Emotion-Preserved Face Recognition with Privacy Protection

Yinxin Wan **Fengyu Yan**

Weihai Shen **Tianbo Song** **Libin Liu**

Abstract

Face recognition is becoming an increasingly important topic in the area of robotic, computer vision, and artificial intelligence. A typical application of face recognition is to identify and label all of the human faces in a given image or video clip. However, the rapid improvement and widespread deployment of this emerging technology also raise a lot of privacy concerns. In this project we propose a strongly privacy-enhanced face recognition system, which achieves effective privacy protection, while still preserves the important human emotion information. When processing images or videos, we fist locate the human faces and draw corresponding bounding boxes in them. Then we run the emotion classification algorithm and replace the original private face area with certain emojis. Our scheme allows fast and robust operation on both images and videos, and can achieve real-time human face identification and replacement. Our implementation result shows the efficiency and high-accuracy of the proposed system. We also build a website to provide a lively video demonstration and easy-to-use interface.

1 Introduction

Over the past years, the rapid development of computer vision and robotics has enabled accurate and efficient extraction of humans' biometric patterns, e.g. faces, gestures, and emotions. Such important information can then be used in many application scenarios such as movie production, public security, and user authentication. We have already witnessed many benefits that face recognition system brought to us as many countries now apply this technique to enhance current surveillance system so that any potential threats to public security can be identified immediately. Some airports in also utilize face recognition to achieve more secure border control. Of course, the roots of those widespread applications lie in the development of more powerful deep learning techniques such as DNN [1], CNN [2, 3], and Residual Networks [4] to handle the input images or videos more efficiently.

Actually, face recognition is ubiquitously used in our daily lives as it has already been underlying embedded in a lot of applications such as cameras, social networking platforms, and automated driving system. Most images uploading system now also can automatically detect and tag faces.

While the widespread deployment of face recognition systems brings us a lot of convenience, it also raises serious privacy risks. Since face patterns can also be used for authentication uses, and can even be used by attackers to track a person's personal life. One major concern is that such biometric information can be maliciously collected and manipulated to profile targeted users against their will. Those security and privacy issues call for a universal platform to handle the privacy problem which can be easily deployed without bringing to much overhead and cost [5].

Although the privacy issue in face recognition has been raised and studied for a long time [6, 7, 5], those approaches either relies on heavy cryptographic operations, which consume a lot of com-

putation resources and bring unbearable latency, or can hardly be directly attached to current face recognition systems. For example, some of the proposed schemes are based on the concept of secure multiparty computation [8], and utilize techniques like combinatorial circuits [9], ordered binary decision diagrams [10], or branching programs [11]. However, these methods tend to be impractical because of the high computational complexity, especially in the condition that we are trying to deal with complicate image or video data. In [5], it leverages hmomographic encryption scheme to achieve calculations on encrypted data. Specifically, Pailler [12] protocol is used, while again it suffers from the low efficiency problem mentioned above.

In this project we propose a strongly privacy-enhanced face recognition system, which can achieve high quality of service just as the normal face recognition system. In addition, we take users' privacy into consideration, and replace the human faces with the corresponding emojis. With the help of emotion classification technique based on deep learning, our scheme can choose the candidate emoji with high accuracy. The proposed system can thus assure the privacy of individuals in scenarios where face recognition is beneficial for society but also has privacy concerns.

In particular, our approach is a combination of different techniques such as face recognition, face alignment, and emotion classification. For any given images or video clips, we first identify and locate all of the human faces in it. After that we turn the captured face into grey scale images. With the help of VGG-16, we can further transform the image data into vectors. We turn into the public dataset for face images labeled with corresponding emotions. After training and evaluating the model, we eventually have a accurate emotion classifier which can be applied directly to our scheme.

To better demonstrate the whole system and provide a easy-to-use platform, we also build a website which can be accessed by everyone under the ASU network environment, or first connect to ASU's VPN (Virtual Private Network). Users can choose from a variety of demo images, upload a new image, or paste a URL. We also provide a interface to perform the online operation on real-time camera videos. We show the efficiency and accuracy of this part in our demonstration video. Our experiment results show that our system is both efficient and effective.

The main contributions in our work are as follows:

- We proposed a privacy enhanced system for face recognition;
- We combine face recognition and emotion classification techniques together to achieve privacy while still preserve the facial emotion;
- The experiment results prove the efficiency and accuracy of our system.

The remaining of this report is organized as follows. Section 2 gives a detailed description of our approach, in which we discuss both the perception module and control module. In Section 3 we present how the experiment is set up and the experimental results. We conclude this project and give some discussion of it in Section 4.

