mais comumente percebidas como não naturais. Por exemplo, características como movimentos da cabeça e expressões faciais que não se alinham com os padrões humanos naturais são frequentemente percebidas como não naturais em personagens virtuais.

A partir da percepção de necessidade de conforto, trabalhos como o de Dal Molin, Araújo e Musse [DMdAAM22] demonstram que métodos como a transformação de personagens em *cartoons*, como forma de distanciamento de uma imagem realista do ser humano, ou fatores como a familiaridade e o carisma de um personagem, podem aumentar o nível de conforto de um personagem. Outros estudos da literatura também abordam a questão, como Kätsyri, Mäkäräinen e Takala [KMT17], que utilizaram a ferramenta Toonify para realizar a conversão de personagens mais realistas para cartoon, e Choudhary, Kim, Schubert, Bruder e Welch [CKS+20], que apresentaram uma técnica chamada *Big Head*, na qual a cabeça de personagens virtuais é aumentada para dar mais ênfase às feições e movimentos faciais. Esses estudos também corroboram o fato de que personagens não realistas tendem a ser mais confortáveis para a percepção humana.

Nestes mesmos trabalhos foram também explorados outros fatores, como altura dos olhos e altura do pescoço e distância do interlocutor, concluindo que o escalonamento baseado na altura dos olhos e a distância da percepção do interlocutor são fatores de grande relevância ao implementar técnicas para aumentar o conforto percebido. Este conhecimento como um todo é essencial para melhorar as métricas de conforto no design de humanos virtuais e pode influenciar a precisão de modelos de detecção do UV, uma vez que mais fatores do personagem e da composição da cena se tornam relevantes.

Além disso, é importante destacar que outras características, como a idade do público-alvo de uma mídia e fatores como o tipo de sons emitidos em uma cena, que também podem ser considerados ao corrigir o efeito UV. Isso é evidenciado por estudos como o realizado por Tu, Chien e Yeh [TCY20], que aplicaram um questionário em diferentes grupos etários relacionado ao efeito de estranhamento. O questionário foi respondido por participantes com idades entre 18 e 39 anos, 40 e 59 anos e um último grupo etário entre 60 e 87 anos. O estudo concluiu que diferentes faixas etárias podem possuir diferentes percepções de conforto, pois foi observado que o efeito de UV foi percebido apenas nos dois grupos mais jovens. O grupo mais velho demonstrou preferência por robôs semelhantes a humanos em detrimento dos não semelhantes a humanos, independentemente da função do robô.

Esses resultados são reforçados pela pesquisa de Yorgancigil, Urgan e Yildirim [YUY21], onde um estudo relacionado ao efeito do conjunto audiovisual e percepção do UV foi realizado com um grupo de faixa etária variada, apresentando animações com diversos personagens e um de quatro níveis de naturalidade dos estímulos auditivos: robô (irrealista), semi-robô (semi-realista), quase humano (realista) e humano (real). A partir dos resultados, concluiu-se que tanto a naturalidade quanto a incongruência do conjunto audiovisual são relevantes para a percepção do efeito, e que a relevância desses fatores muda com a idade das pessoas.

Considerando a contextualização trazida pela literatura em UV, é evidente que técnicas para análise de imagens com múltiplos fatores são relevantes para alcançar uma solu-

ção do problema. Novos métodos para detecção de objetos, como o desenvolvido por Viola e Jones [VJ01], que usa um novo tipo de representação de imagem chamado de *integral image*, junto de um método de cascata se demonstram excelente para detecção de objetos de maneira ágil. Além disso, há métricas para visão computacional, baseadas na teoria de reconhecimento de padrões visuais, como os *Hu moments* introduzidos por Hu [Hu62], e Zunic, Hirota e Rosin [R10], as quais se mostram promissoras para identificação das diferenças da forma de um objeto em relação a outras formas a partir de *moment invariants*. Outro trabalho interessante é o de Dalal e Triggs [DT05], que demonstra que o uso de *Histograms of Oriented Gradients* (HOG) como um descritor para uso em reconhecimento de faces performa significativamente melhor que outras formas existentes.

Nesse sentido, Chen, Li, Bai, Yang, Jiang e Miao [CLB+21] explicam que algoritmos baseados em CNN vêm há um tempo se tornando as melhores escolhas quanto à classificação de imagens e têm atingido níveis de precisão avançados. Além disso, diversos *frameworks* de classificação como o proposto por Mengmen, Li e Du [ZLD18], chamado de *diverse region-based CNN*, se mostram promissores em detectar um conjunto diversificado de fatores de aparência discriminativos.

