SUBJECT GROUP COMPUTER SCIENCE AND INFORMATION SYSTEMS

Research Proposal for an Honours project

The student and the supervisor must consult the *Manual for Postgraduate Studies* before writing the research proposal. The *Manual for Postgraduate Studies* explains in detail what is expected at each of the subheadings below. The proposal should not be longer than 5 pages.

The Subject Group requires that the research proposal will be submitted through the use of this form and in the format below. Please complete using a computer.

1 Student initials, surname, and student number

Initials	N	Surname	Foxcroft	Student number	28418077

2 Degree for which student is registered

BSc Honours in Computer Science and Information Technology

3 Name of supervisor

Initials and surname	Dr R. Serfontein

4 Proposed title

Title (preferably not more than 12 words)	Identifying StyleGAN images

5 Problem statement and substantiation

Provide the theme and link with gaps in the literature and recent research in the area. Indicate the research question, its actuality and how the research will endeavour to answer the question.

StyleGAN is an open-source Generative Adversarial Network (GAN) that can be used to generate faces of people that do not exist (such as those shown on thispersondoesnotexist.com). This means that fraudsters can use StyleGAN generated faces that normally would pass a visual inspection conducted by a human inspector as part of false identities. The detection of such images with the use of artificial intelligence will be useful because of the factors that currently lead to misidentification.

Possible misidentification of StyleGAN images is a reality that needs to be addressed. Humans in the role of identifying artificially generated human faces may be susceptible to external factors hindering their capabilities and increasing the rate of error in which they identify fraudulent images (Fysh & Bindemann, 2018). Fysh and Bindemann also noted that in the specific use case of passport officers that were tested on passport images captured on the same day against the "traveller" presenting that image, that the officers made substantial errors in a controlled environment when comparing the picture identity to that of the traveller. These results enforced their original statement that humans struggle with unfamiliar face identification.

By looking at how the technology has been used since its release, the possible use-cases for GAN generated images and the always growing cybercrime industry the possible detection of these images is identified as a crucial function in the 4th industrial revolution. With these identified factors will the proposed project aim to solve the problem of detecting StyleGAN images by using an artificial intelligence approach to solve the problem.

6 Research aims and objectives

Provide the different general as well as the specific aspects which will form part of the research.

Aims:

The main purpose of the proposed project is to develop a method that can detect fraudulent human faces created by StyleGAN with relative accuracy. Various techniques and approaches to the detection of GAN generated images will be researched and the simplest implemented approach that can still detect these types of images with relative surety will be selected for the artefact.

Objectives:

The success of the project will be weighed against the completion of the secondary objectives that have been identified as listed below.

- Perform a literature study on GANs and specifically analyse the architecture and function of StyleGAN to understand the technology.
- Develop an approach to the successful identification of generated images.
- Develop an artefact that will use the selected method to detect a fake identity.

The successful completion of the above-identified objectives will aid the researcher in satisfying the aim of the proposed research project.

7 Basic hypotheses (where applicable)

The use of Neural Networks will aid in the successful detection of StyleGAN generated images.

8 Method of investigation

8.1 Literature study

Provide an indication only of which literature will be used in the study with key references. A summary of the literature is not required here.

In this modern world with humanity currently in its 4th industrial revolution, the additions of innovative technologies require original approaches to implement and maintain these technologies. The change from the digital age to the automation age is accelerated by the breakthroughs in the fields of artificial intelligence (AI) and security. (Skilton & Hovsepian, 2017)

One of the big advances in artificial intelligence is the creation of StyleGAN, this new approach to a generative adversarial network allowed for more control in an image than its predecessors. (Karras et al., 2019) StyleGAN uses the principles of a GAN to create new images derived from input that specifies what "styles" need to be included in the image. According to Karras et al. (2019) a style is defined in this contexts as a set of parameters that modifies the input of the image to result in different outputs. If the input received is that the image needs to be in the style of a person with glasses, red hair and must be female, StyleGAN will then generate that image based on images used of that similar styles in the initial training of the model. The resulting image will thus be that of the "styles" required in the input. While this functionality can be utilized for various positive use cases – this breakthrough also creates various challenges and setbacks in the field of security, more specifically the aspects of facial recognition and identity verification as malicious use of this GAN might aid in the creation of fraudulent identities. (Mitra *et al.*, 2021)

