Determining Crowd's Ability to Distinguish Deepfakes in Images and Videos

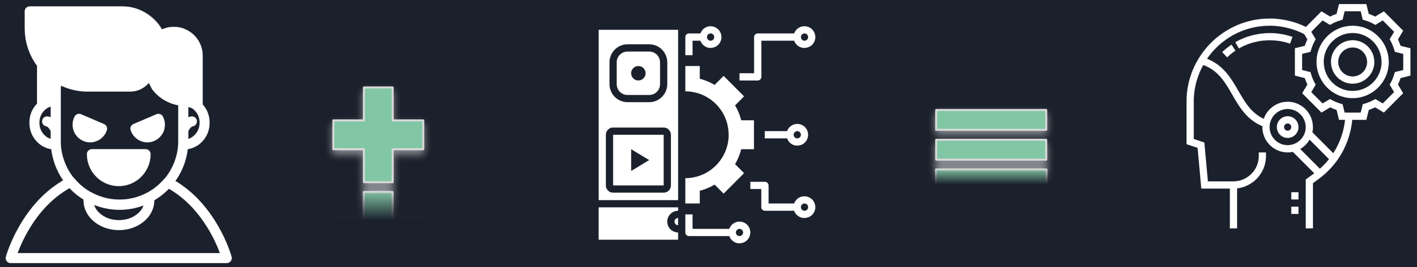
Suhas Dara and Aditya Tyagi

Overview

1. Introduction and Context
2. Related Literature
3. Dataset
4. Task Design
5. Data Analysis
6. Limitations and future work

Introduction and Context

- Misinformation
  - Spread incorrect information - deliberately - regardless of the intend to deceive
- Deep fake Media
  - Usually generated using deep learning algorithms such as Generative Adversarial Networks (GANs)



Introduction and Context

Introduction and Context

Research Questions:

1. Does the crowd detect deepfaked videos better than deepfaked images?
2. Does the crowd reason differently for images and videos?
3. What features does the crowd use to aid in deepfake detection?
4. What demographic of crowd has the highest accuracy?
5. Does performing well on a CRT test imply a higher accuracy at identifying deepfakes?

Related Literature

- Korshunov and Marcel (2020) [1]
  - Conducted studies to compare accuracies between humans and machine detection
  - Focused primarily on automated detection
  - Human subjects were pre-chosen and not randomized
- Yang, Li, and Lyu (2018) [2]
  - Researched for factors that could assist in fake image detection
  - Used Support Vector Machines (SVM) to detect errors when looking at head poses
  - The results were directed towards creation of classifiers, not for any points in manual detection
- Nguyen et al (2019) [3]
  - Conducted surveys to understand key features when attempting to detect deep fake media
  - The survey goes through different modes of detection but does not quantify any human detection rates
  - Again, focused heavily in computer detection

