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

Suhas Dara
Department of Computer Science
The University of Texas at Austin
suhasdara@utexas.edu

Aditya Tyagi
Department of Electrical and Computer Engineering
The University of Texas at Austin
adityatyagi6498@utexas.edu

Abstract

This paper provides a user study to explore how crowdsourcing can be used to detect deepfake media. We quantify our exploration with Amazon Mechanical Turk (AMT) crowd workers and conducting surveys and Human Intelligent Tasks (HITs) through Sagemaker Ground Truth. Workers are asked to distinguish between real and fake media to the best of their knowledge and are asked to fill out a brief demographic survey. We then conduct further analysis to understand how well do the crowdworkers perform and recognize patterns, if any. We then conclude our results and user study with the quantified data and any recommendations for future research involving deepfake detection using human computation.

1 Introduction

Exploitation and weaponization of social media to spread misinformation is a major issue in The United States. When discussing the idea of "fake news", Figueira and Oliveira refers to ideas of news articles that are tagged with catchy headlines and inaccurate or distorted information (Figueira and Oliveira 2017). It should be noted that fake news is just another issue among the current issues of misinformation and there is an ample amount of research going on to mitigate the spread as well as stop it all together.

Although there is a plethora of academic research available when attempting to stop fake news, misinformation spreading through deepfakes comes as a new challenge: deepfake images are face swapped images that can now be done through readily available GPUs. The ease of access to such technology makes it lucrative for people to create deepfakes for entertainment purposes, as well as for target attacks to spread misinformation about certain individuals and institutions (Dolhansky et al. 2020). Such convincing media can be easily spread in social media to be readily available to unsuspecting citizens. One of the more infamous examples is from President Obama and his video posted online [1] where Jordan Peele shares the grim reality and the ease of creating a deepfake video that can make politicians such as the former president of The United States say offensive and inciting statements, as seen in Figure 1. Furthermore, deepfakes can get more detailed as more data is available to train the deep learning algorithms. President Obama's video was created after feeding the algorithm fifty-six hours of sample recording (Cook 2018) indicating that a large number of sample recordings or input can result in more convincing media. This is especially a problem with celebrities and politicians who make numerous public appearances, thereby having more samples available publicly. Such platforms should have the ability to utilize manual power, in addition to their automated systems, such as workers of civil society to aid in thwarting synthetic media.(Leibowicz 2019)

Figure 1: Screenshot captured from the infamous Obama deepfake video where the former president says aggravating words against President Trump, leaving the audience with a eerie reality that they cannot trust everything that they see and hear on the internet. The screenshot is captured by Fagan, with Business Insider (Fagan 2018).

To combat such issues, there has been extensive academic research with datasets created by institutions such as Facebook to understand how to identify deepfake media as well as create algorithms and new ways to mitigate the spread of misinformation through deepfakes.

This paper focuses on conducting a user study in order to gain more information to increase deepfake detection using crowd sourcing. The experiment crowdworkers will be assigned from Amazon Mechanical Turk (AMT) through the use of Amazon Sagemaker Ground Truth (GT). The overall budget of the project will be constrained under USD 300, to ensure that we are able to conduct comprehensive data col-

[1] https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed

lection and pay the workers an appropriate amount as well.

We aim to pose some research questions that enable us to conduct an exhaustive user study that would help future academic interests as well:

RQ1: Does the crowd detect deepfaked videos better than deepfaked images?

RQ2: Does the crowd reason differently for images and videos?

RQ3: What features does the crowd use to aid in deepfake detection?

RQ4: What demographic of crowd has the highest accuracy?

RQ5: Does performing well on a CRT test imply a higher accuracy at identifying deepfakes?

Through this user study, we aim to answer these questions and enable researchers to understand the factors that influence the crowd when it comes to deepfake detection and utilize/exploit the results as need be.

2 Related Works

There has been extensive academic work to understand and mitigate the spread of deepfakes: from the creation of surveys to the creation of new algorithmic approaches in order to incorporate Artificial Intelligence in the combat of spreading misinformation through media.

Korshunov and Marcel compare the accuracy between humans and machines when it comes to deepfake videos (Korshunov and Marcel 2020), that aims to find how well human subjects are able to detect deepfake videos. The research, however, did not focus on the different factors that influence the human detection and how to fine-tune them for best results in accuracy, but rather focused on how the accuracy of machine detection compares to human detection. The study also did not focus on crowdworkers but rather 60 participants who were not randomized and instead pooled from a controlled environment.

