Information Research - Vol. 30 No. iConf (2025)

Uncovering strategies for identifying deepfakes

Celene Neo, Dion Hoe-Lian Goh, Chun Wan Ying Rachel, and Chei Sian Lee

DOI: https://doi.org/10.47989/ir30iConf47209

Abstract

Introduction. The proliferation of generative artificial intelligence tools capable of producing high-quality videos that can masquerade as genuine content has raised concerns about online misinformation. This study investigates human ability to identify deepfake videos, with a focus on identification performance and the strategies employed.

Method. Data was collected through an online survey. Participants were young adults aged 21 to 35. They were shown four videos and asked to identify them as real or deepfake, followed by questions about the identification strategies used.

Results. Our results revealed the diverse range of strategies utilised. Predominant strategies centre around assessing the authenticity of traits pertaining to the video's subject as opposed to peripheral details. Furthermore, we uncovered preferences for intuition and strategies that relate to individual decision-making over consulting other individuals or online materials.

Conclusion. Our results help enhance understanding of how people identify deepfake videos, adding to existing knowledge. These findings also inform initiatives aimed at educating the public about spotting deepfakes.

Introduction

While generative artificial intelligence (AI) has opened a wealth of creative possibilities, its ability to produce convincing misinformation has made it harder to discern artificial content from real content (Park, 2024). In recent times, misinformation has taken the form of deepfakes – artificially generated videos that manipulate individuals to appear to say or do things they never did (Somers, 2020). Deepfakes have gained notoriety in the public sphere primarily for their use in generating pornographic content, and more recently, to erode trust by creating high-quality, fabricated videos that falsely depict influential figures making controversial statements.

Existing studies (Heidari et al., 2024; Jung et al., 2020; Pan et al., 2020) on deepfake detection are primarily focused on those involving deep learning. Few studies investigate human deepfake detection strategies. Among those that do, they are often limited in terms of sample size and generalisability due to the use of qualitative data collection methods (Goh et al., 2022; Zeng et al., 2023). Videos used in these studies also lack diversity and consist largely of entertainment or political videos (Goh, 2024), which may not sufficiently encompass the genres of videos that participants encounter day-to-day.

This study addresses these gaps by determining human ability to identify deepfake videos through an online quantitative survey. The first objective ascertains deepfake and real video identification performance; and the second examines the strategies that people employ to identify deepfake and real videos.

Literature review

Deepfakes pose significant threats to society, particularly through their potential to create fabricated videos for spreading misinformation. The misuse of deepfakes risks undermining public trust and has the potential to deepen social divisions (Westerlund, 2019). The impact of deepfakes is further amplified when considering that individuals generally perceive video content as more credible than textual information (Sundar et al., 2021), making the spread of falsehoods through deepfakes particularly damaging.

Given the rise in deepfakes (Sumsub, 2023), there is a wealth of research on deepfake detection using deep learning models (El-Gayar et al., 2024; Lee et al., 2023). These approaches enable the rapid processing of large volumes of video data while providing objective assessments (Heidari et al., 2024), which may explain the heightened research interest over human detection capabilities.

While there is a growing body of work on human deepfake detection strategies, data collection methods are often limited to exploratory and qualitative modes such as interviews (Goh, 2024) or diary studies (Zeng et al., 2023). Despite yielding detailed insights, they are constrained by small sample sizes and limited generalisability. Findings are also often limited to the most frequently used strategies regardless of video type. For instance, in Goh (2024), strategies for identifying video authenticity are described without differentiation between methods used for detecting deepfakes or real videos. This approach limits the ability to understand which specific strategies are most effective for each type of video.

Methodology

Our study utilized an online survey. Participants were shown four videos — two real and two deepfake — randomly selected from a pool of ten authentic and ten deepfake videos. These videos were publicly available from the Web and covered a range of topics, including entertainment, politics, education, and sports. Video descriptions can be found in Table 1.

After viewing each video, participants were asked to: (1) identify whether the video was real or deepfake, (2) report their confidence level, and (3) select the strategies they used to arrive at their decision. The strategies were categorized into three types: visual (e.g., facial features, background, and environmental issues), auditory (e.g., vocal features, sound quality), and knowledge-based (e.g., online tools, knowledge of video subject). These strategies were adapted from prior research (Goh et al., 2022; Zeng et al., 2023; Goh, 2024). Participants could select multiple strategies.

