# Annotation System For Aiding Automatic Face Detectors

Ethan Wilson
University of Florida

Jenny Skytta
Seattle Children's Hospital and Research Institute

Frederick Shic
Seattle Children's Research Institute

Eakta Jain
University of Florida

## 1 Introduction

Marking the faces in video segments is a complicated task. Multiple solutions have been proposed and implemented over the years: manually marking every face in every frame, manual markings stitched together by interpolating, or more recently, automatic face detection systems. Automatic face detectors are not perfectly accurate, and manual methods are both tedious and time consuming. We propose, implement and demonstrate an annotation system that combines automatic face detection with manual annotation to quickly achieves a total coverage of faces for a single subject of interest in a video clip.

Methods to annotate faces in videos originated in the computer vision and image processing community. Publicly available face detection databases needed to mark bounding boxes around face regions in videos before releasing the datasets in order to provide ground truth. Video processing software additionally used manual markings supplemented by interpolation for use cases such as blurring a face in a documentary to protect the witness. Marking out faces can be a quite tedious task when the videos are long. For example, one minute of 30fps video would have 1800 frames to be marked. If the subject moves around instead of staying in one spot, the annotator's work goes up substantially.

Our project, which consists around facial privacy protection in a clinical setting, created a need to annotate faces at a scale that has not been done previously in the computer vision community. For reference, the Large-scale CelebFaces Attributes Dataset (CelebA) [1] is one of the largest annotated face datasets with around 200,000 images. Our dataset consists of 50 video observation sessions, where a single session has approximately 225,000 frames. Annotating all frames in our dataset is a task over 50 times the size of CelebA.

We hypothesized that it would be a less time and human-labor-intensive approach to run an automatic face detection method as a first pass, and then have

an annotator correct/clean the detected faces reported by this method. However, in practice, even if automatic detection marked 95%, manual annotation was not feasible: if we estimate an annotator marking one frame per second and ignore the transition times between frames, marking the final five percent of frames would still take the annotator two hours per observation session.

## 2 Method

We design a four-pass annotation system that closely intertwines algorithmic results with an annotator's markings. Our key insight is to incorporate **key frames** that allow the annotator to very quickly pinpoint areas of interest that may need manual correction.

In the *first pass*, the annotator scrubs through the entire video clip, marking simple start and end flags to indicate the frames where the key subject's face enters or leaves the frame. If there are multiple subjects in a video clip, the annotator ignores all others and only annotates according to the key subject. These flags are used to compute 'green' regions of interest, along with 'red' regions that can be ignored as they don't contain the main subject's face. Additionally, by the time of the first pass automatic face detection has been run. The bounding boxes are linked spatially across time by looking at nearby positions in nearby frames to create face chains.

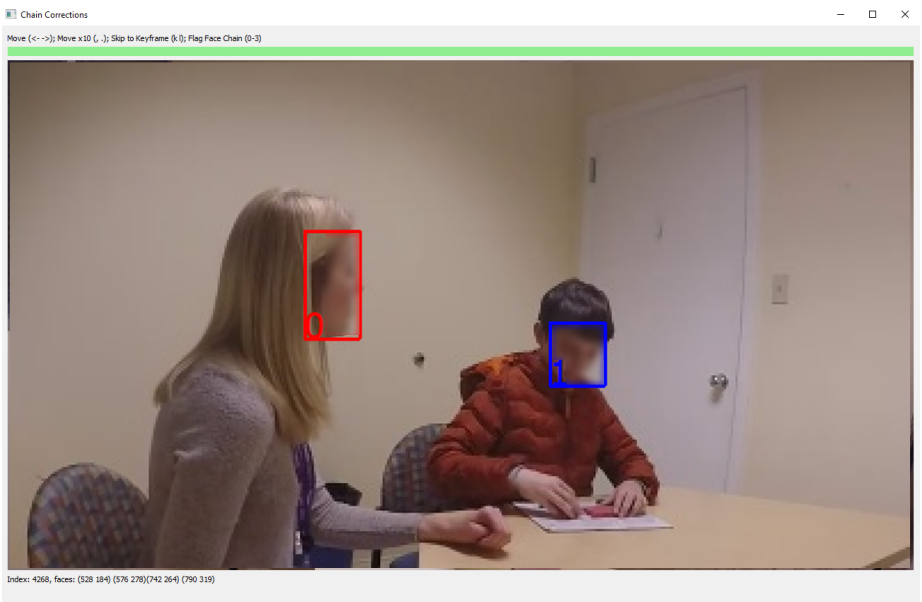


Figure 1: A screenshot of the GUI used for passes two and three. The top bar indicates being in a green region, and faces are tagged such that blue corresponds to the main subject. The image was zoomed in for visualization purposes.