Further, there have been surveys such as Nguyen et al. that aim to understand the key features that a machine model would need to lookout for in order to conclude if an image or a video is fabricated (Nguyen et al. 2019). The survey goes through different modes of detection of deep fake media, including physiological indicators. However, they do not conduct quantified experiments of any human detection and features that enable them to see if the media is a deepfake or not.

Moreover, Yang, Li, and Lyu explored the creation of deepfake images to detect inconsistent head poses as indicators of a fake image (Yang, Li, and Lyu 2018). The authors trained Support Vector Machine (SVM) classifiers that detect the error that is posed when AI creates an image by splicing synthesized face regions on an original image. However, the paper quantifies its experiment by studying the detection only through head orientation vectors and attempting to create a classifier, it does not incorporate the human element when detecting fake close up images that the authors trained their model to do.

Overall, there are publications that aimed to mitigate the spread of incorrect information through detection of fake images and media. However, a lot of the research is focused on the creation of better classifiers and automated detectors. Although it is vital to enforce the automation for detection, Leibowicz believes that the systems would need to be integrated with manual knowledge of journalists and other worker to provide more context that AI currently cannot (2019.) Therefore, we aim to utilize the pool of crowdsourcing workers to divide detection in micro tasks as opposed to using classifiers or AI as parts of our contributions to the deepfake detection research.

3 Experiment

We conducted heavy data analysis offline. From the crowdsourcing perspective, we used the USD300 budget to create task designs that were aimed to gather basic demographic information from the workers, as well as their annotations. We used techniques such as attention checks to ensure that we filtered out bad data points as well as calculated inter-annotator agreement data to ensure that if we observed any outliers in the data points, we were either able to explain the cause or are able to replace it. We also tapped into different dimensions of a crowdsourcing task as described by Sakamoto et al. (2011) in an effort to effectively communicate with the workers as requesters. This ensured that we were able to use the worker pool efficiently by creating tasks that suit them and convey our expectations well (Sakamoto et al. 2011). The task was iterated over to ensure that all instructions and other aspects of the task design are clear. We also planned to release versions of the task at different times to ensure that we attract majority workers from the United States and India, AMT's biggest population of workers (McAllister Byun, Halpin, and Szeredi 2015). This should have also ensured that we do not overuse the worker population by publishing multiple tasks over a short period of time.

3.1 Dataset

For this project, we used two different datasets. The first dataset consisted of videos sampled from the Facebook Deepfake Detection Challenge Dataset (DFDC)[2], which featured 124,000 videos. These videos either contain deepfakes or real footage of people. We utilized these videos to analyze the capabilities of crowdworkers in detecting deepfakes from videos. The second dataset consisted of deepfake and real images of people. There is currently a lack of image datasets for deepfakes, which is the reason we sampled our own dataset for deepfake images. We used a website [3] that compiles deepfake and real images of people in a game format. The deepfake images are created using a Generative Adversarial Network (GAN)[4] and are open source. The real images are sampled from the Flickr-Faces-HQ (FFHQ) Dataset[5], which is public domain.

We chose 20 real videos and 20 deepfake videos randomly from the DFDC dataset that were short in duration and had a file size of lower than 5MB. Additionally, 20 pairs of real

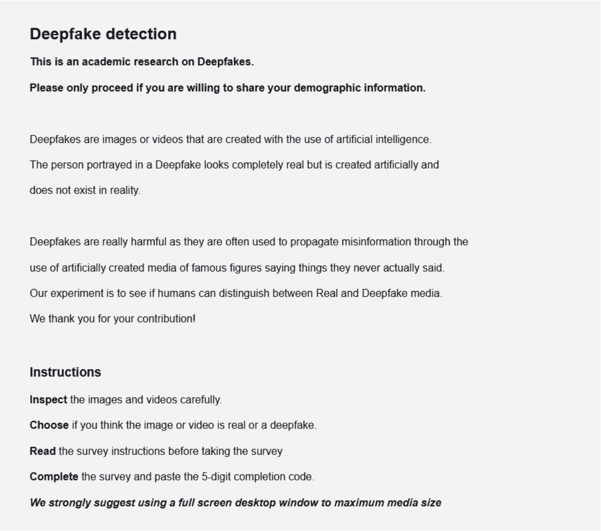
[2] https://ai.facebook.com/datasets/dfdc/

[3] https://whichfaceisreal.com

[4] https://thispersondoesnotexist.com/

[5] https://github.com/NVlabs/ffhq-dataset

and deepfake images were sampled from the game website. With this, we compiled a total of 40 videos and 40 images. To avoid bias in aggregating the images from the website, no pairs of images were skipped regardless of whether it seemed easy or difficult to identify which image in the pair was a deepfake. Consequently, the first 20 pairs were all chosen.



