

Plan Overview

A Data Management Plan created using DMPTool

Title: Uso de técnicas computacionais para identificar traços de personalidade por meio de imagens e/ou vídeos 2D e/ou 3D

Creator: Alan Goncalves

Affiliation: State University of Campinas (unicamp.br)

Principal Investigator: Alan Gonçalves

Project Administrator: Marco Antonio Garcia de Carvalho

Template: UNICAMP-GENERIC: Aplicável a todas as áreas

Project abstract:

Este plano visa apresentar uma proposta de gestão de dados para este projeto que tem como objetivo a aplicação do uso de técnicas computacionais, como Visão Computacional, Machine Learning e Deep Learning, para a análise de traços de personalidade usando imagens 2D e/ou 3D da face e do corpo humano.

A análise da personalidade tem sido muito estudada nas áreas de Psicologia, Neuropsicologia e Processamento de Sinais. Os psicólogos há muito tempo pesquisam a personalidade humana e ao longo dos anos, diferentes teorias foram propostas sobre esse assunto. Além disso, os pesquisadores entendem que estes traços são compostos por emoções, comportamentos e padrões relacionados ao pensamento e eles acreditam que existam cinco **traços de personalidade** essenciais.

Em geral, a psicologia descreve estes traços de personalidade como 5 áreas, também conhecidas como as "**Cinco dimensões básicas da personalidade**", ou os "**Big Five**". Os cinco grandes traços de personalidade descritos pela teoria são: *Abertura, Consciência, Extroversão, Amabilidade e Neuroticismo*. Esta teoria ganhou força e representatividade a partir de 1949 por seu criador Donald W. Fiske e foi expandida posteriormente por outros pesquisadores, entre eles Norman e

Goldberg.

Entretanto, com o surgimento de novas técnicas oriundas a partir do uso da computação visual, essa nova área de pesquisa tornou-se bastante atrativa para os pesquisadores.

Finalmente, sabemos que o texto e a fala predominam nos trabalhos de pesquisa, porém a análise da personalidade por meio de dados visuais como imagens e vídeos, abriu novas oportunidades para os pesquisadores, sendo assim o objetivo deste trabalho de pesquisa.

Start date: 02-22-2022

End date: 06-27-2023

Last modified: 06-22-2022

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Uso de técnicas computacionais para identificar traços de personalidade por meio de imagens e/ou vídeos 2D e/ou 3D

Descrição dos Dados e Metadados

Quais serão os dados coletados?

O conjunto de dados utilizados serão do tipo Imagens, Vídeos e Áudios e compreende cerca de 10.000 cliques (duração média de 15s) extraídos de mais de 3.000 vídeos de alta definição (HD) do YouTube de pessoas olhando e falando em inglês para uma câmera. Os vídeos são divididos em conjuntos de treinamento, validação e teste com uma proporção de 3:1:1. As pessoas nos vídeos mostram diferentes gêneros, idades, nacionalidades e etnias.

Que metadados serão anotados e qual padrão será seguido?

Os vídeos são rotulados com variáveis de traços de personalidade. O Amazon Mechanical Turk (AMT) foi usado para gerar os rótulos e foi utilizado um procedimento baseado em princípios e para garantir a confiabilidade dos rótulos.

Os traços de personalidade considerados foram os do Modelo dos Cinco Fatores (também conhecido como Big Five), que é o paradigma dominante na pesquisa de personalidade. Ele modela a personalidade humana em cinco dimensões: Extroversão, Amabilidade, Conscienciosidade, Neuroticismo e Abertura.

Assim, cada clipe tem rótulos para esses cinco traços representados com um valor.

Em relação as transcrições, todas as palavras dos videoclipes foram transcritas pelo serviço profissional de transcrição. No total, foram transcritas 435984 palavras (183861 non-stopwords), o que corresponde a 43 palavras por vídeo em média (18 non-stopwords). Entre essas palavras, 14535 eram únicas (14386 non-stopwords).

Os arquivos utilizam o formato **roundtruth**, onde as anotações e transcrições são armazenadas em dicionários. Existe um arquivo para anotações e um arquivo para transcrições por fase. Cada vídeo tem uma transcrição (se não houver nada para transcrever em um vídeo, sua transcrição correspondente será uma string vazia). Cada transcrição é um objeto unicode. O arquivo de transcrição é um único dicionário. Ou seja, suas chaves são os nomes dos vídeos, e seus valores são as transcrições correspondentes. Por exemplo: `transcription['a_video_name']` nomeará as transcrições do vídeo como 'a_video_name'.

Aspectos Legais e Facilidade de Acesso aos Dados

Quais são as questões legais e éticas associadas aos dados e relevantes a este projeto?

Não existem questões legais e éticas de relevância envolvidas no presente projeto, ou seja, os dados não estão relacionados a espécies em extinção ou outros que requeiram cuidados especiais, pois o tema investigado não se trata de nenhum segredo comercial ou de patente. Trata de dados de seres humanos que por vontade própria se disponibilizaram a realizar as gravações dos vídeos. Sendo assim, não se fez necessário consultas a comissões de ética em pesquisa ou órgãos de qualquer outra natureza.

Quais são as políticas a serem utilizadas para o compartilhamento de dados?

Não existem restrições ou questões éticas relevantes que devam ser consideradas para o compartilhamento dos dados, ressalvando-se apenas o interesse particular dos colaboradores do projeto no ato da decisão do compartilhamento.

Todos os dados envolvidos podem ser compartilhados sem nenhuma restrição, desde que dentro dos grupos

de pesquisas que estão colaborando no âmbito das análises do projeto como alunos de Mestrado e Doutorado da Faculdade de Tecnologia da Unicamp e da Universidade de Calgary no Canada, desde que com prévia comunicação dos pesquisadores responsáveis.

Os dados eventualmente compartilhados serão armazenados no repositório Zenodo e quaisquer referências em literatura devem respeitar o crédito do autor do dado ou resultado publicado.

Gestão de Dados e Armazenamento

Em que formatos serão armazenados os arquivos resultantes da pesquisa em questão? Que software poderá ser utilizado para a manipulação de cada um dos formatos listados?

Os arquivos serão armazenados em:

1. **DOC/DOCX/PDF:** para arquivos do tipo Textos com imagens e tabelas em formato independente de plataforma computacional;

- **Software(s) utilizado(s):** Word e Adobe Acrobat Reader

2. **XLS/XLSL:** para planilhas eletrônicas;

- **Software(s) utilizado(s):** Excel.

3. JPEG/TIFF/PNG: para imagens;

- Software(s) utilizado(s): Paint 3D, Photoshop, Corel, Adobe Premier

4. MP4/WMV/OGV: Vídeo coletado por câmeras, telefones celulares, etc.

- Software(s) utilizado(s): Windows Media Player

**Como e onde estes arquivos serão mantidos? Por quanto tempo ocorrerá esta preservação?
Como será realizado o backup destes dados?**

Os arquivos serão mantidos no repositório Zenodo, pelo período de dois anos, com o título: Personality Trait Storage.

Os Metadados e identificadores persistentes no Zenodo são armazenados em uma instância PostgreSQL operada na infraestrutura Database on Demand do CERN com ciclo de backup de 12 horas com um backup enviado para armazenamento em fita uma vez por semana, ou seja, este repositório possui sistema de backup automático.

Os dados serão organizados e classificados de acordo com a fase correspondente do projeto de pesquisa. O acesso será preservado a partir de um DOI - Digital Object Identifier.