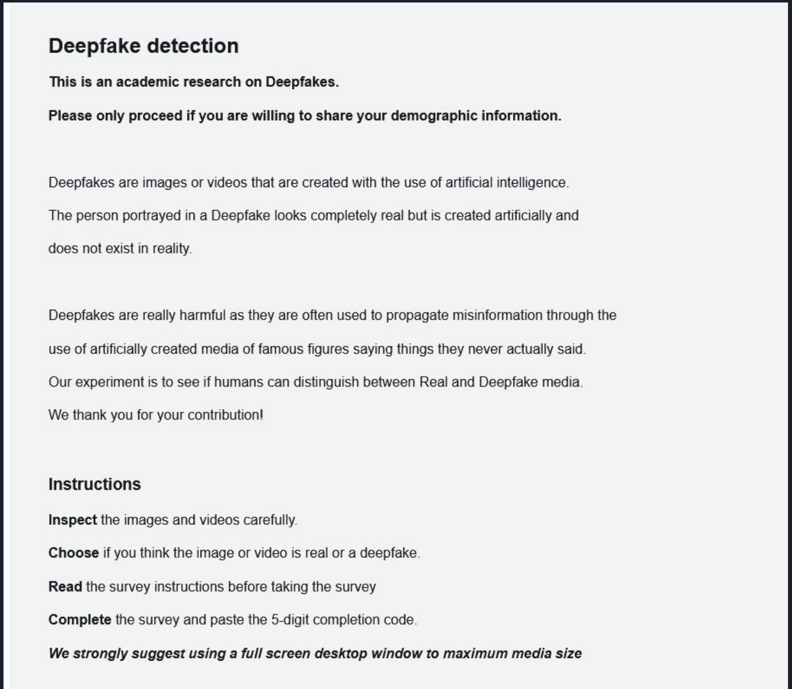
Dataset

- Video dataset
  - Sampled from the Facebook Deep-Fake Detection Challenge (DFDC) dataset
  - Randomly sampled 20 real and 20 deep fake videos out of 124,000
- Image dataset
  - Created manually using data from the website https://whichfaceisreal.com
    - Presents real and deep fake images in a game format for people to test their skills
    - The website itself samples deep fake images from another website https://thispersondoesnotexist.com which generates them using GANs.
    - The website samples real images from the Flickr-Faces-HQ (FFHQ) public dataset
  - Sampled 20 real and 20 deep fake images in the order they appear in the website. No images were skipped in order to avoid bias from our end

Task Design

- Batch of 2 images and 2 videos to annotate
    - Important: It may not be the case that one image is real and one is a deep fake. It is randomized
    - Provide a label - real or deep fake
    - Provide one rationale for image classification and a separate rationale for video classification
- Each batch of media is labeled by 5 workers to receive multiple annotations
- A single worker may label multiple batches but not a single batch multiple times
- Workers provide demographic and other information through a Qualtrics survey
    - Age, Gender, Education level, and Race(s)
    - Standard 3-question CRT according to Toplak, West, and Stanovich [4]
    - Workers receive a completion code which they can report back in the HIT
- Workers face one attention check that asks them to click an appropriate button

Task Design



Deepfake detection

This is an academic research on Deepfakes.

Please only proceed if you are willing to share your demographic information.

Deepfakes are images or videos that are created with the use of artificial intelligence.

The person portrayed in a Deepfake looks completely real but is created artificially and does not exist in reality.

Deepfakes are really harmful as they are often used to propagate misinformation through the use of artificially created media of famous figures saying things they never actually said.

Our experiment is to see if humans can distinguish between Real and Deepfake media.

We thank you for your contribution!

Instructions

Inspect the images and videos carefully.

Choose if you think the image or video is real or a deepfake.

Read the survey instructions before taking the survey

Complete the survey and paste the 5-digit completion code.