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

Figure 2: Screenshot of the current HIT instruction design that the workers on AMT see.

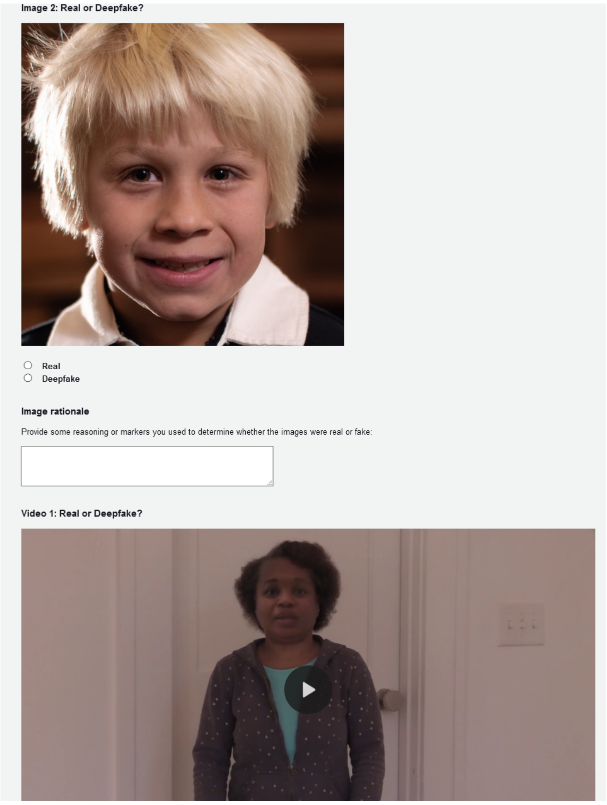


Image 2: Real or Deepfake?

○ 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?

Figure 3: Sample of the HIT that consists of two images and two videos for annotation, two free-form fields to provide rationale for images and videos, and a link to the demographic survey.

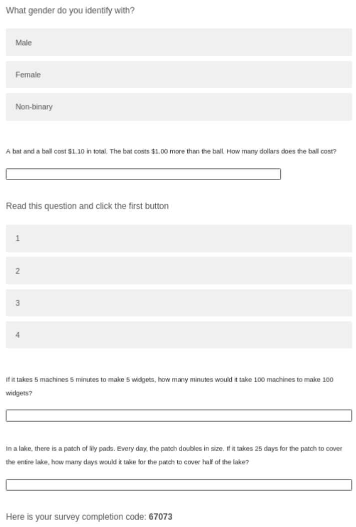
3.2 Task design

Data was collected by publishing Human Intelligence Tasks (HITs) on Amazon Sagemaker GT, which publishes the tasks to AMT for crowdworkers to annotate. Crowdworkers were asked to annotate the images and videos as real or fake. We used the available ground truth extracted from the game website and the DFDC dataset to analyze the responses of the crowdworkers. To understand trends in the crowdworkers performing our HITs, the HITs contained a batch of media - two images and two videos each. Additionally, each HIT was annotated by five different crowdworkers. The images and videos selected for each HIT wer completely randomized from the dataset and it was possible that a single HIT could have two real images or two deepfake images instead of being balanced with one of each. This was to mitigate correct answers based on any potential uniformity bias that could corrupt the data.

The crowdworkers were also asked for their reasoning behind why they thought the images or videos were real or deepfakes. They were only asked for this rationale after annotating the two images in the batch, and again after annotating the two videos in the batch. This was to avoid fatigue among the crowdworkers by answering multiple free response questions. It additionally provided us with an opportunity to analyze potential differences in their thinking processes when evaluating images versus evaluating videos. A part of an example HIT can be seen in Figures 2 and 3.

Along with the annotations, crowdworkers were also asked for their demographic information and to answer three CRT questions. This data was used in the analysis of the annotations. This data was not asked directly in the HIT to ensure that crowdworkers do not have to fill the demographic information multiple times if they chose to perform multiple HITs. Instead, the method of collecting the demographic information and the CRT responses was through a Qualtrics survey. A link to the survey was available in the HIT. The survey provides a “Survey Completion Code” at the end of the survey which the crowdworkers were asked to paste in their HIT response. If a crowdworker performed multiple HITs, they could use the same completion code each time. The survey responses were combined with the annotation responses during analysis with the help of the completion codes. A portion of the survey can be seen in Figure 4.

