AnonEmoFace: Emotion Preserving Facial Anonymization

Jan Hintz[a], Jacob Rühe[b] and Ingo Siegert[c]
Mobile Dialog Systems, Otto-von-Guericke University, Magdeburg, Germany

Keywords: Anonymization, Privacy, Face Recognition, Emotion Recognition, Human Recognition, Human-Machine-Interaction.

Abstract: Seeking therapy often implies a major hurdle, especially when it comes to addressing personal problems that cause shame or are socially stigmatized. This is where the recent developments of remote therapy can help. To further reduce this barrier, it can be accommodating to carry out the therapy anonymously. This paper present a proof of concept for such an anonymization of remote therapy video calls. The aim is to enable video calls for subjects without the risk of being identified by their face. The challenge lies in the contradiction of preserving emotional content and successful anonymization. To achieve this goal, avatarization by facial landmark detection is employed. In a user study with 30 participants we achieved an unweighted average recall of 48.6% for facial recognition task, confirming anonymity, while preserving emotional expressivity with an accuracy of 93.3% for happiness, 68.3% for fear, 50.05% for anger and 35.5% for disgust. Thus creating a safe environment for the user, while preserving emotional content for therapeutic purposes.

1 INTRODUCTION

There are many conditions for which people seek professional help but are reluctant or even afraid to receive it because of the fear of social stigmatization (Owen et al., 2013). With the growing need for psychological treatment, and the digitalization of the health sector, remote therapy rises as a new opportunity, making therapy more accessible and creating a protected space. Text-based electronic therapy for at-risk patients has already been successfully applied by Quackenbusch et al. (Quackenbush and Krasner, 2012). The method allows patients to disclose themselves without risking identification. The therapy was conducted in a virtual office where patients could talk to a therapist through an Avatar on text-basis. While this approach grants high anonymity, it is also not as natural as a face-to-face conversation (Böck et al., 2012).

The aim of this paper is to provide a proof of concept for visual anonymization while preserving the emotional content. The objective is to implement anonymization techniques to safeguard the identity of individuals who contact a helpline seeking remote assessment, treatment, and prevention services related to child sexual abuse. Given the sensitive nature of the subject and the need for help-seekers to maintain a high level of anonymity while disclosing personal inclinations and preferences, the implementation of anonymization measures on the client side (user terminal) becomes imperative. Our approach aims for natural face-to-face interaction while granting anonymity, with a focus on emotion preservation in facial expression.

2 RELATED WORK

There exists a variety of techniques to anonymize visual data, ranging from applying a bar covering the eyes (Roberts et al., 2016), blurring or pixelating (Fan, 2019) to applying generative models (Sun et al., 2018; Hellmann et al., 2023). All of these approaches have different advantages and use cases (Siegert et al., 2020).

The problem with these approaches is that they can be reverted (Cichowski and Czyzewski, 2011). Therefore, modern approaches use either some form of random dithering, making it impossible to obtain the facial expression, or face-swapping, needing high computing power but can result in unnatural faces having an impact on the conversation. To address the aforementioned challenges, we suggest employ-

[a] https://orcid.org/0000-0003-2910-8202
[b] https://orcid.org/0009-0004-5293-9105
[c] https://orcid.org/0000-0001-7447-7141

Hintz, J., Rühe, J. and Siegert, I.
AnonEmoFace: Emotion Preserving Facial Anonymization.
DOI: 10.5220/0012410600003648

In *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, pages 785-788
ISBN: 978-989-758-683-5; ISSN: 2184-4356

ing a technique known as avatarization of the patient's face. This involves tracking facial landmarks and transferring them onto an avatar (e.g., Apple's "Memoji"). It facilitates the identification of facial expressions while it is not seen as unnatural during conversation. To the best of our knowledge, this form of facial motion retargeting has been researched extensively (Aneja et al., 2018; Chaudhuri et al., 2019), but has not been validated in an anonymization scenario with human interaction.

3 PROPOSED METHOD

In the following, we elaborate on the processing pipeline of our approach. We aim to produce a representation that meets two primary criteria. First and foremost, it is imperative that the Video does not contain any identifiable visual information, in order to guarantee the privacy and anonymity of the individuals involved. The second requirement is that, the generated representation should contain all essential information required for emotional assessment. By fulfilling these two requirements, our approach aims to strike a balance between privacy preservation and maintaining facial expressions for therapy purposes. Facial landmarks have been proven to fulfill the second criteria in multiple studies (Aneja et al., 2018; Gupta, 2018; Hellmann et al., 2023; Ko, 2018; Nguyen et al., 2017; Sun et al., 2018). (Sun et al., 2018) and (Hellmann et al., 2023) have also proven that landmark-based, generative approaches grant high anonymity.