We strongly suggest using a full screen desktop window to maximum media size

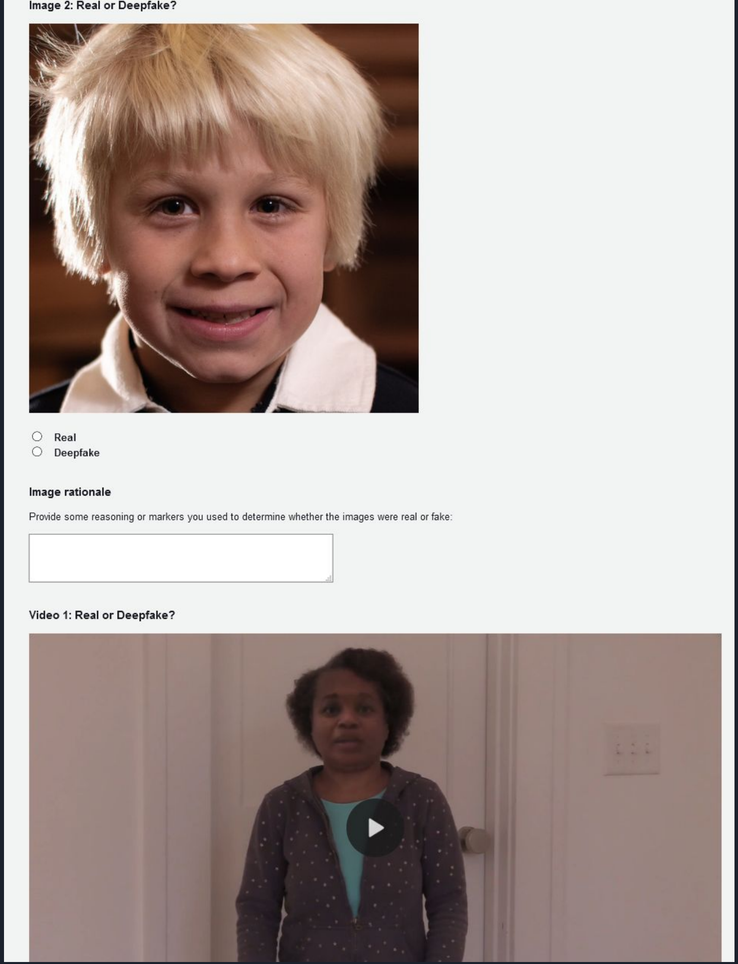


Image 2: Real or Deepfake?

○ Real
○ Deepfake

Image rationale

Provide some reasoning or markers you used to determine whether the images were real or fake:

Video 1: Real or Deepfake?

Data Analysis

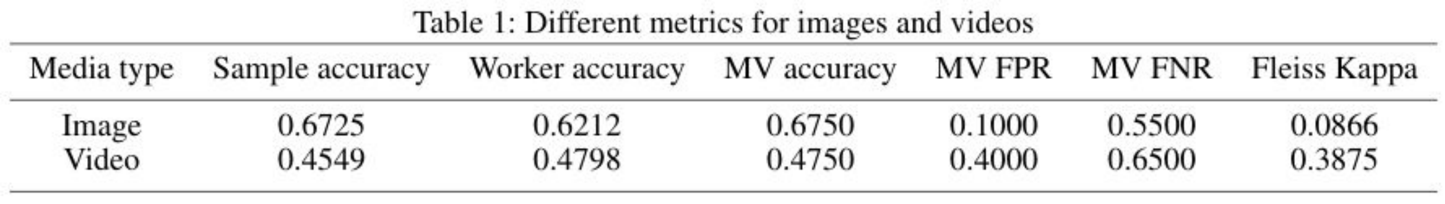


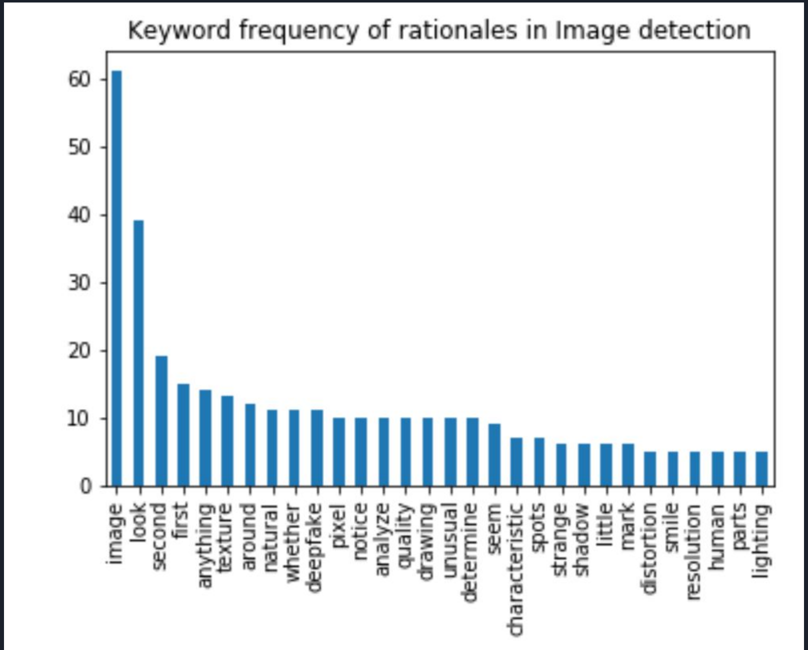
Table 1: Different metrics for images and videos