Outro trabalho baseado em CNN que lida com reconhecimento facial e detecção do UV é o artigo de Imaizumi [ILU23], que utiliza FaceNet para detecção facial e o método Grad-CAM para criar um mapa de calor das características relevantes na detecção do UV. Como base para o trabalho de Imaizumi, foi realizada uma pesquisa de campo onde participantes classificaram a *likability* de 182 imagens faciais. Isso resultou em uma pontuação média humana (pontuação MH) para as imagens do questionário. Porém, como conclusão deste trabalho, ficou evidente que apenas parte do efeito do UV pode ser replicada utilizando o método FaceNet. Observou-se discrepância sobre quais partes do rosto são relevantes para a classificação. O conceito do UV sugere que olhos e bocas são pontos relevantes para um humano detectar o vale. Em contraste, o mapa de calor do estudo de Imaizumi indica que o queixo e a boca são as regiões mais importantes para a classificação segundo o aprendizado de máquina, o que sugere que humanos e máquinas podem ter áreas diferentes de foco.

AbuRass, Huneiti e Al-Zoubi [AHAZ20] implementam um modelo de aprendizado profundo utilizando CNN e Hu moments. Demonstram não somente que é possível a utilização de Hu moments em implementações baseadas em CNN como também que é possível ultrapassar as limitações de invariância associadas a CNN com esse método. Considerando que a orientação de um personagem pode variar durante uma cena, é possível que essa consideração seja importante para a avaliação de imagens contendo personagens.

Mais um caso relevante é o estudo de Laine, Karras, Aila, Herva, Saito, Yu, Li e Lehtinen [LKA+17] que realizaram uma implementação de um software de reconhecimento de faces utilizando um modelo de aprendizagem profunda baseado em CNN. Este trabalho é um exemplo de uma aplicação de reconhecimento de imagens no nível de produção e de-

monstra que é possível realizar implementações como essa para hardware de dispositivos móveis.

3. OBJETIVO

3.1 Problema de Pesquisa

Como um modelo de ML pode contribuir para estimar o nível de conforto de personagens virtuais 3D baseado na métrica CCS? Quais outros modelos podem ser utilizados para esta finalidade?

3.2 Objetivos

3.2.1 Objetivo Geral

Sugerir e avaliar um novo modelo de ML baseado em CNN que utiliza a métrica CCS para estimar o nível de conforto de personagens virtuais 3D e comparar os resultados com o melhor modelo baseado em SVR do trabalho original que utiliza a mesma métrica, fornecendo uma comparação relevante entre os dois tipos de modelos e qual possivelmente seria mais adequado para o problema.

3.2.2 Objetivos Específicos

a) Replicar o melhor modelo de ML baseado em SVR abordado no trabalho de Dal Molin, Araújo e Musse [DMND+21];

b) Propor e implementar novos modelos de ML baseados em CNN;

c) Comparar e avaliar os resultados dos modelos implementados e a aplicabilidade da métrica CCS;

4. METODOLOGIA

4.1 Natureza da Pesquisa

Este capítulo descreve os dados e técnicas utilizados para a realização deste estudo. Porém, antes de detalhar esses aspectos, é necessário esclarecer que a natureza desta pesquisa é tanto aplicada quanto exploratória. Caracteriza-se como aplicada por aplicar conceitos de CNN para resolver um problema específico, previamente explorado através de SVR. Simultaneamente, é exploratória, pois investiga o potencial das CNNs em um trabalho que, até então, utilizava somente SVR. Por fim, a pesquisa possui um caráter comparativo, avaliando o desempenho das CNNs comparado ao melhor método de SVR, contribuindo com *insights* valiosos sobre a eficiência e precisão de diferentes abordagens de modelagem.

4.2 Dataset

O conjunto de dados utilizado neste trabalho é o mesmo que foi criado e explorado no trabalho de Dal Molin, Nomura, Dalmoro, de A. Araújo e Musse [DMND+21]. Contém 22 personagens consistindo de fotos e pequenos vídeos (posteriormente divididos em *frames*) e os valores de conforto de cada personagem. Cada foto e *frame* constitui uma imagem da qual serão retiradas as características do personagem. Para garantir uma boa variação de similaridade humana presente no Uncanny Valley, alguns personagens estão representados de maneira *cartoon* e outros mais realistas. Os valores de conforto de cada personagem foram estabelecidos a partir de uma pesquisa, realizada por Dal Molin et al.(2022), com 119 participantes, sendo estes 58% homens e 42% mulheres, com 77.3% com até 31 anos e 33.7% com mais de 31 anos.

A cada participante foram feitas as perguntas da Tabela 4.1 referentes a cada personagem. A média das respostas de Q1 foi utilizada para determinar o nível de realismo percebido nos três níveis mencionados nas escolhas da questão, enquanto a porcentagem de respostas 'Não' de Q2 é utilizada como conforto percebido. Nosso trabalho em específico não utiliza os resultados de Q3 pois não entramos no mérito de avaliar as partes da face de um personagem que podem causar maior ou menor conforto/desconforto.

Qn	Pergunta	Escolhas
Q1	Quão realista é esse personagem?	"Irrealista" "Moderadamente Realista" "Muito Realista"
Q2	Você sente algum desconforto olhando para este personagem?	"Sim" "Não"
Q3	Em que partes da face você sente maior estranheza?	"Olhos" "Boca" "Nariz" "Cabelo" "Outros" "Eu não senti desconforto"

Tabela 4.1 – Perguntas da Pesquisa de Conforto Percebido

4.3 Pré-processamento dos Dados

O pré-processamento dos dados nesse trabalho é feito de acordo com o do trabalho original de Dal Molin, Araújo e Musse (2021) [DMND+21], portanto são realizados quatro processos para o preparo dos dados.