Young adults aged 21 to 35 were recruited as they represent a group that is active online and would likely have experience with deepfake videos (Petrosyan, 2024). A total of 195 participants were recruited via convenience and snowball sampling.

Video Type	Topic	Description
Deepfake	Entertainment	Mark Zuckerberg says he controls billions of peoples' confidential data and thus owns their future.
		Kim Kardashian tells how she likes making money by manipulating her fans.
		Tom Cruise's daily life.
		The Shining movie clip.
	Politics	Manoj Tiwari criticized an opposing political party and encouraged people to vote for his party.
		Jeremy Corbyn supports Boris Johnson as Prime Minister.
		A speech for the Apollo 11 mission gone wrong.
	Educational	Obama reminds people to be more alert to fake news.
		Morgan Freeman asks people: Is seeing believing?
	Sports	Jose Mourinho comments on soccer.
Real	Entertainment	Mark Zuckerberg says he could ascertain people's online behaviours.
		Kim Kardashian claims that she cheated on an exam.
		Tom Holland taking a break from social media.
		Another *The Shining* movie clip.
	Politics	Biden criticized MAGA Republicans.
		Trump blamed congressional attackers and told his supporters to calm down.
		President Uhuru mourns former Kenyan President Mwai Kibaki.
	Educational	Hillary Clinton talks about fake news dangers.
		Ellen warns people about fake news.
	Sports	Jose Mourinho on Sir Alex Ferguson's response after Porto's Champions League win over Manchester United.

Table 1. Description of videos used in the study

Results

Participant demographics

Participants comprised 83 males and 112 females aged 21 to 35 years. The majority were from fields such as social sciences, finance, engineering, sciences, and computing. YouTube was the most used video platform among participants, followed by Instagram and TikTok. A majority (86%) watched videos daily, with most coming across dubious content '*once in a while*' when watching videos.

Video identification performance

To address the first objective, we found that the distribution of participants skewed towards a higher number of correct identifications. Of 195 participants, the majority (73%) were able to identify more than half of the videos correctly. Only 1% identified zero videos correctly, 6%

correctly identified one video, 21% correctly identified two videos, 34% correctly identified three videos, and 38% correctly identified all four videos (Figure 1).

Figure 2 shows identification accuracy by authenticity type. Here, 53% of participants were able to correctly identify both deepfake videos, while 65% correctly identified both real videos.

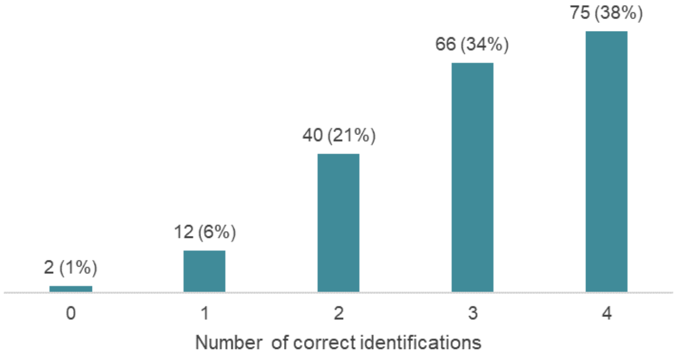

Figure 1. Frequency of number of videos identified correctly

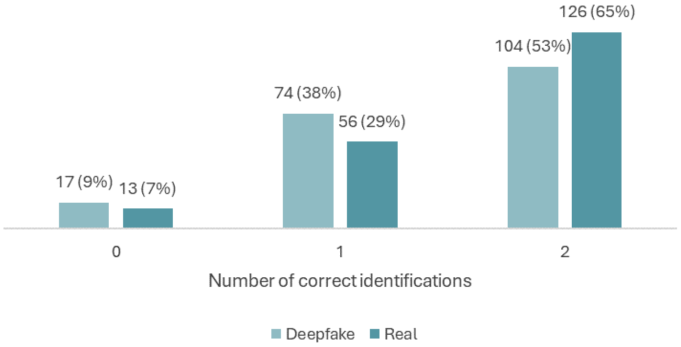

Figure 2. Frequency of correct identifications by authenticity type

Strategies used in deepfake video identification

Table 2 shows the five most and least used strategies associated with correct identification of deepfake videos. The most frequently used methods typically focused on examining the characteristics of the subject in the video, while the least used methods focused on the peripheral details of the video. Among the most used methods, physical and behavioural characteristics, and intuition and emotions were used half of the time.