As part of the user study, filtering bad data was an important precursor to the analysis of the data. This bad data may be intentional or unintentional (outliers). One way of knowing whether crowdworkers were annotating the dataset diligently is by the use of attention checks. Attention checks would help filter out inattentive crowdworkers, but malicious workers would still be hard to catch. An attention check was to be added to the survey instead of the HIT itself. This was to prevent spam demographic information. Additionally, the input field asking for the survey comple-



What gender do you identify with?

Male

Female

Non-binary

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How many dollars does the ball cost?

Read this question and click the first button

1

2

3

4

If it takes 5 machines 5 minutes to make 5 widgets, how many minutes would it take 100 machines to make 100 widgets?

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 25 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake?

Here is your survey completion code: **67073**

Figure 4: Screenshot of a portion of the survey that the workers on AMT will fill out. The HIT consists of three demographic questions, three CRT questions, and an attention check. The survey completion code can be seen at the bottom.

tion code in the HIT would act as another attention check in itself. This was because it is not a completely obvious step and if the crowdworker does not properly read the instructions, they might not realize to save the code from the survey.

Another way of filtering bad data is through the use of honeypot questions. However, considering the complexity of the task, honeypots would be difficult to create/choose from our dataset and may not be a very effective metric to filter data. The deepfake game website provides some useful insights about how to detect deepfakes[6]. According to the website, some dead giveaways are water splotches and bad backgrounds. Maybe images and videos with these issues can be carefully handpicked from the datasets to be chosen as honeypots for future works. We decided not to utilize honeypots as part of this user study due to a lack of unbiased videos that had the said dead giveaways and could be easily detected, and for the purposes of consistency, we chose to eliminate honeypots for images as well.

3.3 Deployment of tasks

To ensure that we do not overuse the worker pool, the HITs were released in batches. 50% of the batches were released in the working hours of the United States to cater to the worker pool here. To ensure that all our work is seen by the US population, the second batch was delayed by at least 5 business days in US and then published in the working hours of India to ensure that we capture the largest pool of workers by catering to India and US. Although there was a risk that some Indian workers would see the tasks deployed for US workers due to latency or because the tasks are "in queue", this method would try to ensure that the task is viewed by a diverse audience and has the opportunity to be completed with the most minimal overlap for a study that is bound to take a few weeks without violating any privacy issues and filtering workers through some personally identifiable information.

4 Hypotheses

This study aims to answer the five research questions presented in section 1. For RQ1, we hypothesize that the crowd is better at detecting deepfake videos than deepfake images because of the additional motion information being available for reasoning.

For RQ2, we hypothesize that the rationales for videos will contain a lot more responses that provide reasoning based on the video subject's motion rather than the physical features of the subject in the video. This information is unavailable in images, so the crowd will have to rely on the physical features. However, for RQ3, we hypothesize that for videos, the crowd will utilize physical features such as facial structure, environment features such as the background, and motion features to detect deepfakes, while for images, only physical features and environment features will be utilized.

For RQ4, we hypothesize that race and gender will not have any impact on the ability to detect deepfakes. However, we expect age and education to affect the ability to detect deepfakes. We expect younger workers to perform better as they may have a sharper eye for artifacts in deepfakes, and we expect workers with a higher education to perform better as they may have more honed reasoning skills through the academic process.

For RQ5, the CRT scores are measured on a scale of 0 - 3 based on the standard three cognitive ability questions (Toplak, West, and Stanovich 2011). We hypothesize that a higher CRT score implies a higher accuracy at detecting deepfakes, as the higher score implies better reasoning skills.

5 Results

With the data collected[7], several different analyses were conducted to answer the research questions presented in section 1 and test the hypotheses presented in section 4. The data collected was in the post-processed form from Amazon Sagemaker GT, and in CSV format from the Qualtrics survey. For each HIT, data included four annotations (two images, two videos), two rationales, and a survey completion code.

The survey responses were first matched with their appropriate completed HITs. Surveys with completion codes that did not have associated HITs as well as multiple surveys from the same worker were discarded. A total of 76 survey responses were collected, however, only 25 were associated with workers. Additionally, HITs that contained survey completion codes that were invalid were discarded. A total of 400 annotations were collected for images and videos combined, and 4 annotations were discarded due to

[6] https://www.whichfaceisreal.com/learn.html

[7] https://github.com/suhasdara/Deepfake-Detection

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

Table 1: Different metrics for images and videos. MV - Majority Voting, FPR - False Positive Rate, FNR - False Negative Rate.

invalid survey completion codes. The responses were then filtered through an attention check that all workers successfully passed. The remaining analysis is based on 396 annotations performed by 25 unique workers.