O primeiro se trata da extração de cada frame dos vídeos no dataset de cada personagem e da associação de cada frame com as métricas de conforto obtidas na pesquisa do trabalho original, que estão disponíveis em um CSV. Deste modo, teremos um vasto dataset para treinarmos nossos modelos.

O segundo processo envolve a detecção facial, onde originalmente se utilizava o método de Paul Viola e Michael Jones [VJ01]. Neste trabalho, optamos por atualizar esta etapa utilizando o modelo Face Mesh do MediaPipe [HSN23]. Esta mudança se justifica pela capacidade superior do MediaPipe em capturar uma vasta gama de pontos de referência faciais de forma precisa, nos permitindo uma extração mais detalhada e precisa de características faciais tais como olhos (direito e esquerdo), sobrancelhas (direita e esquerda), nariz e boca. Imagens onde o rosto ou qualquer uma dessas características não são claramente detectadas são descartadas.

O terceiro processo é a extração das características. Nesta etapa, cada imagem é redimensionada para um múltiplo de dois e particionada em blocos de 8x8. São então localmente computadas as características de entropia espaciais e espectrais de cada bloco e cada região de interesse, como a face, boca e nariz.

Finalmente, o quarto processo utiliza a computação de entropia da etapa anterior para calcular outras características para cada bloco e região de interesse das imagens. As características incluem a média, desvio padrão, distorção, curtose, variância, Hu moments [R10] e Histogram of Oriented Gradients(HOG) [DT05].

4.4 Predição usando SVR

Após extrair as propriedades essenciais dos dados de cada personagem, estas são usadas para a realização das etapas de treino, teste e validação do algoritmo SVR. Primeiro são separados 20% dos dados para validação, o restante dos dados são variados de forma que todos os personagens participem das etapas de treino e teste. Essa separação leva em conta a separação dos dados de cada personagem, garantindo que nenhum personagem apareça em mais de um grupo. Conforme os resultados do trabalho de Dal Molin, Araújo e Musse [DMND+21], o modelo com os melhores resultados de performance na tarefa de predição foi o modelo que incorpora todos os dados de características da imagem, portanto, é implementado apenas o modelo que utiliza todas as características para representação das formas das imagens faciais na modelagem deste caso.

O algoritmo então computa os dados e avalia as características alimentadas para representar o conforto percebido, gerando valores individuais de conforto, CCS, para cada imagem (*frame*) do pequeno vídeo de cada personagem. Baseado nesses valores, calculamos o CCS final como sendo a média dos valores de conforto obtidos em cada imagem do vídeo em que aquele personagem aparece. Neste modelo, utilizamos o método GroupKFold com 4 splits para a separação dos dados entre treino e teste, garantindo que os dados de cada personagem não fossem misturados entre os conjuntos. Os seguintes parâmetros foram utilizados: C=0.2, epsilon=0.5, PolynomialFeaturesDegree=2. O kernel utilizado é o padrão, Radial Basis Function (RBF).

4.5 Predição usando CNN

Para realizar a predição com CNN, partimos do mesmo processo de separação e variação dos dados entre treino, teste e validação usado com o algoritmo de SVR, porém alimentamos esses dados para o algoritmo dos modelos de CNN. Nesta etapa são implementados 9 modelos, seguindo o exemplo do trabalho original, onde as características alimentadas aos modelos variam com o intuito de avaliar seus impactos e encontrar a combinação mais relevante para classificação. Todos os modelos CNN seguiram a mesma arquitetura detalhadas na Tabela 4.2. Usamos 'mean_absolute_error' como loss e 'adam' como optimizer, cada modelo foi treinado em 50 epochs, utilizando um early stopping de 'val_loss' com paciência 10. As combinações de características utilizadas são de acordo com a Tabela 4.3.

A predição dos modelos de CNN é tratada da mesma forma que a do modelo usando SVR, com a diferença de que cada modelo é alimentado somente com os dados das características de imagens que estes foram designados. Assim, cada modelo computa

Camada	Tipo	Parâmetros Principais	Função de Ativação
1	Conv1D	64, kernel_size=2	ReLU
2	MaxPooling1D	pool_size=2	-
3	Conv1D	128, kernel_size=2	ReLU
4	MaxPooling1D	pool_size=2	-
5	Flatten	-	-
6	Dense	128	ReLU
7	Dropout	0.5	-
8	Dense	1	Sigmoid
9	Lambda	x: x * 100	-

Tabela 4.2 – Arquitetura dos Modelos CNN

os valores de CCS do mesmo modo, porém levando em conta apenas a sua combinação de dados.

