学 位 論 文 要 旨

学位申請者：
張重陽

学位論文題目

Personal Photo Privacy Protection by Attacking Face Detectors

As artificial intelligence becomes increasingly integrated into people's daily lives, researchers are increasingly concerned with how to use these technologies safely and responsibly. The malicious face-swapping behavior that this thesis focuses on is a serious problem of personal privacy security being violated due to the indiscriminate use of artificial intelligence technology, and it is also one of the hardest hit areas in computer vision security issues. The social significance of this thesis is to prevent the occurrence of such infringements, and the decision to face-swapping should be in the hands of the owner of the photo. This thesis analyzes the current popular face swap program, faceswap, and finds the weak points in its workflow to carry out research.

Facewswap is a complete face swap software, which includes the core algorithm of face-swapping and the entire face-swapping pipeline from face detection to face fusion. This thesis believes face detection should be the focus of research in this face-swapping pipeline. If the face detector loses its effectiveness, not only the face detection link cannot provide an effective face area for the face-swapping algorithm, but it can also play a certain protective role in other dangerous fields of face recognition and detection. Therefore, the specific purpose of this thesis is to make the MTCNN, SSD, and S3FD face detectors used in the face detection link in faceswap lose their effectiveness and fail to detect faces in images.

Based on previous research methods, the image style transfer technique is employed to enhance the probability of detecting the background portion in an image, and the change of MTCNN facial detection result is observed. It is confirmed that the addition of perturbations to the background under these conditions does not impact the MTCNN facial detection result.

Existing adversarial attacks can successfully attack SSD-based face detectors in the digital domain, and since S3FD is also based on SSD architecture, it can also be broken. However, there is no research on using invisible perturbation to attack MTCNN in the digital domain. The first network in the MTCNN cascade network, P-Net, uses a fully convolutional neural network that can process image data at multiple scales at the same time. We think this is why MTCNN is challenging to break because it can process multi-scale data simultaneously and aggregate the results. According to the characteristics of multi-scale input data, experiment proposes an attack method that directly interpolates and fuses perturbations. In the experiment, P-Net is used as the feedback judgment network, and the attack method is as follows.

Attack area. The experiment uses the detection area roughly detected by p-net as the range of adding perturbation to reduce the pollution to the image. The detection frames' position, number, and size may change after each attack, making our method more accurate than other attack methods.

Attack Steps. The experiment searches for effective perturbations using the form of a double loop. The first layer loops images of various sizes and the second layer loops to find the effective perturbation at the current size. In the second layer loops, we keep adjusting the size of the perturbations until we find a valid perturbation.

Loss Function. The experiment sums the confidences of the detection boxes detected in the second-layer loop in proportion. It takes the negative sum as the loss value to reduce confidence continuously.

Search Rate. MTCNN normalizes the image value to between -1 and 1 during image preprocessing. The

gradient values obtained by sign are -1 and 1, so we choose to design the search rate as a ratio of the original value of the image—for example, 1/255, 0.1/255.

This method is able to add perturbations to images of different sizes and complete the attack on MTCNN. The CelebA data test's detection success rate dropped to 23.7%. It is also the first time that MTCNN has been successfully attacked using invisible perturbations in the digital domain. Also, for the first time in the research of attacking MTCNN with invisible perturbations, a baseline for image quality assessment based on CelebA data is proposed. The baseline scores for PSNR, SSIM, and LPIPS image quality assessment algorithms are 30.25, 0.9, and 0.06, respectively.