5.1 Accuracy and agreement

Worker responses were analyzed under several different metrics to understand and compare their performance on detecting deepfake images and videos. All the evaluated metrics are presented in Table 1. The sample accuracy represents the weighted mean of the accuracy of the 25 workers. The sample accuracy is influenced more by workers who performed more annotations. A higher sample accuracy was observed for images compared to videos. The worker accuracy represents the unweighted mean of the accuracy of the 25 workers. A higher worker accuracy was observed for images compared to videos. The images sample accuracy was slightly higher than the images worker accuracy, implying workers who did more image annotations were more accurate. The opposite is observed with videos, implying workers who did more video annotations were less accurate.

For each image and video in the dataset, a majority vote was calculated. Because each image and video was labeled by 5 different workers, the majority vote cannot be tied. A higher majority voting accuracy was observed for images compared to videos. In all three accuracy metrics, workers performed better on images than videos. Interestingly, worker accuracy on videos was lower than random guess across all metrics.

With majority voting aggregation, it was observed that the false negative rate (deepfake media annotated as real) was quite high for both images and videos. This demonstrates the exact dangers of deepfakes, as workers are highly susceptible to think deepfakes are real. However, it was also observed that the false positive rate (real media annotated as deepfakes) was also relatively high for videos. This could indicate that workers were more cautious about videos being deepfakes, but the extra caution did not help as seen with the accuracy.

Lastly, the workers' inter-annotator agreement was measured using Fleiss Kappa to understand whether the workers agreed on their annotations. The Fleiss Kappa value for videos demonstrated a slight agreement among the workers. This is an interesting result because this implies that workers often agreed on incorrect annotations as the accuracy is low. It is a possibility that certain deepfake videos in the dataset were created very well and tricked a lot of the AMT workers. The Fleiss Kappa value for images was close to 0 implying there was only agreement equivalent of random chance among the workers. However, even with low agreement, workers performed better at detecting deepfake images compared to deepfake videos.

To answer RQ1, the crowd does not perform better at detecting deepfake videos, and even perform worse than random guess. Hence, our initial hypothesis is not supported by the results. One possibility is that there is too much information in a deepfake video to process at once and workers may get confused by certain combinations of features. Other potential reasons for this result are discussed in the limitations of this research in section 6.

5.2 Rationales

To analyze the rationales provided by workers, Li's approach (2018) was utilized to tokenize the sentences. A word is considered to be a keyword when it is at least 4 characters long. The tokenized data was cleaned accordingly and the frequencies of the keywords were calculated. For both videos and images, there were more than 180 unique keywords that were recorded. The top 30 most frequent keywords for both videos and images were chosen to be the most used and relevant keywords. The frequent keyword boundary was chosen arbitrarily.

When looking at the rationales provided for images (Figure 5), it is hard to decipher what the general pattern of key features the workers were observing. Although certain keywords with higher frequency are "texture" and "pixel", it is still unclear to requesters what useful instructions can be constructed out of these keywords to ensure that the crowd looks for these particular features when attempting to detect deepfake images. Through reviewing the top keywords, it can be concluded that the general populace of the workers would rationalize their decisions based on texture, spots, distortions, and other image quality features rather than the subject of the image.

However, looking at Figure 6 for video rationales keywords, we notice that the keywords in high frequency are important features that aid in detection of deepfake videos. With keywords such as "movement", "gesture", and "shadow" appearing in high frequency, we observe that crowd workers are using motion features equally or more often than image features when attempting to detect deepfake videos.

To answer RQ2 and RQ3, the crowd utilizes different features when trying to detect deepfake images and videos. The utilized features for images are more vague compared to utilized features for videos, though they tend to focus on the quality of the image rather than the subject of the image. The specific keywords extracted from the video rationales revolve around the details of the subject in the video including shadows, movement, gestures, and the environment around the subject in addition to the standard image quality keywords observed for images such as distortions. Our hypothesis for RQ2 is supported with the presented evidence. However, our hypothesis for RQ3 is supported for videos, but only partially supported for images as workers did not often utilize physical features of the subject.