Model	Spatial Entropy	Spectral Entropy	Statistics Features	HOG	Hu Moments
1	X	X	X	X	X
2	X	X	X	X	
3	X	X	X		X
4	X		X	X	X
5	X		X	X	
6	X		X		X
7		X	X	X	X
8		X	X	X	
9		X	X		X

Tabela 4.3 – Variações de Características por Modelo

4.6 Comparação dos Modelos

A performance de ambos os modelos será avaliada usando as seguintes métricas:

- **Precisão da Classificação Binária:** Mede a capacidade do modelo de utilizar o CCS para categorizar corretamente os personagens como confortáveis ou desconfortáveis. Para esta análise, diferentes pontos de corte foram considerados (35%, 40%, 60%, 70% e 75%), com o objetivo de avaliar a precisão em diversas faixas de conforto percebido. Um CCS abaixo do ponto de corte definido indica desconforto, enquanto valores iguais ou superiores ao ponto de corte sugerem conforto.
- **Erro Absoluto Médio (MAE):** Calcula a média das diferenças absolutas entre os valores de CCS previstos pelo modelo e as notas de conforto percebidas, fornecendo uma medida direta da acurácia das previsões de conforto.

- **Erro Absoluto Percentual Médio (MAPE):** Calcula a porcentagem da média das diferenças absolutas entre os valores de CCS previstos pelo modelo e as notas de conforto percebidas, fornecendo uma medida relativa da acurácia das previsões de conforto.

Além disso, adotaremos a mesma abordagem do estudo original para avaliar a significância das diferenças entre os modelos, assegurando que as métricas de performance Precisão da Classificação Binária e Erro Absoluto Médio (MAE) sejam aplicadas de maneira consistente para uma comparação válida e direta.

5. RESULTADOS

5.1 Pré-processamento e Detecção Facial

Após o carregamento dos dados dos 22 personagens contidos no dataset, realizamos uma filtragem baseada na detecção facial. Utilizando a biblioteca MediaPipe, analisamos cada personagem para garantir que de cada um fosse possível detectar todos os aspectos relevantes para o nosso modelo, esses sendo rosto, olhos, boca e nariz, além de outros como sobrancelhas, removendo personagens onde não conseguimos realizar a detecção.

Para um segundo processo de filtragem, analisamos os frames individuais dos vídeos de cada personagem restante e mantivemos apenas aqueles em que pelo menos a face do personagem está presente. Esse processo foi aplicado aos 22 personagens do dataset, com os personagens onde não houve detecções adicionais sendo removidos dos dados. No final, restaram um total de 21 personagens restantes para posterior análise pelos modelos, o personagem não detectado neste caso foi o personagem 21 que aparece na figura 5.2. A figura abaixo ilustra alguns exemplos de detecção de faces nas imagens do dataset. As faces detectadas são marcadas com caixas delimitadoras verdes, indicando o reconhecimento bem-sucedido das características faciais.

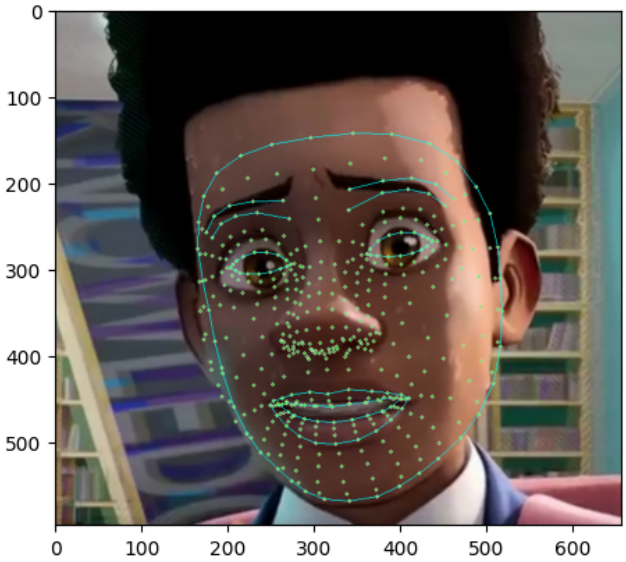

Figura 5.1 – Detecção de faces e suas características em um personagem virtual.

Em sequência, submetemos cada imagem ao processo de extração de características. Primeiramente, redimensionamos cada imagem para seus próximos múltiplos de dois mais baixos e então as dividimos em blocos de 8x8, segundo os resultados do trabalho de Liu L, Liu B, Huang e Bovik [LLHB14]. Para cada bloco, realizamos a extração de

Figura 5.2 – Personagem 21, Falha de Detecção Facial

características de Média, Desvio Padrão, Variância, Skewness, Curtose, Entropia Espacial e Espectral, Hu Moments e HOG. Para as características de Skewness, Curtose e Entropia, utilizamos a biblioteca scipy; para Exposure e HOG, a biblioteca skimage; e para Hu Moments, a biblioteca cv2.

As características extraídas foram posteriormente correlacionadas com as pontuações de conforto extraídas do dataset, com as característcas de cada frame sendo associados ao valor de conforto do personagem da imagem da qual essas foram extraidas. Esse processo resultou em um dataframe que combina as características e valores de conforto para cada imagem de cada personagem. No total, obtivemos 8211 pontos de dados contemplando 22 personagens.