Media type	Sample accuracy	Worker accuracy	MV accuracy	MV FPR	MV FNR	Fleiss Kappa
Image	0.6725	0.6212	0.6750	0.1000	0.5500	0.0866
Video	0.4549	0.4798	0.4750	0.4000	0.6500	0.3875

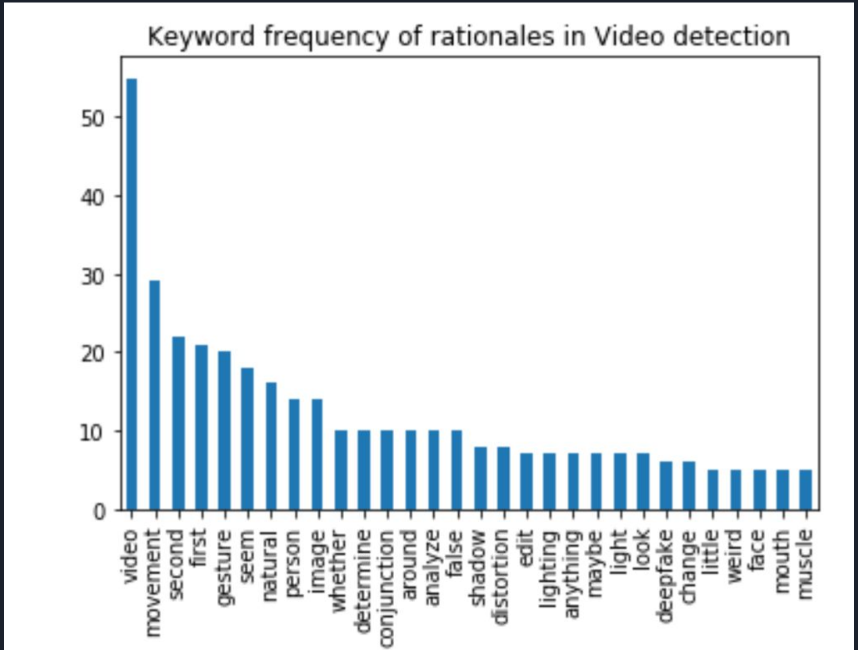
MV - Majority Voting | FPR - False positive rate (real labeled as fake) | FNR - False negative rate (fake labeled as real)

- Accuracy in images is higher than in videos across all accuracy metrics. Accuracy for videos is worse than random guess
- Inter-annotator agreement is better for videos compared to images even though accuracy is lower for videos
- False Negative Rate is quite high for both images and videos, which shows that humans are susceptible to think deep-fakes are real
- False Positive Rate is low for images, but relatively high for videos. This shows humans were more cautious about fake videos, but the extra caution didn't help