Note that the frequency values presented in the tables reflect the total number of times each strategy was employed across both deepfake videos watched (Tables 2 and 3) as well as real videos (Tables 4 and 5). Each of the 195 participants viewed two real and two deepfake videos, resulting in 390 video assessments per video type. The percentages for each strategy are calculated using this total number of video assessments as the base.

Table 3 shows the top and bottom five strategies associated with incorrect identification of deepfake videos. Strategies were similar to those used in the correct identification of deepfake videos, except for intelligibility and language, and knowledge of video content which were not present in Table 2.

Strategy		Frequency
Most used	Physical and behavioural characteristics *(Analysing individuals' expressions, body gestures, and movements in the video.)*	203 (52.1%)
	Intuition and emotions *(Based on own instinct, opinions, or emotions.)*	202 (51.8%)
	Vocal features *(Assessing modulation and naturalness in the speaker's voice.)*	185 (47.4%)
	Facial features *(Analysing facial characteristics, such as skin tone, facial symmetry, and hairstyle.)*	173 (44.4%)
	Knowledge of person *(Prior knowledge of people in video.)*	124 (31.8%)
Least used	Colour and lighting inconsistencies *(Evaluating the video's lighting conditions.)*	76 (19.5%)
	Production issues *(Editing issues, camera angle/work issues, shakiness, and jitter.)*	67 (17.2%)
	Background sound issues *(Background reverberation/echoes, overall noise, mechanical noises.)*	66 (16.9%)
	Use of multiple sources *(Consulting multiple sources when using online tools or communicating with others.)*	57 (14.6%)
	Communication with others *(Checking with family or friends either offline or online.)*	53 (13.6%)

Table 2. Strategies associated with correct deepfake identification

Strategy		Frequency
Most used	Vocal features	86 (22.1%)
	Intuition and emotions	72 (18.5%)
	Facial features	69 (17.7%)
	Physical and behavioural characteristics	68 (17.4%)
	Intelligibility and language *(Assessing the language used, fluency, pronunciation, and intelligibility of speech.)*	66 (16.9%)
Least used	Knowledge of video content *(Familiarity with events in the video.)*	32 (8.2%)
	Production issues	31 (7.9%)
	Colour and lighting inconsistencies	30 (7.7%)
	Use of multiple sources	25 (6.4%)
	Communication with others	21 (5.4%)

Table 3. Strategies associated with incorrect deepfake identification

Strategies used in real video identification

Table 4 shows the five most and least used strategies associated with correct identification of real videos. The strategies aligned with those used to correctly identify deepfakes, with the only differing strategy being intelligibility and language.

In terms of the most used strategies, it was interesting to observe that vocal features and facial features saw higher frequency of use as compared to strategies associated with correct identification of deepfake videos. For example, use of vocal features was the predominant strategy, with a frequency of nearly 70%. Compared to Table 2, intuition and emotions fell from the second to fifth most used strategy despite maintaining a similar frequency of use. A similar change was observed for physical and behavioural characteristics.

Strategy		**Frequency**
Most used	Vocal features	267 (68.5%)
	Facial features	206 (52.8%)
	Physical and behavioural characteristics	199 (51.0%)
	Intelligibility and language	190 (48.7%)
	Intuition and emotions	190 (48.7%)
Least used	Online tools *(Using search engines like Google or social media platforms like Facebook and YouTube.)*	94 (24.1%)
	Production issues	76 (19.5%)
	Colour and lighting inconsistencies	66 (16.9%)
	Use of multiple sources	52 (13.3%)
	Communication with others	42 (10.8%)

Table 4. Strategies associated with correct real video identification

Table 5 shows the five most and least used strategies associated with incorrect identification of real videos. Once again, strategies were similar to those used in the correct identification of real videos, which is a common observation between correct and incorrect identification of deepfakes. Notably, this was the only instance where background and environmental details were among the most used strategies.