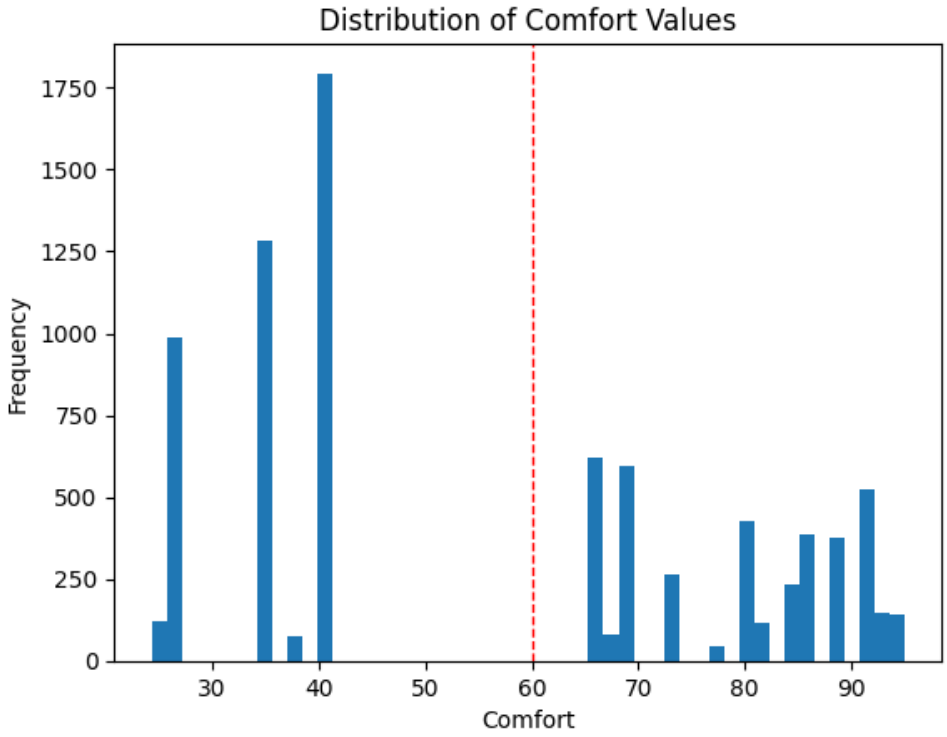

Figura 5.3 – Detecção de faces e suas características em um personagem virtual.

5.2 Desempenho do Modelo SVR

Utilizando os dados do dataframe resultante do processo de pré-processamento, treinamos, testamos e validamos nosso modelo de SVR, variando os personagens dos os grupos de treino e teste até que todos os personagens tenham participado de cada grupo. Na execução do modelo, utilizamos o melhor conjunto de características com base nos resultados do trabalho de Dal Molin, Nomura, Dalmoro, de A. Araújo e Musse [DMND+21], que considera todas as características extraídas das imagens.

O modelo resultante é então usado para gerar valores de conforto individuais para cada imagem de cada personagem, a métrica CCS proposta no trabalho de Dal Molin, Nomura, Dalmoro, de A. Araújo e Musse [DMND+21]. A média dos valores de cada imagem de um personagem é considerada o valor de CCS atribuído àquele personagem pelo modelo.

Com isso, realizamos a classificação binária dos personagens, considerando que valores de conforto(CCS) <60 podem gerar desconforto na percepção humana, enquanto valores >=60 não geram desconforto. Avaliamos a acurácia do modelo comparando as classificações feitas por este em relação aos resultados da pesquisa original. Nesta avaliação, obtivemos uma acurácia de 88,04%, com um MAE de 7.96 e um MAPE de 15,90%.

É interessante ressaltar que os resultados de acurácia do modelo podem variar a depender do ponto de corte usado para determinar caso um personagem é ou não confortável, a acurácia em diferentes pontos é demonstrada na Tabela 5.1. Esses pontos foram escolhidos por terem uma maior densidade de pontos de dados nos intervalos adjacentes e não serem tão próximos aos extremos. Observando o aumento de acúracia quando diminuimos o ponto de corte pode indicar uma tendência do modelo de avaliar positivamente os personagens.

Pontos de Corte	35	40	60	70	75
Accurácia	91,7%	91,4%	88,0%	80,4%	73.7%

Tabela 5.1 – Accurácias do Modelo SVR com diferentes pontos de corte

5.3 Desempenho dos Modelos CNN

Utilizando os mesmos dados usados pelo modelo SVR, preparamos um novo dataset adequado para utilização pelo nosso modelo CNN. Após ajustarmos os dados, seguimos o processo de treino, teste e validação, inicialmente implementando o mesmo procedimento descrito por Dal Molin, Nomura, Dalmoro, de A. Araújo e Musse [DMND+21], em que a va-

lidação também variava. No entanto, essa abordagem levou a problemas de overfitting durante os testes, prejudicando a capacidade dos modelos de generalizar os resultados.