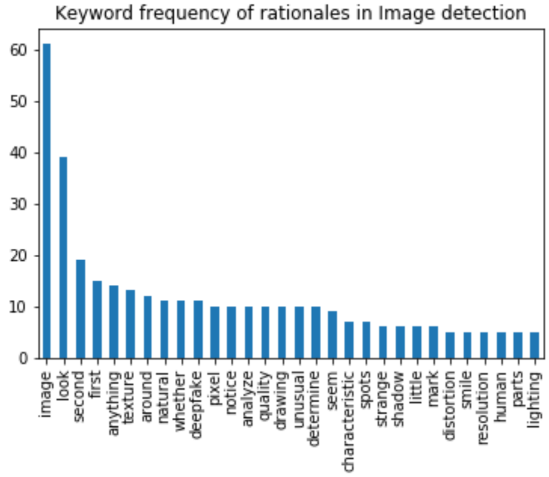


Figure 5: Top 30 keywords used, with their frequency, when AMT workers rationalized their reasoning for image detection.

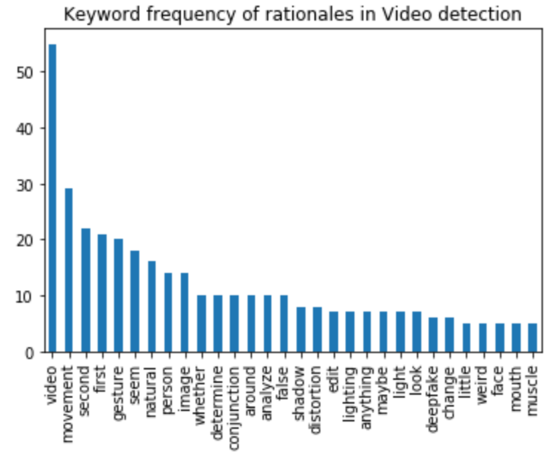


Figure 6: Top 30 keywords used, with their frequency, when AMT workers rationalized their reasoning for video detection.

5.3 Demographic and CRT analysis

The demographic information from the survey data was utilized to understand the correlation between worker accuracy and the demographic classes that they belong to. This analysis was also conducted on the workers' CRT scores as seen in Figure 7.

However, the demographic information was skewed for several of the demographic classes as seen in Appendix A[8]. This was most prevalent in Gender where females contributed only 28 out of the 396 samples. Additionally, Race

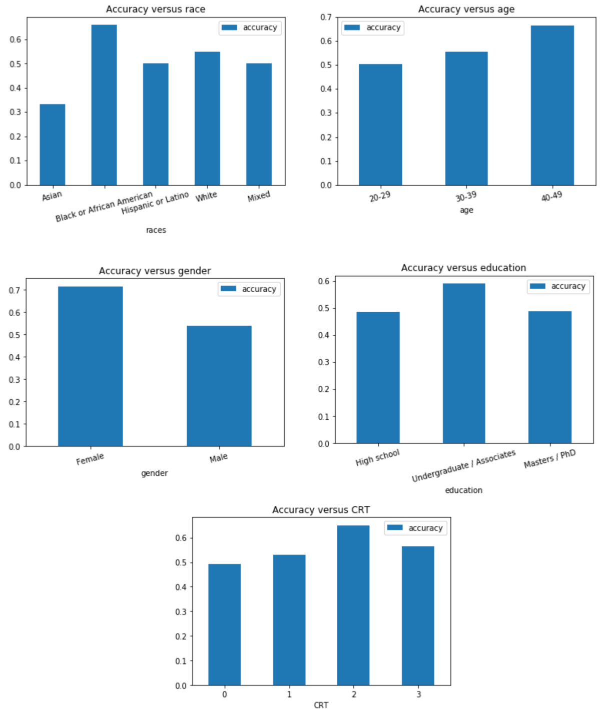


Figure 7: Different accuracy graphs plotted based on demographic information.

was predominantly comprised of White workers. Education and Age are slightly less skewed but dominated by Undergraduate and 20-29 year old workers respectively.

Gender It was observed that the female accuracy was higher than male accuracy. As previously mentioned, this may be because of a lack of similar amounts of samples from the female demographic as compared to the male demographic. However, after conducting a one-way ANOVA test, a P-value of 0.2148 (P-value $>$ 0.05) was observed. Hence, there was no statistical difference in workers' accuracy based on their gender, proving our initial hypothesis correct. A worker's accuracy is independent of their gender.