Strategy		**Frequency**
Most used	Physical and behavioural characteristics	49 (12.6%)
	Vocal features	47 (12.1%)
	Intuition and emotions	41 (10.5%)
	Background and environmental details *(Scene settings, unusual backgrounds, issues with watermarks, logos, or subtitles.)*	33 (8.5%)
	Facial features	31 (7.9%)
Least used	Online tools	20 (5.1%)
	Knowledge of video content	19 (4.9%)
	Colour and lighting inconsistencies	17 (4.4%)
	Communication with others	17 (4.4%)
	Use of multiple sources	12 (3.1%)

Table 5. Strategies associated with incorrect real video identification

Discussion

Overall, participants demonstrated an ability to differentiate between deepfake and authentic videos. When examining performance based on authenticity type, participants exhibited stronger proficiency in detecting real videos over deepfakes.

Across all videos watched, participants used a variety of strategies. This highlights the difficulty in human detection as there is no singular strategy that individuals can rely on to accurately distinguish deepfakes from real videos (Goh, 2024; Groh, 2020). This finding thus highlights the importance of media literacy and the use of multiple methods to ascertain the authenticity of videos.

Participants largely focused on the subjects in the video, as seen by facial features, and physical and behavioural characteristics emerging as frequently used strategies. In contrast, strategies that considered general video and audio attributes such as production quality were less frequently utilised. This may stem from the understanding that the subject is often the main focus of deepfake manipulation (Sundar et al., 2021), prompting participants to ignore other details (Groh et al., 2021). The danger however is that as the quality of deepfake manipulation increases, people relying primarily on such strategies may fail to spot falsified content.

The strategies used for identifying deepfake videos differed greatly from those used for image and text identification. A study on deepfake images found that participants often relied on peripheral features, such as accessories or clothing texture, to identify manipulated images (Bray et al., 2023). Similarly, strategies for evaluating the credibility of blog posts emphasised the importance of arrangement and alignment features (Jo et al., 2019). In contrast, deepfake video detection requires the analysis of dynamic components, such as facial expressions and body movement, which adds a layer of complexity compared to static media types. This distinction illustrates the need for a more nuanced approach to video detection, focusing on dynamic strategies that are not as relevant in text or image analysis.

Vocal features were the most used strategy in the identification of real videos. This suggests that participants perceive the human voice as a convincing feature associated with real videos. The complexity of the human voice and the difficulty of deepfakes to replicate these complexities makes vocal features a critical differentiator (Kulangareth et al., 2024), aiding participants in distinguishing between real and fabricated content. However as with the concern about visual features, deepfake technology is increasingly able to accurately clone voices (Mai et al., 2023). Reliance on vocal features only would again pose misidentification dangers.

Notably, regardless of video type and the accuracy of identification, the two least used strategies were employing multiple sources for verification and communication with others. Preference for individual decision-making may be explained by confidence in individual ability to identify deepfakes (Köbis et al, 2021). However, as pointed out by Goh (2024), the use of more cognitively demanding strategies such as referencing multiple sources increases the likelihood of correct identifications. This is especially important due to the rapid advances in deepfake generation technology.

Conclusion

This study reveals the strategies used in deepfake video identification, addressing a gap in the current literature. The findings underscore the complexity of detecting deepfakes, as evidenced by the array of identification methods utilized.

Our study offers theoretical contributions to the field of human deepfake detection and more generally, information credibility assessment, by revealing the strategies young adults use. In contrast to existing research which often reported on most used strategies (Goh et al., 2022; Zeng et al., 2023; Goh, 2024), our larger participant sample allowed for a detailed analysis of both most and least used ones. The exploration of lesser-preferred strategies offers a more comprehensive understanding of identification approaches and their potential shortcomings. Our findings provide a nuanced, multi-dimensional perspective that enhances existing literature on deepfake detection.

In terms of practical implications, our findings can inform educators in the development of media literacy curricula aimed at enhancing digital wellness and safety. Additionally, authorities and online platforms can use these insights to craft targeted strategies for combating deepfakes, including public service announcements, improving resilience against misinformation.

Despite the insights uncovered, several limitations should be acknowledged. First, the strategies recorded in the survey may not reflect the full suite of those employed by individuals in real world scenarios. Participants were informed that they had the option to rewatch videos multiple times, potentially leading to heightened scrutiny than typical online encounters, inflating the accuracy rates. Future research should aim to capture more naturalistic viewing behaviours. Second, as deepfakes become increasingly sophisticated, new detection strategies may emerge while existing ones may become less effective. Ongoing research is thus essential in providing insights into these evolving methods, ensuring that detection strategies remain relevant and effective in countering ever advancing deepfake technologies.