Lastly, a method of disrupting feature extraction continuity by adding black lines to the face is proposed. The attack ability is enhanced as the width of the line increases. Quantitative experiments were conducted using MTCNN, SSD, and S3FD, when the line width was 10 pixels, the detection rates in CelebA data dropped to 6.15%, 9.47%, and 32.1%, respectively; when the line width was 8 pixels, the detection rates in FFHQ data dropped to 7.2%, 4.3%, and 9%, respectively. Due to the reduced usability caused by the coverage of facial features by black lines, this study conducted optimization experiments on the structure and image quality of this method. According to the results of the structure optimization experiment, it is difficult to optimize the black line structure manually. Short line structures can expose the eyes and corners of the mouth, which can be optimized to some extent, but the usability is still affected. In the image quality optimization experiment, random perturbations were added to the coordinates between the two points of the line, and the coverage range of the perturbations was reduced to below 45%. This experiment takes images where the face is completely covered by black lines as baseline images for image quality evaluation. Based on the PSNR, SSIM, and LPIPS image quality assessment methods, the scores on the CelebA data are 9.27, 0.18, and 0.21 higher than the baseline of this experiment, and the scores on the FFHQ data are 8.56, 0.3, and 0.33 higher than the baseline. Combining the questionnaire survey on user usage requirements, the generated adversarial examples using this method have a 34.2% higher acceptance rate compared to the baseline images. Based on the results of the aforementioned studies, it can be concluded that our method is capable of providing assistance to users who have the need for facial privacy security.

Summary

Applicant for degree:
Zhang Chongyang

Title of thesis :
Personal Photo Privacy Protection by Attacking Face Detectors

As artificial intelligence becomes increasingly integrated into people's daily lives, researchers are increasingly concerned with how to use these technologies safely and responsibly. Malicious face swapping is caused by the abuse of AI technology and seriously endangers individuals' privacy security. Based on previous research and the workflow of the popular face-swapping program, Faceswap, this study argues that attacking the facial detection stage is an effective solution to this issue. However, there is currently no attack method that can simultaneously make MTCNN, SSD, and S3FD provided by Faceswap ineffective. This study presents a solution, with the specific work content summarized as follows.

In Part One, based on previous research methods, the image style transfer technique is employed to enhance the probability of detecting the background portion in an image, and the change of MTCNN facial detection result is observed. It is confirmed that the addition of perturbations to the background under these conditions does not impact the MTCNN facial detection result. The second part, a method of attacking the MTCNN facial detector with invisible perturbations is proposed for the first time, and the reasons for its difficulty in being attacked are explained. Based on the feature of using image pyramids to process input images, an attack method that fuses perturbations of multiple scales is constructed. This is a white-box attack method, with a success rate of 76.3% in quantitative testing on CelebA data. Also, for the first time in the research of attacking MTCNN with invisible perturbations, a baseline for image quality assessment based on CelebA data is proposed. The baseline scores for PSNR, SSIM, and LPIPS image quality assessment algorithms are 30.25, 0.9, and 0.06, respectively.

In the third part, a method of disrupting feature extraction continuity by adding black lines to the face is proposed. The attack ability is enhanced as the width of the line increases. Quantitative experiments were conducted using MTCNN, SSD, and S3FD, when the line width was 10 pixels, the detection rates in CelebA data dropped to 6.15%, 9.47%, and 32.1%, respectively; when the line width was 8 pixels, the detection rates in FFHQ data dropped to 7.2%, 4.3%, and 9%, respectively. Due to the reduced usability caused by the coverage of facial features by black lines, this study conducted optimization experiments on the structure and image quality of this method. According to the results of the structure optimization experiment, it is difficult to optimize the black line structure manually. Short line structures can expose the eyes and corners of the mouth, which can be optimized to some extent, but the usability is still affected. In the image quality optimization experiment, random perturbations were added to the coordinates between the two points of the line, and the coverage range of the perturbations was reduced to below 45%. This experiment takes images where the face is completely covered by black lines as baseline images for image quality evaluation. Based on the PSNR, SSIM, and LPIPS image quality assessment methods, the scores on the CelebA data are 9.27, 0.18, and 0.21 higher than the baseline of this experiment, and the scores on the FFHQ data are 8.56, 0.3, and 0.33 higher than the baseline. Combining the questionnaire survey on user usage requirements, the generated adversarial examples using this method have a 34.2% higher acceptance rate compared to the baseline images. Based on the results of the aforementioned studies, it can be concluded that our method is capable of providing assistance to users who have the need for facial privacy security.