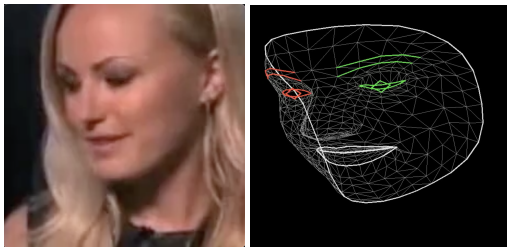

Figure 1: Video anonymization (left original taken from Voxceleb2 (Chung et al., 2018), right anonymized, cropped).

We are using Mediapipe's face-mesh model (Kartynnik et al., 2019), extracting 468 3-dimensional Landmarks. To ensure that any additional information that could exhibit a possible indirect identification is disregarded, the extracted points are then plotted onto a 2-dimensional blank canvas (720×1280). The result can be seen in Figure 1

To allow seamless integration into streaming platforms and video conference tools, we generate an OBS virtual camera by using the pyvirtualcam library. In the later phase of the project, this will be integrated into a stand-alone virtual camera environment for easy use.

4 USER STUDY

A user study with 30 participants was conducted to evaluate subjective verification and emotion preservation in anonymized data. The test dataset consist of six individuals, with three recordings each for anonymization purposes where the person is moving their head from left to right and two recordings each for the 4 acted emotions (happiness, anger, fear, and disgust). The participants were presented a video of a person next to an anonymized video. The objective was to verify the identity (True/False) and rate the confidence in this decision (11-point scale). In doing so, we tested whether it is possible to identify the original person from the anonymized videos. In the following, participants were given the task to label a video with emotion. As baseline, the raters had to assess non-anonymized recordings of acted emotions before evaluating the anonymized material. As before, the participants also rated their confidence.

5 RESULTS

5.1 Face Verification

If the anonymized and non-anonymized faces belonged to the same individual, the unweighted average recall (UAR) for face recognition was 30%. However, if the IDs did not match, the UAR increased to 67.3%. This resulted in 48.6% UAR overall, in a two-class problem this is as good as random chance which implies high anonymity.

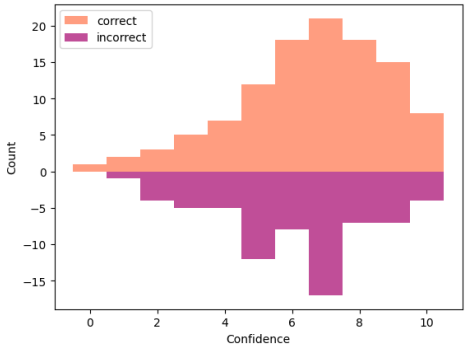

Figure 2: Bi-histogram of participants' confidence by correct/incorrect match of ground truth with anonymized data.

The stated confidence did subtly reflect the participants' performance, with an average confidence score of 6.59 in case of correct verification results, and 6.14 for incorrect verification results. The bi-histogram in Figure 2 visualizes this as shift of the histogram of incorrect matches with the ground truth, displayed on the negative y-axis.

5.2 Emotion Recognition

The human raters were capable of recognizing all acted emotions in the non-anonymized scenario, with happiness (99.4%) and fear (94%) being correctly recognized by the majority, while disgust (93.8%) and anger (81.6%) were slightly more difficult to distinguish. The average confidence score was 8.56.

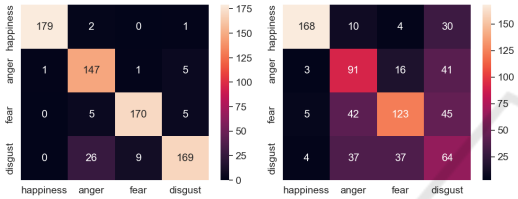

Figure 3: Confusion plot non-anonymized (left) and anonymized (right) emotion recognition.

In case of the anonymized stimuli, happiness (93.3%) and fear (68.3%) were recognized correctly by most, while anger (50.05%) and disgust (35.5%) were more difficult to discriminate from the other emotions. This can also be observed by comparing the confusion plot over the tested emotions, shown in Figure 3. The average confidence score was 6.14.

The combined bi-histogram shown in Figure 4 demonstrates also reflects the average confidence scores and overall recognition rates.