In the *second pass*, face chains are presented with associated numeric labels. The annotator can mark face chains as belonging to the key subject or not. All chains are assumed to be the key subject initially, and are represented in blue. Once the annotator has toggled a chain, all faces belonging to that chain are represented in red. To speed up the process, the annotator can jump between **key frames**. Key frames are made up of the start and ending frame of each face chain, and the start and ending frame of every 'green' region. The annotator can focus on the key frames, ensuring they see every face chain while drastically reducing the amount of frames they need to view.

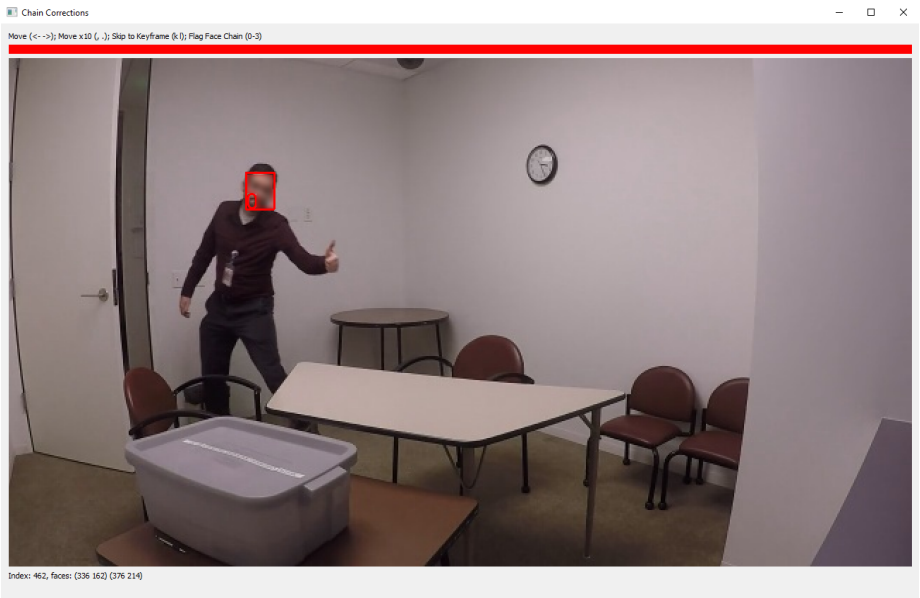


Figure 2: A screenshot of the GUI during a red zone when the main subject is not in frame.

The *third pass* utilizes the same GUI as the second and can be performed at the same time, but is considered separately as initial annotators chose to perform the passes separately for simplicity. In the third pass, the annotator can optionally create new face chains for the child. The annotator again scrubs through the video by key frames. If a key frame does not contain a face, the annotator marks it and nearby regions manually, thus catching face regions that were missed by the automatic face detection method. The annotator also keeps an eye out for segments containing a lot of movement and creates new face chains accordingly to improve the accuracy of the interpolation in the fourth pass.

The *fourth pass* is an automated interpolation to fill in gaps between face chains. Any frame within a green region that is incomplete computes a new bounding box by using information from the nearest (by frames) faces by inter-

polating their positions and sizes. These interpolations are tight enough from the annotated faces added during the third pass to where a consistent bounding box accurately follows the patient's face at any point within the green regions. The bounding boxes in the fourth pass are malleable and quickly computed at render time. This allows the annotator to go back and correct any mistakes in earlier steps without worrying about the fourth pass saving over previous work.

# 3 Results

This annotation system has been incorporated into our greater workflow. For the videos we have fully processed and annotated, we are able to see a noticeable improvement in the detection rates of the main subject after annotation. These detection statistics can be considered the ground truth, as every frame that contains the main subject is now marked out. With this information, we are able to compare our annotations to the starting automatic detection data to see the error of automatic detection. Table 1 shows results for three of our ADOS sessions. We can see that the error for the automatic detection varies greatly depending on individual session, but is a highly significant value.

| Session | Automatic Detection | Detection with Annotations | Ground Truth Error |
|---|---|---|---|
| A | 89.79% | 98.50% | 8.84% |
| B | 63.75% | 93.26% | 31.64% |
| C | 96.49% | 98.94% | 2.48% |

Table 1: Percentage of video frames containing the subject's face. Results from the S3FD automatic detection algorithm and after completing the manual annotation pipeline. The final column measures the error the automatic detection produced when compared against the final annotated data, which has masked every face and can be considered as a ground truth.

# 4 Discussion

Our four-pass method will be used for the processing of recorded clinical ADOS sessions. It is expected to take 20-40 minutes per session, depending on the length of the session and the quality of the automatic detection results that are being used as a starting point. As mentioned in the previous manuscript, there are two main layouts in these sessions. For layout A, where the camera has a clear view of the participant, annotation should be a straight-forward process, mainly only needing to clean up portions where the subject enters/exits. For videos in the less ideal layout B, annotating could be much more involved. If the subject is at a distance where automatic detection fails, many more face

chains will need to be created and annotation times could drastically increase. As a result, we may need to cull the stimuli set such that we only use the more accurately detected videos from layout A.

# References

[1] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015).