Data Analysis



Keyword frequency of rationales in Image detection



Keyword frequency of rationales in Video detection

Data Analysis

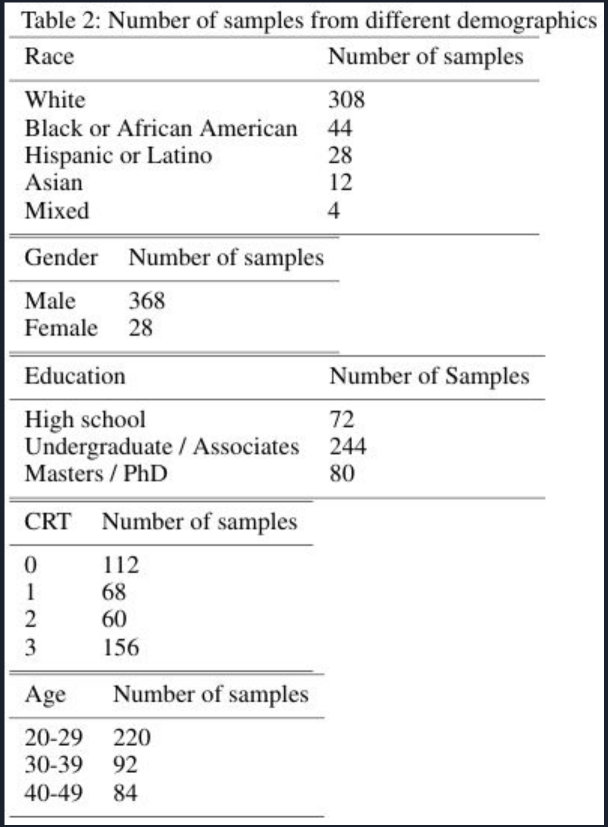


Table 2: Number of samples from different demographics

Race	Number of samples
White	308
Black or African American	44
Hispanic or Latino	28
Asian	12
Mixed	4

Gender	Number of samples
Male	368
Female	28

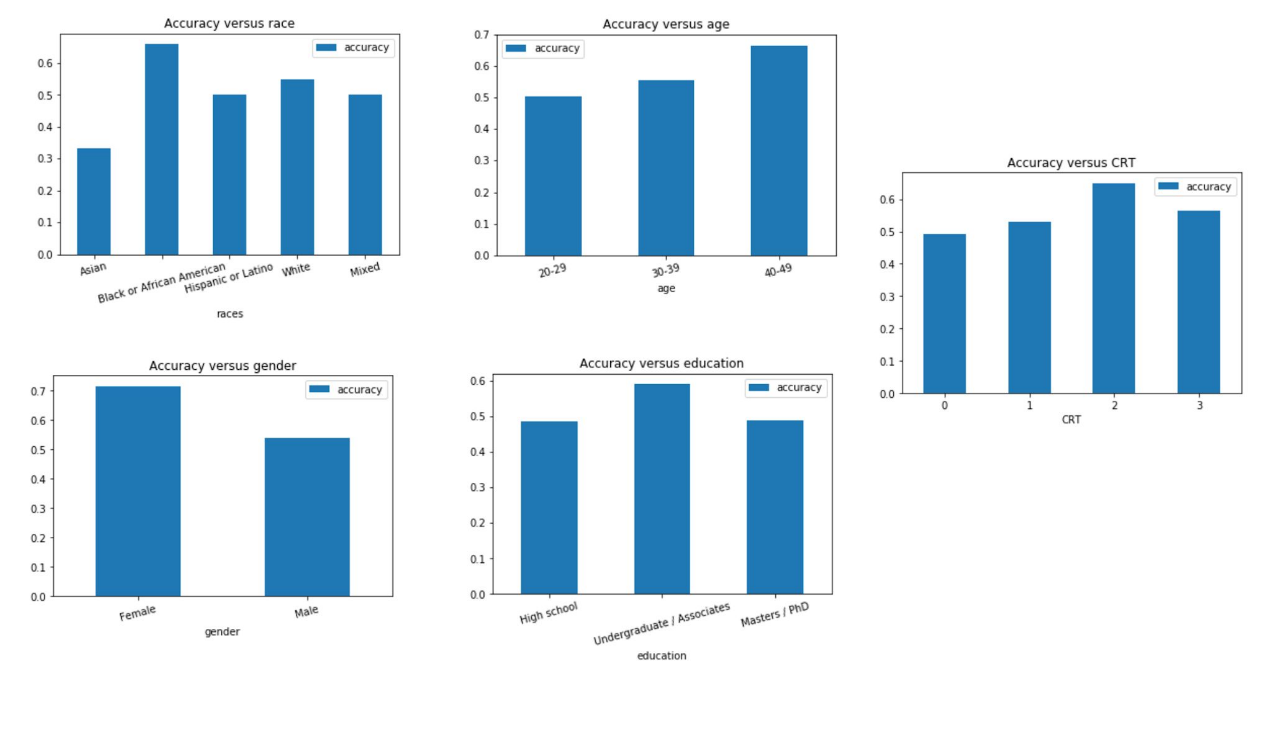
Education	Number of Samples
High school	72
Undergraduate / Associates	244
Masters / PhD	80