Para mitigar esse problema, ajustamos a metodologia, optando por manter a validação fixa enquanto variávamos os dados de treino e teste. Essa mudança buscou garantir maior consistência na avaliação do desempenho dos modelos CNN, reduzindo o impacto do overfitting e proporcionando resultados mais confiáveis.

Novamente, seguindo o mesmo processo do modelo SVR, os modelos CNN resultantes são usados para gerar valores de conforto, ou CCS, individuais para cada imagem de cada personagem, sendo a média dos valores de cada imagem de um personagem o valor de CCS atribuído àquele personagem.

Com esses resultados em mãos, realizamos a classificação binária dos personagens utilizando o mesmo ponto de corte de 60 para os valores de CCS preditos, com personagens acima do valor de corte sendo considerados como não gerando desconforto e personagens abaixo como gerando desconforto. A acurácia obtida por cada um dos 9 modelos é demonstrada na Tabela 5.2, junto dos valores de MAE e MAPE de cada modelo.

M	SpatialEntropy	SpectralEntropy	StatisticsFeatures	HOG	HuMoments	Accurácia	MAE	MAPE
1	X	X	X	X	X	56,80%	23.94	50.3%
2	X	X	X	X		49,29%	25.74	55.6%
3	X	X	X		X	64,45%	26.62	61.3%
4	X		X	X	X	49,52%	26.21	61.5%
5	X		X	X		68,03%	22.56	46.9%
6	X		X		X	54,75%	22.22	46.7%
7		X	X	X	X	51,90%	21.85	46.3%
8		X	X	X		51,28%	23.62	49.8%
9		X	X		X	55,83%	24.87	57.8%

Tabela 5.2 – Variação de Features por Modelo e Suas accurácias(60%)

Seguindo as suspeitas geradas pelo modelo SVR, também testamos a classificação binária dos nove modelos CNN com pontos de corte mais próximos dos intervalos com maior densidade de valores de CCS. Nesses casos, o mesmo efeito observado no modelo anterior, onde a acurácia oscila quando é escolhido um ponto de corte diferente. Porém, diferente do modelo SVR, os modelos baseados em CNN geralmente obtiveram resultados melhores quando os pontos de corte são movidos para áreas mais densas do dataset, principalmente quando o ponto de corte é maior. Os resultados aparecem na Tabela 5.3.

Quanto à performance dos modelos os resultados gerais não foram satisfatórios com a maioria dos modelos se mantendo por perto de 50% de acurácia no ponto de corte base de 60. Também não fica clara a superioridade de nenhum dos modelos, o modelo 5 possui a melhor acurácia porem apresenta um MAE elevado. Se considerarmos MAE e MAPE o modelo 7 se demonstra superior, em contrapartida não demonstra melhor acurácia do que a de outros modelos. Considerando ambos os resultados de acurácia e MAE o modelo que acaba se mostrando mais consistente é o modelo 5, tendo a melhor acurácia em 35, 40 e 60 e a segunda melhor em 70 e 75, com o terceiro melhor MAE. Em relação às

		Pontos de Corte				
		35	40	60	70	75
Modelos	1	55,34%	33,78%	56,86%	83,32%	76,07%
	2	65,26%	44,12%	49,29%	72,44%	82,65%
	3	70,97%	48,63%	64,45%	77,51%	80,08%
	4	63,64%	41,17%	49,52%	75,59%	77,61%
	5	72,20%	80,17%	68,03%	84,93%	86,94%
	6	62,29%	47,22%	54,75%	85,93%	89,13%
	7	67,06%	55,39%	51,90%	79,27%	83,07%
	8	69,11%	52,55%	51,28%	77,23%	80,24%
	9	67,95%	34,12%	55,83%	81,79%	77,08%

Tabela 5.3 – Variação de Accurácia por Ponto de Corte dos Modelos CNN

features utilizadas, os resultados obtidos não demonstraram diferença o bastante para que se realizasse uma análise relevante da influência das features para a predição dos modelos.

5.4 Comparação entre Modelos

Considerando os modelos produzidos neste trabalho podemos ver, pela Tabela 5.4, que a nossa reprodução do modelo SVR por Dal Molin, Nomura, Dalmoro, de A. Araújo e Musse [DMND+21] demonstra resultados de acurácia geralmente melhores dos que os obtidos pelo nosso modelo CNN mais consistente, perdendo apenas nos pontos de corte de 70 e 75. No entanto, é a diferença de MAE e MAPE que apresenta a maior distinção do modelo SVR, possuindo menos da metade dos scores obtidos pelos modelos de CNN, indicando uma capacidade muito maior de aproximar os valores de CCS.

Modelo	Acc.35	Acc.40	Acc.60	Acc.70	Acc.75	MAE	MAPE
SVR Reproduzido	91,7%	91,4%	88,0%	80,4%	73,7%	7,9	15,9%
CNN Modelo 5	72,20%	80,17%	68,03%	84,93%	86,94%	22,56	46,7%
SVR Original	-	-	80,00%	-	-	23,59	-%

Tabela 5.4 – Métricas SVR Reproduzido e CNN Modelo 5

Para a comparação com o SVR original utilizamos apenas as métricas acurácia e MAE que possuímos, os dados restantes do modelo original focavam também na capacidade do modelo em avaliar outros fatores dos personagens, como Realismo, que não abordamos diretamente neste trabalho.