Race It was observed that certain races performed better than others, but this might be because of a lack of samples again. Similar to gender, after conducting a one-way ANOVA test, a P-value of 0.2682 (P-value $>$ 0.05) was observed, indicating that there was no statistical difference in workers' accuracy based on their race either, proving our initial hypothesis correct. A worker's accuracy is independent of their race. A lower accuracy score for Asian race was probably observed due to the fact that we only collected samples from a single Asian worker, which led the score to be skewed.

Age It was observed that with an increasing age bracket, the accuracy scores increased as well. This was supported by a linear regression test which reveals a weak positive correlation of 0.4193. This proved our initial hypothesis incorrect. One plausible explanation for this pattern could be explained

[8] Also posted on Github repository

through a study conducted by McDowd, Epstein, and Craik (1988) where reaction times were quantified for older people versus young people. Their subjects included sixteen young adults (mean age of 19.4) and sixteen older people (mean age of 69) all in good health. The results showed that when looking at attention tasks, older people tended to complete them slower than young adults. This result could be applicable for deepfake detection as longer duration spent performing the task can be useful to focus on a larger set of keyword features as outlined in the rationales.

Education A linear regression test revealed a correlation coefficient for education and accuracy of 0.1705 which was very weak, indicating little to no correlation. This implied that workers could have any level of education and that would not affect their accuracy, proving our initial hypothesis incorrect. However, it is interesting that the number of samples and accuracy were very similar for workers with high school degrees and masters or PhDs. When visualized in Figure 7, they seem to be lower than undergraduates. One possible explanation for this could be derived from a survey conducted by Kaufmann, Schulze, and Veit (2011) where the primary motivator for many workers is not money but rather other factors such as education, challenge, and such. Workers with these degree levels tend to not pursue crowd work as full time roles and may have other intrinsic motivations such as task variety or other educational purposes to complete tasks. With financial gain not being the primary motivator, the workers could possibly be less cautious due to the risk of work rejection and not getting paid not being their primary motivation factors. However, previous experience with deepfakes could potentially influence accuracy more as compared to educational degrees, and is something that can be explored further.

CRT A linear regression test revealed a correlation coefficient for CRT scores and accuracy of -0.1248, indicating no correlation or a very weak negative correlation. This was unexpected and proved our initial hypothesis incorrect because previous studies have supported that worker accuracy has a positive correlation with the CRT scores (Hettiachchi et al. 2019). The tasks provided to the workers in Hettiachchi et al.'s study, such as proof reading or transcription, are at par with deepfake detection since all these tasks require high attention to detail for better accuracy. The lack of positive correlation could be attributed to the low number of workers contributing responses. We also suspect that some workers may have utilized a search engine to answer the CRT questions, achieve a higher CRT accuracy, and skew the results as a consequence.

To answer RQ4 and RQ5 based on the different demographic information and CRT scores collected from the responses, we conclude that in order to get the highest accuracy possible, crowdsourcing requesters should filter out their workers based on age. We find that workers' race and gender are statistically insignificant in determining their accuracy at deepfake detection. We also find that workers' education level and CRT scores are not correlated with their accuracy. Requesters should aim to send deepfake detection tasks to workers who are above the age of 30 in an attempt to get the best detection accuracy, as supported by our 396 samples.

6 Limitations and future work

We studied the performance of the crowd at distinguishing deepfake images and videos from real images and videos. However, there were some inherent limitations of our study that we discuss in this section.

The first limitation is regarding the dataset, which had direct consequences on our results, especially for RQ1 as seen in section 5.1. We predicted that the crowd would perform better at detecting deepfake videos than deepfake images because of the extra motion information available in videos. However, we noticed that crowd did significantly better on images that videos. This is potentially due to the origins of the datasets. The videos were sampled from a state-of-the-art challenge dataset provided by Facebook that often contained a subject's full body in frame. On the other hand, the images were sampled from a website that gamifies deepfake detection where the deepfakes were often noticeably of poor quality because of splotches and background. Additionally, all the images were headshots and did not have the rest of body in frame. This may have caused the discrepancy of higher accuracy on images, as the images were zoomed in and easy to look for the features presented in section 5.2. One possible future solution to this problem is to pull individual frames from deepfake videos not included in the video dataset (avoiding overlap) and to use them for the image dataset. Some extra quality control would be necessary to avoid blurry images but it might provide a more reasonable accuracy comparison.

