📰 Detecting faces in historical newspapers

This project aims to detect faces in historical newspapers through adoption machine learning methodologies. The main goal is to analyze the occurance and frequency of human faces, grouped by newspaper page, in historical newspaper archives from three historic swiss newspapers: Journal de Genève (JDG, 1826-1994); the Gazette de Lausanne (GDL, 1804-1991); and the Impartial (IMP, 1881-2017). The intent behind the project is providing insights into societal trends through computational analysis of cultural data, aligning with digital humanities methodology.

📈 Data

📋 Dataset

This project utilizes the historical newspaper dataset stemming from paper 'Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers' by Barman et al., (2021). The dataset's individual sub folders, each containing the images for seperate newspapers, should be placed in the `in` directory. The dataset can be sourced from here. You can download the dataset directly here. The script only looks at .jpg files, so you can leave any other files.

🤖 Model

The face detection model used in this project is `facenet-pytorch`, a PyTorch implementation of the deep learning CNN FaceNet face recognition model. This model is capable of detecting faces in images.

The MTCNN (Multi-task Cascaded Convolutional Networks) is initialized for performing face detection. Additionally, embeddings could be extracted using the InceptionResnetV1 model. Embeddings are not utilized in this project.

🗁 Project structure

```
└── face_detection_historical_newspapers
    ├── in/
    |    ├── GDL/
    |    ├── IMP/
    |    └── JDG/
    |
    ├── out/
    |    ├── csv_results/
    |    └── plots/
    |
    ├── src/
    |    ├── base_utilities.py
    |    ├── code_utilities.py
    |    ├── data_processing_utilities.py
    |    ├── face_detection.py
    |    └── plotting_utilities.py
    |
    ├── README.md
    ├── requirements.txt
    ├── setup_unix.sh
    └── setup_win.sh
```

⚙️ Setup

To set up the project, you need to create a virtual environment and install the required packages. You can do this by running the appropriate setup script for your operating system.

Dependencies

Please ensure you have the following dependencies installed on your system:

- **Python**: `version 3.12.3`

💾 Installation

1. Clone the repository

```
git clone https://github.com/apathriel/cds-vis-analytics
```

2. Navigate to the project directory

```
cd assignments
cd face_detection_historical_newspapers
```

3. Run the setup script to install dependencies, depending on OS.

```
bash setup_unix.sh
```

4. Activate virtual environment (OS-specific) and run main py scripts.

```
source env/bin/activate
python src/face_detection.py
```

🚀 Usage

After setting up the project, you can run the main script (`src/face_detection.py`) to start the face detection process. The results will be stored in the out/ directory. By default, the script visualizes the results using pyplot. Alternatively, the results can be preloaded using the `preload_and_visualize_results()` function. Examples of an interactive visualization is implemented in plotly.

📊 Results

When examining the resulting visualizations, a clear trend emerges: There's a clear correlation between decade and percentage of pages containing faces. This might be due to technological developments, or other shifts in media culture. The results presented by the methodology clearly conveys a societal change, as explored through cultural data. The approach allows for a computational exploration of the historically salient question:

> How has the prevelance of images of human faces changed in print media over the last roughly 200 years? Are there any significant differences and what might this mean?

The methodology provides a robust, reproducable answer to this question. It allows us to quantify and visualize the presence of human faces in the newspapers, contributing to our understanding of culture, examined through understanding print media as a cultural medium.

One of the challenges in this project was the idea of evaluating the performance of the face detection model. Unlike labeled supervised machine learning, there's no ground truth on which to evaluate the model performance. Since it might unlikely to find a model specically trained on historical newspaper data, you could potentially incorporate transfer learning to fine-tune a pre-trained model, or use it for feature extraction. You'd have to have a manually annotated dataset from which a custom model could be trained on. Another possibility might be incorporating an additional model to evaluate embeddings for the boxes returned by the InceptionResnetV1 model (If the FaceNt model doesn't already incorporate this).

While this computational approach provides valuable insights into the prevalence of human faces in historical newspapers, it's important to acknowledge its limitations:

1. The quality of the dataset images can vary, impacting the accuracy of the face detection, leading to inconclusive results.
2. Bias in training data and/or dataset. The data on which the FaceNet model was trained on might contain biases, which might impact its application to the target image data.
3. Imbalance in dataset. The dataset largely contains more samples of newer images. Ideally you'd have a larger dataset, which would theoretically lead to accurate results.
4. Interpreting the results requires historical context. Ideally, the method would be coupled with traditional humanities methodologies to contextualize the findings.

Despite limitations, this computational approach offers a novel way to explore cultural trends and changes over time using large-scale image data that would be impossible for humans to manually process. The symbiosis between humanities and computation provides a unique avenue for making sense of human culture.

📖 References

- Medium article on facenet-pytorch
- Barman et al. - Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers (2021)
- Datasets and Models for Historical Newspaper Article Segmentation