Real-Time Reflection Reduction from Glasses in Videoconferences

Marc-André Tucholke* Marie Christoph*
Lasse Anders* Raven Ochlich*
TU Braunschweig
Mühlenpfordtstr. 23
38106, Braunschweig, Germany
{m.tucholke, marie.christoph, l.anders, r.ochlich}@tu-bs.de

Steve Grogorick Martin Eisemann
TU Braunschweig
Mühlenpfordtstr. 23
38106, Braunschweig, Germany
{grogorick, eisemann}@cg.cs.tu-bs.de

ABSTRACT

Surrounding lighting conditions cannot always be sufficiently controlled during videoconferences, yielding situations in which disturbing reflections might appear on the participants glasses. In this article, we present a retrained neural network to convincingly reduce such reflections. For real time performance we propose an asynchronous processing pipeline accompanied by a head pose-based caching strategy to reuse intermediate processing results. The implementation as virtual webcam allows the system to be used with arbitrary videoconferencing systems.

Keywords
reflection, glasses, video conferencing, image processing, deep learning, face detection, real time

1 INTRODUCTION

In the last years video conferences have undergone a huge rise in usage and popularity. Wearers of glasses often experience reflections in their glasses that distract their counterparts or could reveal sensible information. The aim of this work is a reduction of these reflections in real time. To achieve this we integrate an existing neural network for reflection removal, that is not real-time capable, in a real-time context.

Existing techniques for reflection reduction [LLY+23] from a single input image are currently still far from being real-time capable, often reporting processing times of approximately 400 ms. To remedy this, we propose to reduce the computational load by extracting and processing only the relevant part (glasses) of each frame, and propagating the results to subsequent frames.

We detect the region of interest using a learning-based face detection. The segment of the image that contains the glasses is then processed asynchronously by the reflection removal network. Based on the current head pose a reflection mask is applied to the current frame. Through this optimization the processing time per image is reduced to under 40 milliseconds on commodity hardware.

In Section 2 we introduce the neural network that is used for the actual reflection reduction and differentiate our approach from other methods. We present the goals, approach, and realization of the components of our method in Section 3. An evaluation of the methods performance on multiple metrics is presented in Section 4 before concluding in Section 6.

2 RELATED WORK

Reflection removal has drawn attention during recent years, especially in the field of deep learning [AST+22]. Reflection-aware guidance (RAGNet) [LLY+23] is a neural network to remove reflections from glass surfaces in real or synthetic images of fully occluded objects or persons behind a glass panel. The task of reflection removal under these circumstances is similar to the presented task of reflection removal from spectacle lenses, but not identical. The main difference is the partial occlusion of the object and the curvature of the spectacle lenses. The V-DESIRR network [PSB+21] surpasses RAGNet in terms of reflection removal quality and inference time but both target solely reflections on plain glass. Neither the data set nor the code of the V-DESIRR network have yet been made available to the public, preventing its use in any subsequent research. Another promising approach was shown by Wan et. al. in [WSL+21] by removing reflections from images containing partially occluded persons behind a glass panel. Their approach focused solely on faces and incorporated specific facial priors.

* Authors contributed equally

https://www.doi.org/10.24132/CSRN.3301.20

The task is quite similar to the presented task, but the missing open source implementation is again preventing its application in research. Besides single-image reflection removal, multi image methods exist such as [LCL21]. Those methods are not applicable to our problem as we assume input from a single webcam.

The presented work differs from the aforementioned works by focusing on the real-time aspect and the curved surface of glasses.

3 METHOD

In this paper we investigate whether an existing reflection removal algorithm can be adapted to reduce reflections on a user's glasses in real-time. For this purpose we make use of RAGNet [LLY$^+$23] an open-sourced current state-of-the-art reflection removal method.

As input we assume a simple RGB image stream from a 30 Hz live stream or input video of size 1920×1080 pixels. The output is a video or a video stream (virtual webcam) with a maximum resolution of 1920×1080 pixels. We further assume, that there is only one person in the image and the person is in focus and decently illuminated.

3.1 Overview

In this section we give an overview of our technique. The flowchart in Figure 1 shows the per-frame steps of our proposed procedure, separated in two asynchronous threads. The main thread reads the input image and computes the position of significant features in the persons' face, usually referred to as facial landmarks. If glasses are detected, the section of the image that contains the glasses (further referred to as glass-section) is extracted and scaled to a fixed resolution.

The glass-section and landmarks are fed to the the side thread. There, the RAGNet generates two output images: the reflection map and the reflection reduced image. The output images as well as the landmarks are stored in a cache. The RAGNet distorts the original colors of the image and a color correction has to be applied prior to the storage process [RAGS01].

The main thread detects motion relative to the previous frame. If no motion is recognized, the previously detected reflection mask is reused. If, otherwise, motion exceeds a certain threshold, the cache is searched for previous results of a similar pose. If a matching pose is found, the cached reflection mask is warped to fit the current input image and is then applied to it. In favor of a real time frame rate, the frame is left unchanged if no matching pose was found in the cache.

For evaluation later on, we also implemented a synchronous mode running the RAGNet on each frame without motion detection and cache. Because of the long processing time of the RAGNet this mode is not real-time capable by itself.