2 Approach

Our proposed scheme can roughly be divided into two independent sub-systems. When a image or video clip is set as input, we fist perform face recognition to identify the human faces using the pre-trained model [13]. Then we run facial expression classification algorithm to identify the correct emotion of the selected human faces. In the last step, we choose the corresponding emoji to replace the human faces in the original images or video clips.

In the next two sub-sections, we will present the detailed design of our scheme from perspectives of perception module and control module.

2.1 Perception Module

When our system receives an input image or video clip, it first make a gray-scale copy of it. Then it utilizes pre-trained face alignment model to get 68 2D facial landmarks, which can generally represent the main features of a given human face.

Figure 1: 2D Facial Landmarks.

As we can see from Figure 1, from the 68 facial landmarks we can capture most of the important gestures and boundaries of a human face. For example the shapes of noise, eyes, and mouth are preserved precisely when we add some connection lines between adjacent landmarks. Although the face alignment algorithm can also be applied to detect 3D facial landmarks, we only demonstrate 2D example here since most of the input images are taken from a single camera.

Those 68 landmarks then can be used to help us to determine a bounding box which cover the area of the whole human faces. The contents in the bounding box is what we are intended to protect and should be replaced by a corresponding emoji in the next step. The later process is a classical classification problem, since we want to determine the best match of the emotion in human face from a set of candidates.

To correctly decide the emotion of the person in the picture, we applied the CNN as the training model and used real-world dataset of high quality. Our model can distinguishes from emotions like anger, disgust, fear, happy, neutral, sad, and surprise. The training dataset is categorize into seven subset accordingly and to ease the problem of over-fitting and make best use of the whole images, we run cross evaluation by using the 'train_test_split' in the 'scikit-learn' package.

Our CNN model is described as follows. The input is 42×42 gray-scale images stored in a '.csv' file. The kernel's size in first convolution layer is 5×5, which means we select a smaller area of size 5×5 every time and then move one step to its neighbors after padding the original image to a $46 \times 46 \times 1$ size. Now each image have a larger size of $42 \times 42 \times 32$ but is later reduced to $21 \times 21 \times 32$ after the first pooling layer. The second convolution layer is a little bit different as its kernel size is now only 4 and the padding size is 1, so the image reduced to 10×32 after another polling layer. The third convolution and pooling layer are the same as the first ones. Now we get a 5×64 vector, but it needs to be further processed by two fully-connected layers. Finally, it reduced to the the size of $1 \times 1 \times 7$, which corresponds to seven kinds of emotions we want to classify. To achieve high-speed and efficient training, we also applied BN (batch normalization) after each layer.

Active function is what we use to get the output of node, which is also known asthe Transfer Function. The Sigmoid Function's curve looks like a S-shape. This function isdifferentiable, which means that we can find the slope of the sigmoid curve at any two points. The function ismonotonicbut functions derivative is not. Tanh is like logistic Sigmoid as it also has a 'S' shape but the range of the tanh function is from -1 to 1. The ReLU is the most widely used activation function because

it is easier to computer and converges very fast. Taking these issues into consideration, we choose ReLU in the end and it actually performs quite very and help us train the model more efficiently.

2.2 Control Module

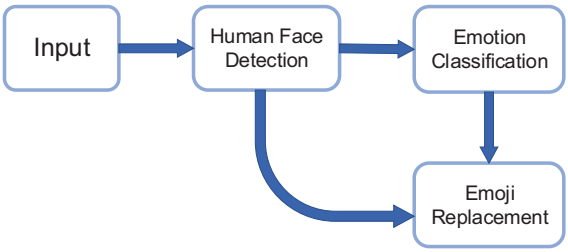

Figure 2: System Overview.

As show in Figure 2, our system mainly consists of three parts, namely face detection, emotion classification, and emoji replacement. Firstly, we use the camera to capture a image or video stream, then save it for the down-streaming detection and classification. The face detector locates the face, by using x, y, w, h to record the face location in each image, i.e. the position, width and height of the bounding box. The original image and these variables are the input of the classifier. The classifier output the the possibility values of each kind of emotion, which can help the system to make a decision. Finally, we use the corresponding emoji to replace the original face area. The modified image will be showed in the window. The system is based on the OpenCV video caption architecture.

3 Experiments

3.1 Data Collection

The training dataset for human face detection is from [13, 14, 15, 16, 17], which is all pre-labeled and large enough for our task. The training dataset for the emotion classification is from [18] and Kaggle challenge on Facial Expression Recognition. To better demonstrate our results, the test data can be captured in real-time by a web-camera or any images on the Internet with a link. The dataset contains pre-labeled images of seven different emotion classes, so our model can recognize seven emotions.