Acknowledgements

This research project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme as well as a Ministry of Education (Singapore) Tier 2 grant (T2EP40122-0015).

About the authors

Celene Neo is a final-year Communication Studies student at Nanyang Technological University's Wee Kim Wee School of Communication and Information in Singapore. She can be contacted at cneo018@e.ntu.edu.sg.

Dion H. Goh is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include social media practices and perceptions, game-based techniques for shaping user perceptions and motivating behaviour, and online information sharing and seeking. He can be contacted at ashlgoh@ntu.edu.sg.

Rachel Chun is a Research Associate at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. She is dedicated to extending care and respect to the community through her research and professional endeavours. She can be contacted at rachel.chunwy@ntu.edu.sg.

Chei Sian Lee is a Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University in Singapore. Her research focuses on generative AI and digital nudging, enhancing engagement, decision-making, and learning with an ethical, information-oriented approach. She can be contacted at leecs@ntu.edu.sg.

References

Bray, S. D., Johnson, S. D., Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. Journal of Cybersecurity, 9(1). https://doi.org/10.1093/cybsec/tyad011

El-Gayar, M.M., Abouhawwash, M., Askar, S.S., & Sweidan, S. (2024). A novel approach for detecting deep fake videos using graph neural network. J Big Data 11, 22. https://doi.org/10.1186/s40537-024-00884-y

Goh, D.H., Lee, C.S., Chen, Z., Kuah, X.W., & Pan, Y.L. (2022). Understanding users' deepfake video verification strategies. Communications in Computer and Information Science, 1655. https://doi.org/10.1007/978-3-031-19682-9_4

Goh, D. (2024). He looks very real: Media, knowledge, and search-based strategies for deepfake identification. Journal of the Association for Information Science and Technology; 75(6):643-654. https://doi.org/10.1002/asi.24867

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. Psychological and Cognitive Sciences, 119(1). https://doi.org/10.1073/pnas.2110013119

Groh, M. (2020). Detect DeepFakes: How to counteract misinformation created by AI. MIT Media Lab. https://www.media.mit.edu/projects/detect-fakes/overview/

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520

Jo, Y., Kim, M., Han, K. (2019). How Do Humans Assess the Credibility on Web Blogs: Qualifying and Verifying Human Factors with Machine Learning. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). https://doi.org/10.1145/3290605.3300904

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access, 8, 83144-83154. https://doi.org/10.1109/ACCESS.2020.2988660

Köbis, N. C., Dolezalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. iScience, 24(11), Article 103364. https://doi.org/10.1016/j.isci.2021.103364

Kulangareth, N.V., Kaufman, J., Oreskovic, J., & Fossat, Y. (2024). Investigation of Deepfake Voice Detection Using Speech Pause Patterns: Algorithm Development and Validation. JMIR Biomed Eng, 9. https://doi.org/10.2196/56245

Lee, E. G., Lee, I., & Yoo, S. B. (2023). ClueCatcher: Catching Domain-Wise Independent Clues for Deepfake Detection. Mathematics. 2023; 11(18):3952. https://doi.org/10.3390/math11183952

Mai, K. T., Bray, S., Davies, T., Griffin L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLoS One, 18(8), e0285333. https://doi.org/10.1371/journal.pone.0285333

Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R.O. (2020). Deepfake Detection through Deep Learning. 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT),134 - 143. https://doi.org/10.1109/BDCAT50828.2020.00001

Park, H. J. (2024). The rise of generative artificial intelligence and the threat of fake news and disinformation online: Perspectives from sexual medicine. Investig Clin Urol, 65(3), 199-201. https://doi.org/10.4111/icu.20240015

Petrosyan, A. (2024). Age distribution of internet users worldwide 2024. Statista. https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

Somers, M. (2020). Deepfakes, explained. MIT Sloan School of Management. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Sumsub. (2023). Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023. Sumsub. https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps. Journal of Computer-Mediated Communication, 26(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11): 40-53. http://doi.org/10.22215/timreview/1282

Zeng, R., Song, S., Guo, Z., Goh, D.H., & Lee, C.S. (2023). Real or Fake: Eliciting Deepfake Identification Strategies Through a Diary Study. Proceedings of the Association for Information Science and Technology, 60(1), 1206-1208. https://doi.org/10.1002/pra2.993