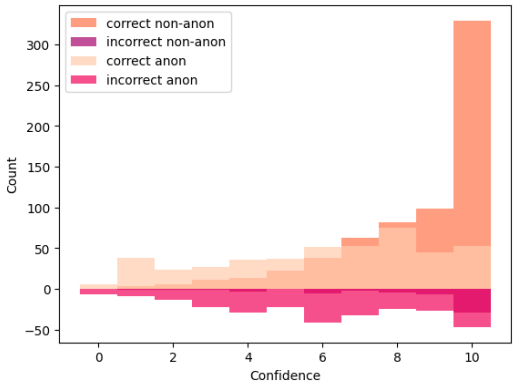

Figure 4: Combined bi-histogram of participants' confidence by correct/incorrect match of ground truth with emotional non-anonymized and anonymized data.

6 LIMITATIONS

The conducted study has certain limitations. The generated dataset is relatively small and lacks sufficient variety that would be required for a broader study. The evaluation also lacks extension by machine recognition to accurately assess whether the anonymization is also robust in this regard, as human recognition can be outperformed (O'Toole et al., 2012). This was disregarded as the current concept aims for human to human communication in a streaming setup.

Furthermore, the dataset only includes four emotions, performed by people without acting experience. Therefore, the study lacks subtle, more nuanced emotions that might be valuable for therapy.

7 CONCLUSIONS

The presented method of avatarization has proven to anonymize sufficiently while still preserving facial expressivity. The participants of our preliminary study were unable to accurately verify the identity of an individual, therefore proving anonymity. Similarly to the findings of Aneja et al. (Aneja et al., 2018) some emotions seem to be more difficult for humans to both perceive and act out. The effect on perception can also be observed in our findings, as anger and disgust are sometimes confused with other emotions within the non-anonymized data. The impact is even more notable with anonymized data. This could be due to missing properties of the anonymization method, as the mask does not transfer the wrinkles of frowning as well as mouth movement during smiling. Furthermore, the anonymization also has a notable influence on the overall confidence of the participants in their decisions.

Future works will include extended experiments with deeper and broader evaluation of emotion transfer and anonymity, addressing the limitations mentioned in Section 6.

REFERENCES

Aneja, D., Chaudhuri, B., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2018). Learning to generate 3d stylized character expressions from humans. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 160–169. IEEE.

Böck, Limbrecht, Siegert, Glüge, Walter, and Wendemuth (2012). Combining mimic and prosodic analyses for user disposition classification. In Wolff, M., editor, *Elektronische Sprachsignalverarbeitung 2012*.

Tagungsband der 23. Konferenz, volume 64 of *Studientexte zur Sprachkommunikation*, pages 220–227, Cottbus, Germany. TUDpress.

Chaudhuri, B., Vesdapunt, N., and Wang, B. (2019). Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Cichowski, J. and Czyzewski, A. (2011). Reversible video stream anonymization for video surveillance systems based on pixels relocation and watermarking. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1971–1977. IEEE.

Fan, L. (2019). Practical image obfuscation with provable privacy. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 784–789.

Gupta, S. (2018). Facial emotion recognition in real-time and static images. In *2018 2nd international conference on inventive systems and control (ICISC)*, pages 553–560. IEEE.

Hellmann, F., Mertes, S., Benouis, M., Hustinx, A., Hsieh, T.-C., Conati, C., Krawitz, P., and André, E. (2023). Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *arXiv preprint arXiv:2305.02143*.

Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

Nguyen, B. T., Trinh, M. H., Phan, T. V., and Nguyen, H. D. (2017). An efficient real-time emotion detection using camera and facial landmarks. In *2017 seventh international conference on information science and technology (ICIST)*, pages 251–255. IEEE.

O'Toole, A. J., An, X., Dunlop, J., Natu, V., and Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)*, 9(4):1–13.

Owen, J., Thomas, L., and Rodolfa, E. (2013). Stigma for seeking therapy: Self-stigma, social stigma, and therapeutic processes. *The Counseling Psychologist*, 41(6):857–880.

Quackenbush, D. M. and Krasner, A. (2012). Avatar therapy: where technology, symbols, culture, and connection collide. *Journal of Psychiatric Practice®*, 18(6):451–459.

Roberts, E. A., Troiano, C., and Spiegel, J. H. (2016). Standardization of guidelines for patient photograph deidentification. *Annals of plastic surgery*, 76(6):611–614.

Siegert, I., Varod, V. S., Carmi, N., and Kamocki, P. (2020). Personal data protection and academia: GDPR issues and multi-modal data-collections "in the wild". *The Online Journal of Applied Knowledge Management: OJAKM*, 8:16 – 31.

Sun, Q., Ma, L., Oh, S. J., Van Gool, L., Schiele, B., and Fritz, M. (2018). Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059.