CRT	Number of samples
0	112
1	68
2	60
3	156

Age	Number of samples
20-29	220
30-39	92
40-49	84

- 400 samples labeled by 25 workers
- 4 samples were excluded as one worker provided a fake demographic survey code
- All workers passed attention check
- Imbalance in samples per demographic class
  - More prevalent in gender and race compared to education and age.
  - CRT scores have the least imbalance.
- Certain demographic classes weren’t observed in the age group - not fully representative

Data Analysis



One way-ANOVA for Gender:

F-value 1.6724 P-value 0.2148 >> 0.05 (Statistically insignificant)

One way-ANOVA for Race:

F-value is 1.4056, and P-value is 0.2682 >> 0.05 (Statistically insignificant)

Age vs worker accuracy r-value:

0.4193 (positive correlation)

Education vs worker accuracy r-value:

0.1705 (weak positive correlation)

CRT vs worker accuracy r-value:

-0.1248 (weak negative correlation)

Final points

- Conclusions from user study
    - If a researcher wants to use crowdsourcing to aid in deepfake media detection
    - They should filter out their workers:
        - Cognitive Testing
            - Higher accuracy is better
        - Demographics
            - At least a college degree (undergraduate/graduate)
            - Above the age of 30
    - Rationales
        - Instruct the workers to look , in a video, for:
            - Odd movement, odd gesture, odd shadows, and distortions
            - Can also focus on face and mouth specifically
        - Instruct the worker to look, in an image, for:
            - Odd texture, odd pixel, spots, and distortions

Limitations and Future Work

- Dataset
    - Size: constrained due to manual creation
    - Origin: The origins of the video dataset and the image dataset is vastly different. The videos originate from a state-of-the-art dataset, while the images were sampled from a game website
    - Future: sample frames from videos, but need to make sure not blurry
- Worker pool
    - Not very diversified
    - Demographic data collected was skewed
    - Did not collect a big enough sample size from the female gender or non-white races.
    - Future: Utilize Amazon Mechanical Turk to access a more generalized crowdworker population
- Worker sample size
    - Smaller than expected
    - No limitations for workers to do the number of tasks
    - A few workers completed a lot of HITs contributing a lot more samples toward the demographic classes they represent
    - Future: Utilize Amazon Mechanical Turk to control one HIT per worker

Thank you!

References

1. *(Korshunov & Marcel, Deepfake detection: humans vs. machines 2020 https://arxiv.org/abs/2009.03155)*
2. *(Yang et al., Exposing Deep Fakes Using Inconsistent Head Poses 2018 https://arxiv.org/abs/1811.00661)*
3. *(Nguyen et al., Deep Learning for Deepfakes Creation and Detection: A Survey 2020 https://arxiv.org/abs/1909.11573)*
4. *(McDowd et al., Effects of aging and task difficulty on divided attention performance 1988 https://pubmed.ncbi.nlm.nih.gov/2967880/)*
5. *(Ackerman et al., The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks 1994 https://link.springer.com/article/10.3758/s13421-011-0104-1)*
6. *(Kaufmann et al., More than funand money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk 2011 https://www.researchgate.net/publication/216184483_More_than_fun_and_money_Worker_Motivation_in_Crowdsourcing--A_Study_on_Mechanical_Turk)*
7. *(Hettiachchi et al., Effect of Cognitive Abilities on Crowdsourcing Task Performance 2019 https://link.springer.com/chapter/10.1007/978-3-030-29381-9_28)*