Another limitation of the user study is the diversity of the worker pool that annotated the HITs. The demographic data was very skewed in certain categories such as race and gender (see Appendix A), and in general skewed in the other categories as well. The age question in our HIT demographic survey had a total of six brackets but we only received responses from three brackets. The female gender and non-White races were severely underrepresented and may have contributed to an incomplete analysis. A future solution could be to utilize AMT to publish the HITs instead of Amazon Sagemaker GT to access a more generalized crowd population and also provide more control with the release of HITs to specific populace.

Lastly, another limitation is also concerning the workers that participated in the study. A total of 396 annotations were performed by a very small sample size of 25 workers, that was also skewed demographically. A lot of these workers only performed a small number of HITs and did not contribute much data. On the other hand, a small number of workers performed a large number of HITs and contributed a lot of data. This meant that certain individuals' accuracy contributed a lot more than others. If this set of workers is high-performing, the accuracy is skewed upward, and vice-versa. The solution to this issue is also to utilize AMT as its interface allows restricting the number of HITs a single worker can perform. This functionality is unavailable in Amazon Sagemaker GT making it less viable for conducting user studies.

7 Conclusion

This user study attempted to contribute novel insights into the performance of the crowd at labeling deepfake images and videos, understanding which format of deepfake media is easier to detect, and which demographics perform the best. Based on our results, we realize that the crowd is not very competent at detecting deepfake media from real media. Even though the crowd performed much better on images than videos, this is still not indicative of a good enough accuracy to utilize the crowd for creating ground truth annotations. Additionally, based on the limitations of our user study, we believe that worker accuracy on images would also be lower if the images were sourced from the same origin.

However, should requesters want to utilize the crowd for annotating deepfake media, the best approach is to provide instructions that capture the essence of the important artifacts that the crowd should look for in a given media. This includes odd texture, odd pixels, spots, and distortions for images; and odd movement, odd gestures, and odd shadows other than the image artifacts for videos. Additionally, demographic filtration can be performed on the crowd to receive better results. Workers can be filtered by age, using middle-aged workers to provide annotations for deepfake media. Overall, there is a lot of scope for improvement of this user study as outlined in section 6 to refine the results.

References

Cook, J. 2018. Are deepfakes the new fakenews?

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge dataset.

Fagan, K. 2018. A viral video that appeared to show obama calling trump a 'dips—' shows a disturbing new trend called 'deepfakes'.

Figueira, , and Oliveira, L. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121:817–825.

Hettiachchi, D.; Berkel, N. v.; Hosio, S.; Kostakos, V.; and Goncalves, J. 2019. Effect of cognitive abilities on crowdsourcing task performance.

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing – a study on mechanical turk.

Korshunov, P., and Marcel, S. 2020. Deepfake detection: humans vs. machines.

Leibowicz, C. 2019. On ai amp; media integrity: Insights from the deepfake detection challenge.

Li, S. 2018. Topic modelling in python with nltk and gensim.

McAllister Byun, T.; Halpin, P. F.; and Szeredi, D. 2015. Online crowdsourcing for efficient rating of speech: a validation study.

McDowd, J. M.; Epstein, W.; and Craik, F. I. 1988. Effects of aging and task difficulty on divided attention performance.

Nguyen, T. T.; Nguyen, C. M.; Nguyen, D. T.; Nguyen, D. T.; and Nahavandi, S. 2019. Deep learning for deepfakes creation and detection: A survey.

Sakamoto, Y.; Tanaka, Y.; Yu, L.; and Nickerson, J. V. 2011. The crowdsourcing design space. In Schmorrow, D. D., and Fidopiastis, C. M., eds., *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*, 346–355. Berlin, Heidelberg: Springer Berlin Heidelberg.

Toplak, M. E.; West, R. F.; and Stanovich, K. E. 2011. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition* 39(7):1275.

Yang, X.; Li, Y.; and Lyu, S. 2018. Exposing deep fakes using inconsistent head poses.

Appendix

A Demographic Sample Counts

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

B Group Contributions

Research

- Data Collection - Suhas Dara
- Related Research - Aditya Tyagi
- HIT Design - Suhas Dara and Aditya Tyagi
- Survey Design - Suhas Dara

- Worker Response Collection - Aditya Tyagi
- Metrics Analysis - Suhas Dara
- Rationale Analysis - Aditya Tyagi
- Data Visualization - Aditya Tyagi

Paper Sections Abstract, 1, 2, 3, 3.3, 5.2, and 5.3 are written by Aditya Tyagi.

Sections 3.1, 3.2, 4, 5, 5.1, parts of 5.3, 6, 7, and Appendix are written by Suhas Dara.

Both authors also did contributions to sections other than the ones mentioned.