Os resultados superiores do modelo SVR reproduzido podem ser explicados pela natureza do dataset, que se mostrou mais adequado à utilização de características extraídas manualmente, otimizando o aprendizado do SVR. Em contrapartida, os modelos CNN não obtiveram resultados satisfatórios, o que sugere limitações impostas pelo tamanho re-

duzido do dataset e pela configuração das arquiteturas empregadas, que podem não ter explorado de forma eficiente as características disponíveis.

A exclusão de um personagem no processo de filtragem não comprometeu a representatividade do conjunto de dados, dado que os 21 personagens restantes englobam uma ampla gama de características relevantes ao problema de pesquisa. Contudo, essa redução pode impactar a capacidade de generalização dos modelos, especialmente no caso das CNNs, que usualmente requerem um maior volume de dados para alcançar melhores desempenhos.

Os pontos de corte selecionados (35, 40, 60, 70, 75) foram definidos com base na densidade dos dados dentro do intervalo de valores de CCS, buscando refletir as faixas mais representativas do dataset. Essa abordagem permitiu uma análise mais detalhada da capacidade dos modelos em distinguir níveis de conforto percebido, respeitando a distribuição original dos dados.

Ainda assim, observando as duas métricas na Tabela 5.4, é possível afirmar que o modelo SVR reproduzido apresenta uma melhora em relação ao modelo original, com um pequeno aumento de acurácia e uma redução no MAE. Essas diferenças podem ser explicadas, em parte, pela mudança na técnica de reconhecimento facial empregada, que permitiu utilizar uma porção consideravelmente maior do dataset. Enquanto o modelo original foi treinado com 5730 imagens, o modelo reproduzido utilizou um total de 8211 imagens, ampliando a variabilidade dos dados disponíveis para o treinamento.

Os resultados obtidos confirmam a eficácia do modelo SVR reproduzido, que alcançou maior precisão na estimativa do conforto percebido em comparação aos modelos CNN implementados. Apesar do desempenho inferior, as CNNs oferecem uma alternativa promissora para estudos futuros, especialmente em contextos onde datasets mais amplos e diversificados possam ser explorados.

5.5 Conclusão

Neste trabalho, investigamos o uso de modelos de Machine Learning para a estimativa do conforto percebido em personagens virtuais 3D com base na métrica CCS, comparando abordagens baseadas em Support Vector Regression (SVR) e redes neurais convolucionais (CNN). Os resultados obtidos demonstraram que o modelo SVR reproduzido superou tanto a versão original quanto os modelos CNN, apresentando maior acurácia e menor erro absoluto médio (MAE).

Embora os modelos CNN tenham apresentado resultados inferiores no cenário avaliado, sua capacidade de aprendizado de características complexas diretamente das imagens aponta para um potencial promissor em estudos futuros. Este potencial pode ser explorado especialmente em contextos onde datasets maiores e mais diversificados estejam disponíveis, permitindo uma melhor generalização e maior precisão na estimativa do conforto percebido.

Adicionalmente, a substituição do método de detecção facial utilizado no trabalho original pelo modelo Face Mesh do MediaPipe mostrou-se eficiente, permitindo um aproveitamento mais amplo do dataset. Essa melhoria metodológica destacou a importância de técnicas de pré-processamento na obtenção de dados de maior qualidade, impactando diretamente os resultados dos modelos avaliados.

Com base nos resultados obtidos, é possível afirmar que o modelo SVR permanece uma escolha robusta e eficiente para a tarefa proposta, especialmente em cenários com datasets menores e características pré-definidas. Por outro lado, o desempenho promissor das CNNs em outros contextos sugere que sua aplicação em problemas semelhantes pode se beneficiar de investigações adicionais, como o uso de arquiteturas mais avançadas ou técnicas de transferência de aprendizado.

Por fim, este estudo explora o uso da métrica CCS como uma ferramenta para avaliar o conforto percebido em personagens virtuais, demonstrando sua aplicação prática em modelos de Machine Learning. A investigação realizada contribui para uma melhor compreensão de como a CCS pode ser utilizada em diferentes abordagens, como SVR e CNN, permitindo identificar limitações e oportunidades de aprimoramento. Trabalhos futuros poderão expandir essa exploração, integrando novas métricas e investigando a percepção de conforto em cenários mais diversos e desafiadores.

REFERÊNCIAS BIBLIOGRÁFICAS

[AHAZ20] AbuRass, S.; Huneiti, A.; Al-Zoubi, M. B. “Enhancing convolutional neural network using hu’s moments”, *International Journal of Advanced Computer Science and Applications*, vol. 11–12, 2020.

[AK15] Awad, M.; Khanna, R. “Support Vector Regression”. Berkeley, CA: Apress, 2015, pp. 67–80.

