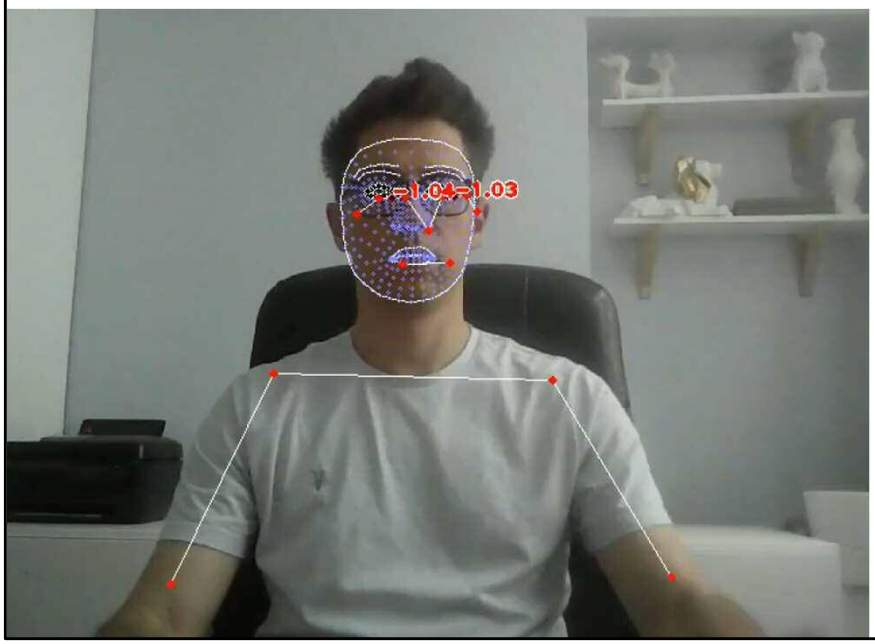
PROTECT Researcher Symposium 5-6 May 2022

PROTECT
A COVID-19 National Core Study

AI-assisted face-touching behavior observation

Tom Komar, Urban Observatory at Newcastle University

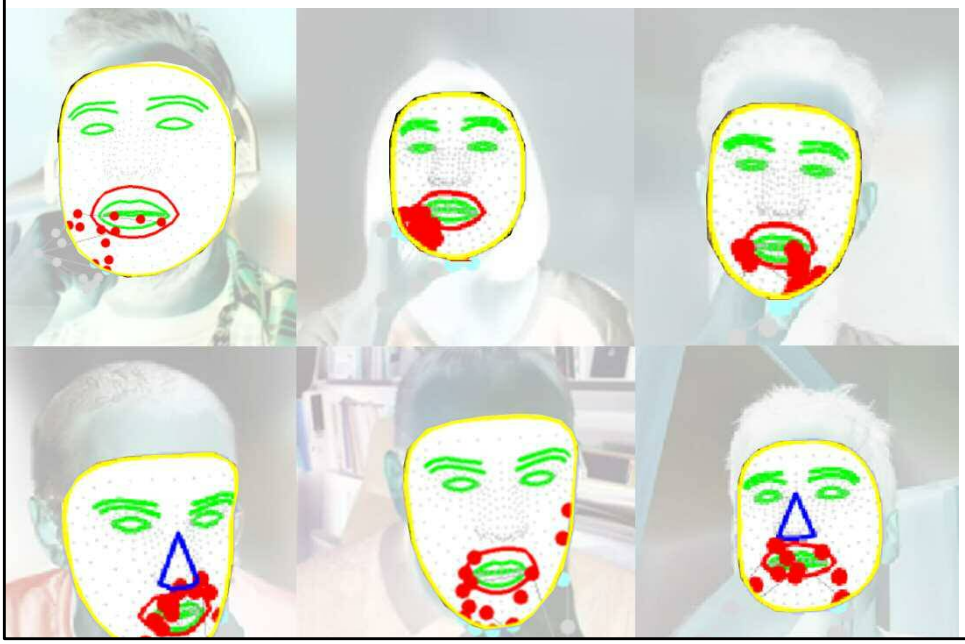


I will tell you about some aspects of AI-assisted observation of human self-face touching behavior. Normally such observations are done manually where people watch recordings or visit a target environment and observe some individuals. That is a time consuming method of data capture and even with a pool of human resources it will not scale well. There are just no simple shortcuts like what we have available in computer science – ability to effortlessly analyse behavior of multiple subjects at once, improving input or code efficiency, enabling more processing power and making the system work 24/7.

