Analyzing Emotional Expression in Micro-video Through Semantic Segmentation

Students: Sean Chen, Nicholas Haisler, Marshall Tentis (Drake Univesity)

Supervised by: Alimoor Reza and Dian Nostikasary (Drake Univesity), Jane Rongerude (Iowa State University)

Motivation

In this research, we explore how digital spaces reflect the ways young people use public parks and recreation areas. Oftentimes, youth express themselves more in digital spaces than in public or to those around them. By understanding this phenomenon, we hope to gain insight into their emotional struggles, allowing for better methods of addressing these issues. This could help the city and civic organizations develop effective response systems for youth. We aim to tackle this problem using a combination of computer vision and statistical analysis techniques across two groups: non-vulnerable and vulnerable.

Visual Analytics from Computer Vision Models

We aimed to analyze the visual cues representing the emotional state of a user from micro-video data. We address several image understanding tasks using existing pre-trained computer vision models for each video.

- **Task#1: Semantic categories recognition**
  - Collect the types of objects present in the video, such as:
    - **Outdoor categories:** building, tree, bench, dog, truck, fence, etc.
    - **Indoor categories:** laptop, chair, table, mug, book, etc.
  - Utilized the Mask2Former model [1].
- **Task#2: Scene recognition from images**
  - Identify the overall scene depicted in the image, including:
    - Baseball fields, parks, lagoons, museums, parking, tennis courts, lawns, etc.
- **Task#3: Weather recognition from images**
  - Classify weather conditions, such as:
    - Rain, snow, or sunny.
  - Enable season classification through weather patterns.
- **Task#4: Illumination conditions**
  - Detect characteristics such as:
    - Time of day, brightness, dawn, or sunset.
- **Task#5: User emotion recognition from an image**
  - Recognize user emotions using methods like:
    - Ekman's 7 emotions: Happiness, Sadness, Surprise, Anger, Disgust, Fear, and Contempt.[2]
- **Task#6: Skin-type classification**
  - Classify skin color using the Fitzpatrick Scale.

Figure: Masks from Mask2Former

Figure: Fitzpatrick Skin color Scale

Micro-video (TikTok) Scrapper

This is a tool we used to collect micro-videos from the TikTok platform. It is an unofficial API developed by users on GitHub. Using this tool, we scraped TikTok for specific tags and created a spreadsheet with links to the corresponding videos. It's important to note that users self-annotate their videos with tags. In this work, we focus on analyzing two emotional states of the users: **happy** and **sad.**

Dataset	Url	Tags
Sad	https://www.tiktok.com/...	couch, icecream
Happy	https://www.tiktok.com/...	birthday, friends
Happy	https://www.tiktok.com/...	dog, food,festival
Sad	https://www.tiktok.com/...	cry,rain
...	...	...

Table: A table with two columns and five rows.

Datasets

We created two datasets with videos corresponding to **Happy** or **Sad**. Each dataset contains approximately 50 videos scraped from TikTok. These were gathered using a scraper to find videos with specific tags, such as friends or beach for the happy category, and pain or lost for the sad category. After collecting the videos, we narrowed them down to two sets of approximately 50.

Figure: Sample Of Happy Videos

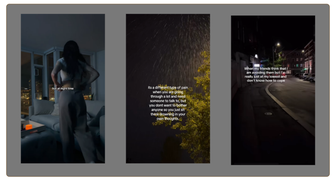

Figure: Sample Of Sad Videos

Semantic Segmentation Model: Mask2Former

Mask2Former is a model released by meta for Universal Image Segmentation. The model excels at segmenting images across a wide range of object which made it an extremely important tool for our work.

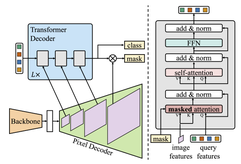

Figure: Mask2Former Architecture

Semantic Segmentation Experiment

Depending on the length and number of videos, the time to compute can take hours to days. Using the previously mentioned Mask2Former we segmented our videos using a 3-step pipeline. We used a single GPU to compute our segments over several days.

Split into Frames

- Cycle through the folder of videos and split each video into frames assuming 30 frames per second. Making a 10 second video for example have 300 frames.

Segment Each Frame

- Pass each of the frames collected from the videos through Mask2Former. This will output a colored mask, along with a .txt file with the color to label mapping.

Analyze Segmentation

- Once we have the segments, we can find the objects in the videos across each resulting dataset by scanning through each of the .txt files to identify objects found.

Results

Below our two graphs of the top 25 images found in the two datasets. There are some obvious similarties such as person and wall, but even here some differences can be spotted. Happy has certain objects such as trees and dogs being seen more, whereas Sad has pillow, tv, and bed.

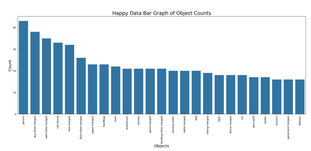

Figure: Sample Of Happy Videos

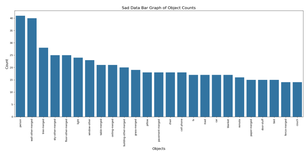

Figure: Sample Of Sad Videos

This next graph shows the difference more clearly with positive blue meaning an object showed up more times in happy and vice versa for the negative red. Here lots of outdoor bojects can be found showing up more in Happy than Sad.

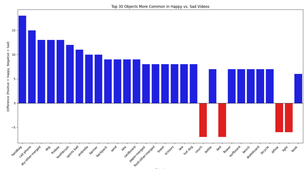

Figure: Difference of object in Happy and Sad

References

[1] Masked-attention Mask Transformer for Universal Image Segmentation, Bowen Cheng and Ishan Misra and Alexander G. Schwing and Alexander Kirillov and Rohit Girdhar, CVPR 2022

[2] Basic emotions, Ekman, Paul and Dalgleish, Tim and Power, M, San Francisco, USA,1999