[BSH+05] Bailenson, J.; Swinth, K.; Hoyt, C.; Persky, S.; Dimov, A.; Blascovich, J. “The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments”, *Presence*, vol. 14, 08 2005, pp. 379–393.

[CKS+20] Choudhary, Z.; Kim, K.; Schubert, R.; Bruder, G.; Welch, G. “Virtual big heads: Analysis of human perception and comfort of head scales in social virtual reality”, 2020, pp. 425–433.

[CLB+21] Chen, L.; Li, S.; Bai, Q.; Yang, J.; Jiang, S.; Miao, Y. “Review of image classification algorithms based on convolutional neural networks”, *Remote Sensing*, vol. 13–22, 2021.

[DMdAAM22] Dal Molin, G. P.; de Andrade Araujo., V. F.; Musse., S. R. “Estimating perceived comfort in virtual humans based on spatial and spectral entropy”. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP, 2022, pp. 436–443.

[DMND+21] Dal Molin, G. P.; Nomura, F. M.; Dalmoro, B. M.; de A. Araújo, V. F.; Musse, S. R. “Can we estimate the perceived comfort of virtual human faces using visual cues?” In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 366–369.

[DT05] Dalal, N.; Triggs, B. “Histograms of oriented gradients for human detection”. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 886–893 vol. 1.

[Han06] Hanson, D. “Exploring the aesthetic range for humanoid robots”, 01 2006.

[HSN23] Hangaragi, S.; Singh, T.; N, N. “Face detection and recognition using face mesh and deep neural network”, *Procedia Computer Science*, vol. 218, 2023, pp. 741–749.

[Hu62] Hu, M.-K. “Visual pattern recognition by moment invariants”, *IRE Transactions on Information Theory*, vol. 8–2, 1962, pp. 179–187.

[ILU23] Imaizumi, T.; Li, L.; Ueda, K. “Does machine learning replicate the uncanny valley? an example using facenet”, *eScholarship*, 2023.

[KMT17] Kätsyri, J.; Mäkäräinen, M.; Takala, T. “Testing the ’uncanny valley’ hypothesis in semirealistic computer-animated film characters: An empirical evaluation of natural film stimuli”, *International Journal of Human-Computer Studies*, vol. 97, 2017, pp. 149–161.

[LKA+17] Laine, S.; Karras, T.; Aila, T.; Herva, A.; Saito, S.; Yu, R.; Li, H.; Lehtinen, J. “Production-level facial performance capture using deep convolutional neural networks”. In: Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation, 2017.

[LLHB14] Liu, L.; Liu, B.; Huang, H.; Bovik, A. “No-reference image quality assessment based on spatial and spectral entropies”, *Signal Processing Image Communication*, vol. 29, 09 2014.

[MCHP10] MacDorman, K. F.; Coram, J. A.; Ho, C.-C.; Patel, H. “Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas”, *Presence: Teleoper. Virtual Environ.*, vol. 19–3, jun 2010, pp. 213–229.

[MI06] MacDorman, K. F.; Ishiguro, H. “Toward social mechanisms of android science”, *Interaction Studies*, vol. 7–2, 2006, pp. 289–296.

[Mor70] Mori, M. “Bukimi no tani [the uncanny valley].”, *Energy*, vol. 7, 1970, pp. 33.

[PR14] Prakash, A.; Rogers, W. “Why some humanoid faces are perceived more positively than others: Effects of human-likeness and task”, *International Journal of Social Robotics*, vol. 7, 12 2014.

[TCY20] Tu, Y.-C.; Chien, S.-E.; Yeh, S.-L. “Age-related differences in the uncanny valley effect”, *Gerontology*, vol. 66–4, 2020, pp. 382–392, epub 2020 Jun 11.

[TF01] Tumblin, J.; Ferwerda, J. “Applied perception”, *IEEE Computer Graphics and Applications*, vol. 21–5, 2001, pp. 20–21.

[TGANW11] Tinwell, A.; Grimshaw-Aagaard, M.; Nabi, D.; Williams, A. “Facial expression of emotion and perception of the uncanny valley in virtual characters”, *Computers in Human Behavior*, vol. 27, 03 2011, pp. 741–749.

[VJ01] Viola, P.; Jones, M. “Rapid object detection using a boosted cascade of simple features.”, *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*, vol. 1, 2001, pp. I–I.

[YUY21] Yorgancigil, E.; Urgan, B. A.; Yildirim, F. “Uncanny valley effect is amplified with multimodal stimuli and varies across ages”, oct 2021.

[ZLD18] Zhang, M.; Li, W.; Du, Q. “Diverse region-based cnn for hyperspectral image classification”, *IEEE Transactions on Image Processing*, vol. 27–6, 2018, pp. 2623–2634.

[ZZM19] Zell, E.; Zibrek, K.; McDonnell, R. “Perception of virtual characters”. In: ACM SIGGRAPH 2019 Courses, 2019.

[R10] Žunić, J.; Hirota, K.; Rosin, P. L. “A hu moment invariant as a shape circularity measure”, *Pattern Recognition*, vol. 43–1, 2010, pp. 47–57.