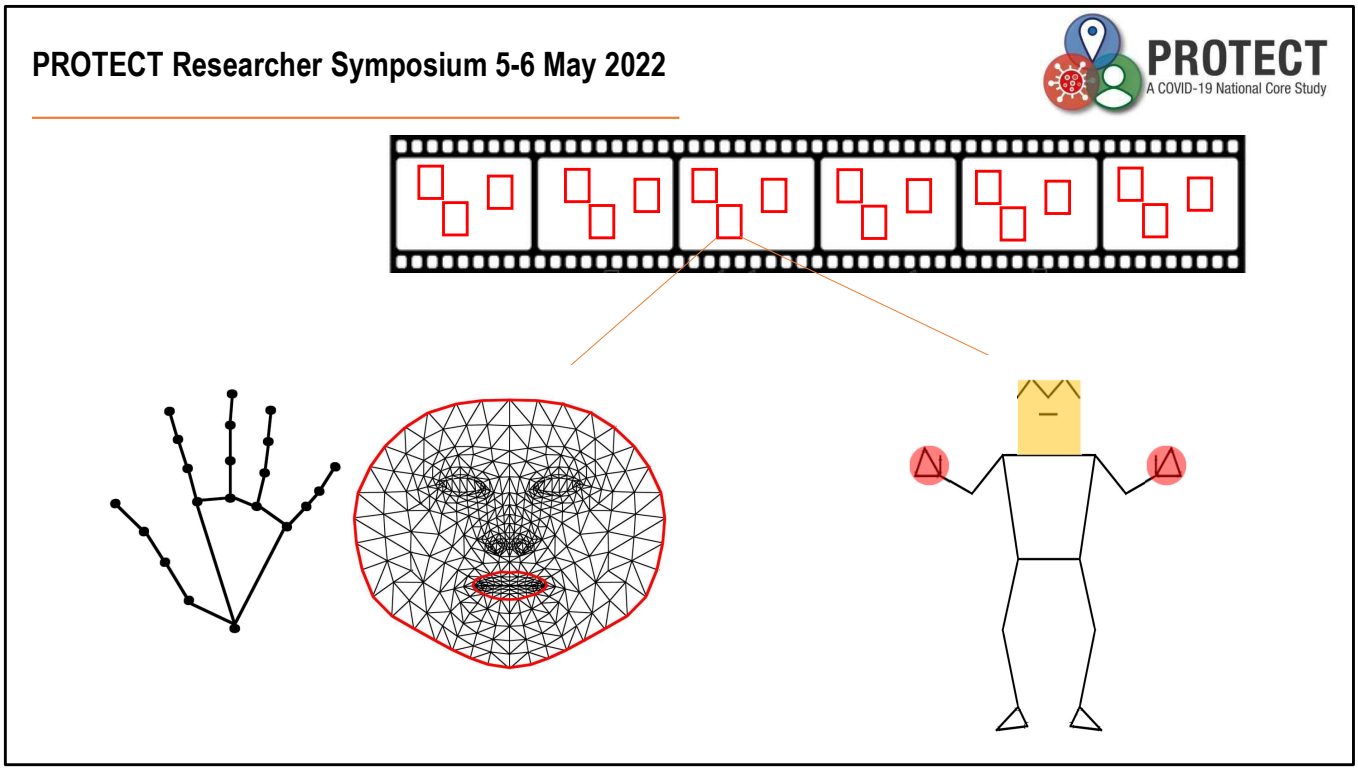
The models we used this time were body, face and hand keypoint detection. Effectively, it allowed us to capture moments when one's hand appears in close proximity to the face. Let me share with you few lessons from using that technology.