3.2 ROS System Setup

In the ROS system, we have to first start the ros core and create a workspace for this project. We run the 'catkin_make' in our workspace, and it will create a CMakeLists.txt link in our 'src' folder for us. Our system have two nodes, the listener and the talker. For the talker node, we apply the camera caption in that part and then transfer the video shots to the listener for further processing. Which means the listener is the more important node here as it need to perform almost all the face recognition and emotions classification operations mentioned above. Our whole system is written in python and requires only acceptable computation and communication overhead once the model has been trained, which means it not only works under ROS, but can also directly applied in many other applications directly. Users can easily use our system as we also build a website for it.

3.3 Experimental Results

In the face recognition part, we choose NME (Normalized Mean Error) as the metric used for face alignment. NME is defined as follows:

$$NME = \frac{1}{N}\sum_{k=1}^{N}\frac{\|\boldsymbol{x_k} - \boldsymbol{y_k}\|_2}{d},$$

where $\boldsymbol{x}$ denotes the ground truth landmarks for a given face, $\boldsymbol{y}$ is the corresponding prediction and d is the squareroot of the ground truth bounding box, computed as $d = \sqrt{wbbox * hbbox}$. Our test results show that we can achieve a NME score as high as 72.1% on the given dataset. Which means our face recognition sub-system can accurately identify the human faces in given images or video clips in most cases.

For the emotion classification part, we used the CNN model described before and set the total number of epoch as 50. The overall accuracy is 67.8%, which is acceptable as there are seven kinds of emotions in tota to categorized.

(a) Neutral (b) Neutral Emoji Replacement

(c) Happy (d) Happy Emoji Replacement

Figure 3: Experiment Results

As shown in Figure3, our system performs quite well and can accurately classify the emotions and replace the original human faces with the corresponding emojis in most cases. With the help of a web-camera and a server, we can run the whole procedure in almost real-time when the input the file is either a image or a video clip directly captured from the camera.

4 Discussion and Conclusion

Face recognition technique is increasingly deployed in civilian, robotic, and artificial intelligence applications. However, the privacy issue is always an important topic that we cannot take a detour. In this paper, we propose a system which takes an input image or video clip, then identify the humans faces in it, and replace it with the emojis which have the corresponding emotions. Our scheme is efficient that it can perform all the operations mentioned before in real-time after finishing training the model. The experiments and the demonstration video shows the practicality of our solution.

However, there are still some problems that we are intended to solve in our future work. This first thing is that we currently can only classify seven kinds of emotions, which seems reasonable but still cannot cover all the possible cases. Another things is that our system chooses emoji as the replacement of the origin human faces. However, the image or the video we get after being processed is not ideal as it looks some kind of weird. One possible solution is to change the human face in a unnoticeable way, which means it still looks normal but cannot be used to identify the people any more. We plan to further study this part and make our system more efficient and more easy to use.

Acknowledgment

We would like to thank Prof. Yezhou Yang and the teaching assistants for they kind support and generous help. We are highly indebted to them as without their guidance and constant supervision as well as for providing necessary information, we cannot finish this project by ourselves.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems (NIPS)*, 2015, pp. 91–99.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 1440–1448.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *International symposium on privacy enhancing technologies symposium (PET)*. Springer, 2009, pp. 235–253.

[6] S. Avidan and M. Butman, "Efficient methods for privacy preserving face detection," in *Advances in neural information processing systems (NIPS)*, 2007, pp. 57–64.

[7] A.-R. Sadeghi, T. Schneider, and I. Wehrenberg, "Efficient privacy-preserving face recognition," in *International Conference on Information Security and Cryptology (ICISC)*. Springer, 2009, pp. 229–244.

[8] A. C.-C. Yao, "Protocols for secure computations," in *FOCS*, vol. 82, 1982, pp. 160–164.

[9] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game," in *Proceedings of the nineteenth annual ACM symposium on Theory of computing (STOC)*. ACM, 1987, pp. 218–229.

[10] L. Kruger, S. Jha, E.-J. Goh, and D. Boneh, "Secure function evaluation with ordered binary decision diagrams," in *Proceedings of the 13th ACM conference on Computer and communications security (CCS)*. ACM, 2006, pp. 410–420.

[11] M. Naor and K. Nissim, "Communication preserving protocols for secure function evaluation," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC)*. ACM, 2001, pp. 590–599.

[12] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Advances in Cryptology–EUROCRYPT'99*. Springer, 1999, p. 223.

[13] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[14] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proceedings the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 2144–2151.

[15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings the 2013 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2013, pp. 397–403.

[16] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings the 2015 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2015, pp. 50–58.

[17] Jain, Vidit and Learned-Miller, Erik, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," *UMass Amherst Technical Report*, p. , 2010.

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2016, pp. 94:1–94:4.