Manually looking for criminals in CCTV footage has traditionally been an extremely tedious and lengthy process, motivating us to create a system of machine learning models to aid in that process. We trained a vision and text transformer together to output aligned embeddings of an image and a text description of a person. We started by using the COCO dataset, an image dataset primarily used for object recognition. We then extracted the images of people, using YOLOv8, a computer vision model. After we had a new dataset of images of people, we labeled each image with a description of the person in the image, using OpenAi’s 4o vision API.

When dealing with crime, especially with governments and businesses, one major way they can catch or identify perpetrators is CCTV footage, which can record the crime, or people related to the crime. . However, a core issue with this footage is the vast amount of it that exists. Traditionally, many hours of footage need to be manually combed over to identify when a particular suspect entered the frame of the camera. Our goal is to use just a text description of a criminal, which can be provided by a potential witness, to find that criminal in CCTV footage. With this technology aiding CCTV footage users, this process can be significantly quickened and the search scope can expand to find criminals that otherwise might not have been caught.

A more common CCTV related search task is using a known image of a criminal to find them in CCTV footage. Previous work by Tagore et al. found success using several variants of CNNs for this task. Sheshkal et al. found more success with transfer learning based siamese network CNNs. In terms of text-based matching, approaches from Niu et. al focus on both global and local scales, while Ma et. al created a combined CNN-text model to produce similarity scores.

We wanted to train text-vision transformer models to take in an input of a text description, along with many images of people, and output embeddings for each text and each image. However, in order to do this we needed a dataset of images of people and their text descriptions, and we could not find any such datasets online. As a result, we decided to synthesize our own dataset based off of the COCO dataset, which is known to include many images with people present in them.

We ran a pre-trained YOLOv8 large model on each image in the dataset to extract each individual person from the frame. We mainly wanted full body images of each person, so we filtered out people with segmentation masks less than 3% the area of the total frame. Additionally, we used a 90% prediction confidence threshold to ensure non-humans didn’t enter the dataset. Finally, we manually went through each image in the dataset (~1000) and removed the few with non-ideal conditions (ex: not full body, too dark to analyze, etc).

We ensured that all background information around each person's image was removed using the precise segmentation masks to ensure that only information about the person and not the background setting was present in the image. That helped to make sure that the vision model didn’t incorrectly use something unrelated for identification of the person. Then, each image was fed through OpenAI's chatGPT 4o vision API to generate a text description of each person. The following prompt was used: "Describe the person in the image in a paragraph. Include features such as build, skin tone, age, hair color and style, eye color, facial features, any notable physical characteristics. Avoid mentioning facial expressions or pose." Many of the API calls resulted in false content violation errors and had to be re-run to obtain text descriptions.

YOLOv8 generates segmentation masks of objects it’s been trained to identify in a frame. At its core, it mainly utilizes a CNN architecture and the majority of the predictive process lies in predicting the coordinates of the edges of bounding boxes surrounding targets at a highly localized scale. Chatgpt 4o vision is a vision transformer-based model that encodes patches of images into embeddings and utilizes multi-headed attention to output appropriate text.

Using our synthesized dataset of text and image pairs, we used contrastive learning to train a text and vision transformer model. We used transfer learning, starting from CLIP (an OpenAI model), and trained the models to generate closer embeddings between matching pairs, and further away embeddings from non-matching pairs. An 80-20 train-test split was used, to ensure the amount of test cases was significant enough to test the model on. We then used cosine similarity to determine the distance from one embedding to another, and a 512-dimensional space was used. Each epoch took a while to run due to the size of the models (~190m parameters), so we were only able to train for 3 epochs.

The graph below shows how accurately these models were able to find the correct image using a text description. Among all the matches the model predicted, the percentile represents where the actual match lies. Before fine-tuning the CLIP model, it showed little predictive power with a median prediction rank of 38th percentile. After fine-tuning the model (figure 1), we achieved a median prediction rank of 6.5th percentile, with a recall-1, recall-5, and recall-10 of 18%, 38%, and 54% respectively.

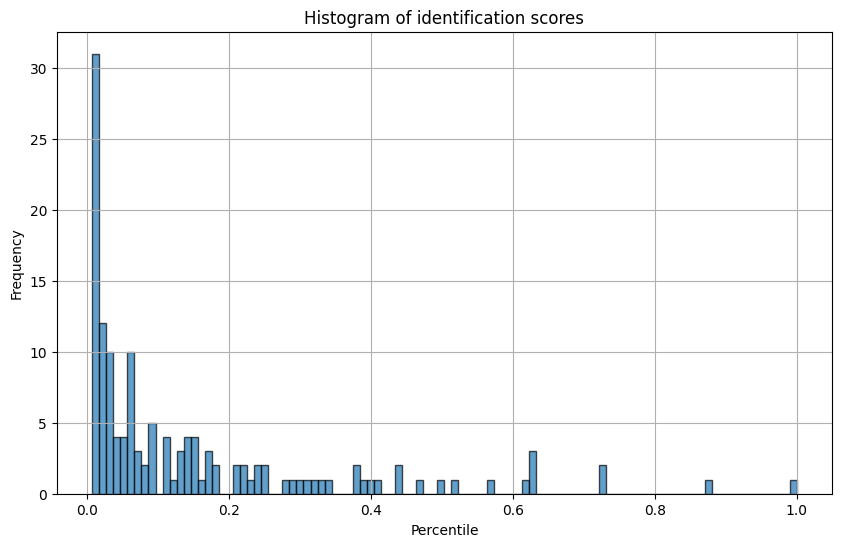


Figure 1: Identification scores of each input in the testing set. The percentile

represents where the actual match lies among all the predicted matches.

While our models were able to achieve decent predictive performance somewhat consistently, they fail to fully identify criminals based on a text description all by itself. As a result, our work should instead be used as a tool to aid people with searching through CCTV footage to quicken the process and expand the scope of the search. Our tools can help CCTV tape reviewers skip past times when no humans pass through the video frame, and can sort identified people to help them find the perpetrator without wasting time on a large portion of humans captured in the video. In the future, we would lock layers on the transformers and opt for smaller models to be able to train them even further.

References:

1. Tagore, A., & Gupta, R. (2020). *Hierarchical re-identification using Siamese network and color histograms*. Retrieved from<https://arxiv.org/abs/2005.03293>
2. Niu, M., He, W., Zhang, L., & Wang, P. (2019). *Multi-granularity image-text alignment for text-based person search*. Retrieved from<https://arxiv.org/abs/1906.09610>
3. Ma, X., Luo, H., Chang, Y., Wang, Y., & Wang, Z. (2020). *Dual-path CNN with Max Gated block for text-based person search*. Retrieved from<https://arxiv.org/abs/2009.09343>
4. Amouei Sheshkal, M., Karimian, S., & Seyyedsalehi, S. A. (2020). *Improved person re-identification using lightweight CNN and transfer learning in Siamese networks*. Retrieved from<https://arxiv.org/abs/2008.09448>