PROTECT Researcher Symposium 5-6 May 2022

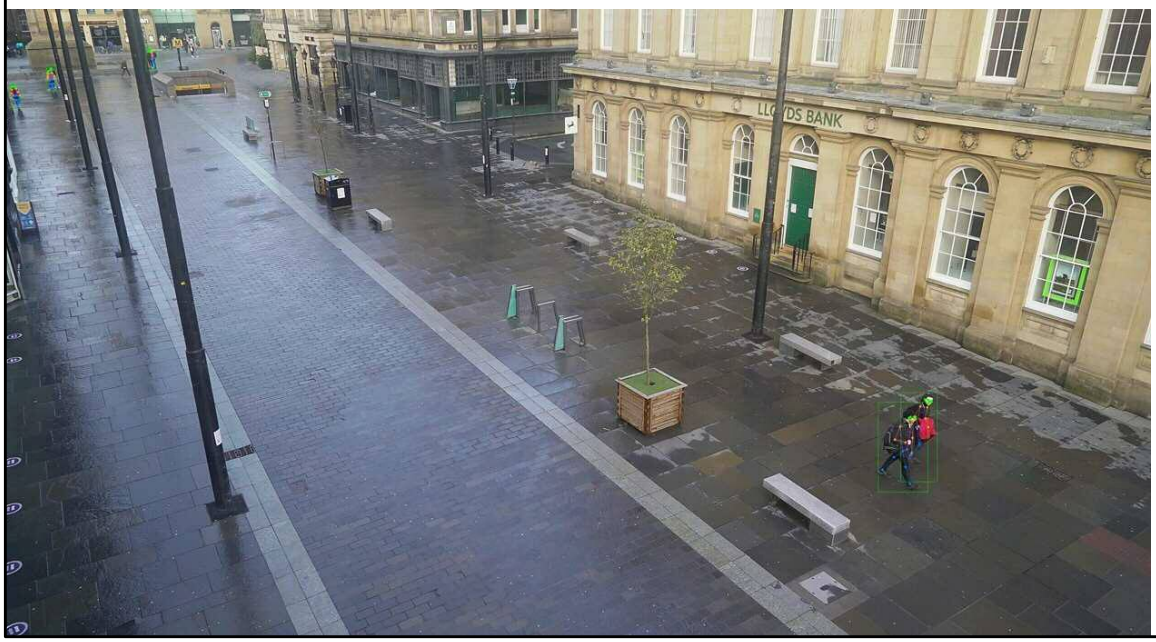
PROTECT
A COVID-19 National Core Study



Initial work with zoom call recordings was a great playground as everyone there was facing directly towards the camera, even allowing for estimation of which exact area of the face was touched. Here the first lesson was that all individuals visible in their little zoom call windows needed to be analysed separately. That's because lighting conditions in each of their own rooms were different, and needed to be contrast-corrected and processed individually.



That's why we split the process into stages:
Person detection and tracking to make sure that we collect data about every individual separately. Image pre-processing to improve clarity and contrast and increase chances of successful detection. And lastly face and hand keypoints estimation. When hand is detected near the face, an image is saved for later verification. In practice that means that we don’t have to watch the whole videos but just verify if candidate images in fact include face-touching.



Second scenario involved observations in public space where a high resolution camera observed a pedestrianised section of a street. Being optimistic after the experiments with zoom calls we went straight into what seems now like the most complicated scenario. People were observed from a far distance, they moved in all directions and faced everywhere except towards the camera. With the use of full-body pose estimation models we attempted to gather some meaningful data but turned out that 60% were non-verifiable or false positives with people oriented away from the camera, resolution being insufficient to confidently asses the action, or with behaviors with hands near the face but far from touching it.

The lesson here was that automation is indeed possible but our naive implementation brought a lot of false-positive candidates. We may want to find out in the future if stacking an additional image classifier to prune the false-positives would improve accuracy of the method applied to that environment.

Another take on the same scenario was made when we moved the camera lower down to see more detail of people's faces and oriented the camera so we had people moving along a uniform path towards the camera.
Here we were able to detect both face and hand keypoints, but the disadvantage was that people moved very quickly across the scene. This prevented the keypoint estimation models outputs from stabilising. We overcame that by test-time augmentation which in practice means presenting the same frame multiple times, applying slight variations, effectively making repeated measurements of the same substance while stirring it. This way we got rid of majority of false positives and captured more positive samples than before.

In summary, key takeaways from our experiments are that image preprocessing plays an important role in improving performance of a model from a different domain, test-time augmentation provides more input to the model which allows for smoother, more accurate results and that it is critical to position the camera so it has a clear, direct, unobscured view of the subjects.