The big advancement in StyleGAN that led to the conception of this proposed project was the release of StyleGAN2 in February 2020. (Karras, Laine, Aittala*, et al.*, 2020) Following the release of version 2, there were improvements in the generation process of these images with this updated version of StyleGAN. StyleGAN2 saw that images were no longer subverted with artefacts or traces left on the image because of the processing method used in the previous iterations. (Karras, Laine, Aittala*, et al.*, 2020) These traces were easy to identify and clearly showed out of place in the context of the image. With the removal of the traces, usually, in the form of drop-like spots, the difficulty in detecting the generated images increased and similarly the need to detect StyleGAN generated images used maliciously in security verification processes.

8.2 Methods of investigation

The proposed design, data acquisition, procedures, data processing, funding sources (but not a budget), mathematical methods, computer methods.

The positivistic research paradigm is suitable for the proposed project because with the creation of the artefact, the data collected will be examined and an unobjective interpretation of the data is necessary. The data collected will be the results of the artefact's successful identification of StyleGAN generated images.

DSRM is an information systems specific methodology that focuses on research and iterative design. (Peffers *et al.*, 2007) Because the researcher is studying the field and technologies in which they want to solve the specific problem the background knowledge of the problem will be explored parallel to the design of the problem solution. DSRM will enable iterative design and development throughout the completion of the proposed project

The artefact that will be developed will aim to detect StyleGAN generated images between a data set of real images of human faces and StyleGAN generated images. The artefact will be hosted online as this will simplify the development of user interaction. Because of these factors, Agile will be the most suitable methodology for artefact development.

Neural networks and computer vision will be used once an extensive literature study identifies these methods suitable for StyleGAN image detection.

9 Provisional chapter division

Chapter 1: Introduction - This chapter will be concluded in the project proposal phase. It will include the research question, the project description and background. It will include the proposed project plan for the entire project.

Chapter 2: Literature Study - This chapter will be comprised of all the necessary research to understand the project and fulfil the project aims and objectives. Mostly in this project will be focusing on the specific workings of StyleGAN to effectively detect fake images.

Chapter 3: Development of the artefact - Chapter 3 will apply the chosen methodologies that were identified and discussed in Chapter 1 to enable the successful development of the proposed project's artefact. This chapter will document the artefact development phase including unfamiliar problems that are identified within the development stage. The success of the artefact will be compared to the Aims and Objectives of Chapter 1 and if they are met.

Chapter 4: Testing and Results - The results of the Artefact will be introduced in this Chapter and the successful identification of StyleGAN images will be determined. The testing in this Chapter will identify the success of the artefact with the comparison in Chapter 3.

Chapter 5: Conclusion - Chapter 5 will summarize the entire proposed project and conclude if the problem was solved with the successful completion of the objectives of the proposed project that allowed it to fulfil the aim of the project. The limitations that impeded the proposed project will be discussed in this section. Future expansion of the project will be discussed and explained in the context of the limitations faced

10 Literature references

Provide complete references to the literature referenced in this proposal only.

Gallagher, F. and Calabrese, E. (2019).Facebook’s latest takedown has a twist - AI-generated profile pictures. ABC News. 31 December 2019. https://abcnews.go.com/US/facebooks-latest-takedown-twist-ai-generated-profile-pictures/story?id=67925292 Date of access: 24 March 2021.

Fysh, M. C. and Bindemann, M. (2018). Human–computer interaction in face matching.Cognitive Science, 42(5):1714–1732.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture forgenerative adversarial networks.IEEE transactions on pattern analysis and machineintelligence, PP.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzingand improving the image quality of stylegan.2020 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR), pages 8107–8116.

Mitra, A., Mohanty, S. P., Corcoran, P., and Kougianos, E. (2021). A machine learningbased approach for deepfake detection in social media through key video frameextraction.SN Computer Science, 2(2)

Peffers, K. E. N., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). Adesign science research methodology for information systems research.Journal ofManagement Information Systems, 24(3):45–77.

Skilton, M. and Hovsepian, F. (2017).The 4th industrial revolution : responding to theimpact of artificial intelligence on business. Springer International Publishing AG.

Foxcroft

....

Student

18 / 04 / 2021

Date

....

Supervisor

....

Date