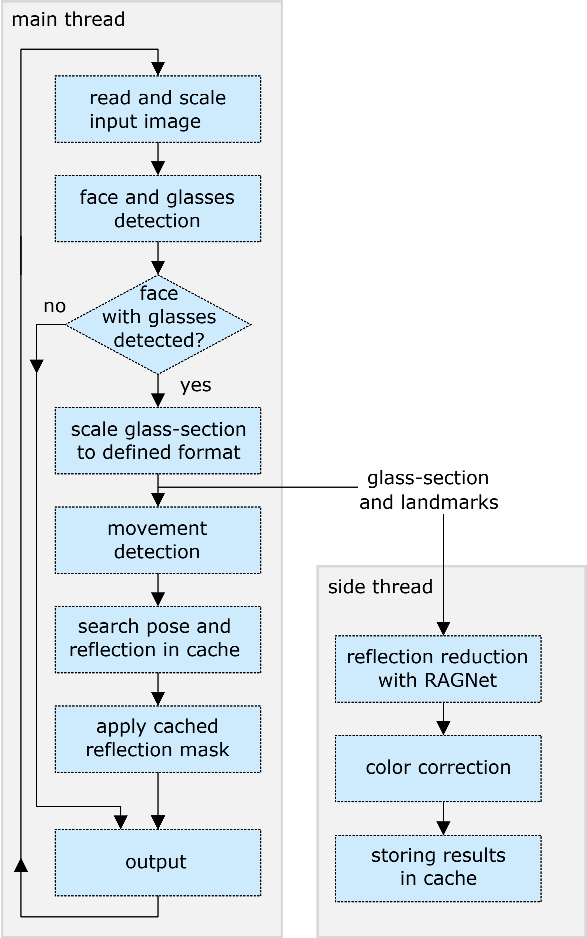
main thread
read and scale input image
face and glasses detection
face with glasses detected?
no
yes
scale glass-section to defined format
glass-section and landmarks
movement detection
search pose and reflection in cache
apply cached reflection mask
output
side thread
reflection reduction with RAGNet
color correction
storing results in cache

Figure 1: Asynchronous processing pipeline

3.2 RAGNet

The neural network RAGNet follows a two step approach [LLY$^+$23]. In the first step the network computes a reflection mask. This mask together with the original image form the input for the second step, where the reflection reduced image is generated. One major task of the second step is to “hallucinate” the content of the image regions where the brightness of the reflection is clipped by the image format limits, i.e., in overexposed regions. This behavior can be seen in Figure 2.

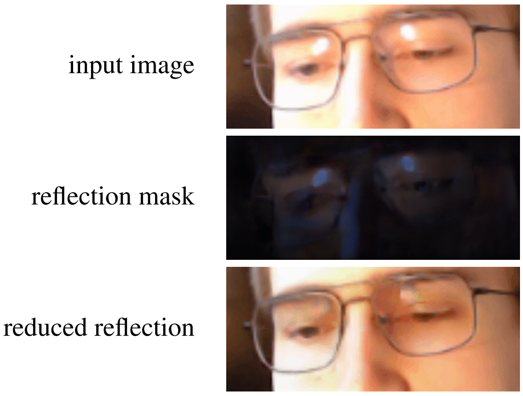

Figure 2: RAGNet hallucinates overexposed regions

3.3 Glasses detection

As applying RAGNet is computationally costly, we extract the region-of-interest containing the glasses and restrict further processing to this region only. The decision if glasses are present in the computed face is based on the presence of edges, i.e., frames of glasses, in three image regions: below each and in-between the eyes (see Figure 3, right). These facial regions, identified based on the landmarks [JBAB00, Tia19, Sid21].

The face detection is realized using DLIB [Kin09], a well established library that robustly handles variations in pose or illumination. Specifically, the landmarks shown in Figure 3 (left) are acquired using the DLIB facial landmark detector [SAT+16, KS14]. The computation speed of the DLIB algorithms can be improved by executing them on the graphics card using the CUDA toolkit [NVI]. The performance can further be improved by downscaling the input image. We empirically chose the resolution (1280 × 720 pixels) such that the rate of correctly identified facial landmarks is nearly equivalent and subsequent computations are not compromised.

The boundaries of the glass-section are determined by the bounding box of the landmarks around the eyes. This yields a robust, accurate and fast detection of the glass-section (see Section 4).

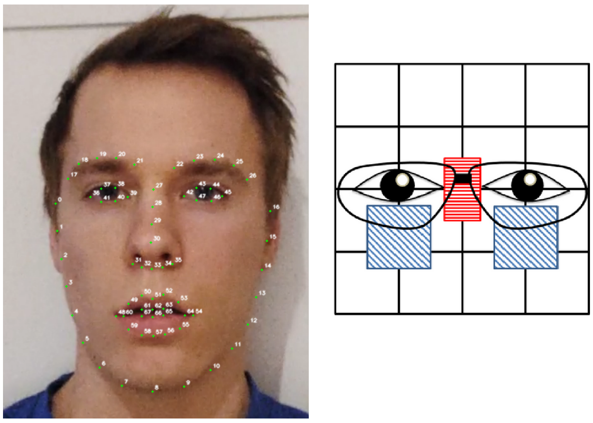

Figure 3: 68 facial landmarks detected with DLIB (left) and examined facial areas according to [Tia19] (right)

3.4 Refined RAGNet

The following sections describe our task-specific transfer learning to optimize RAGNet towards glasses.

3.4.1 *Pretrained weights*

The pretrained RAGNet [LLY+23] removes reflections from images where the content is fully covered by a glass plane. We found that the quality of the reflection removal is still acceptable for our scenario, where only a small part of the image is covered by glass, but the computation times are far from real-time, even when focusing on the glass-section only.

The processing time as well as the quality of the reflection removal depends strongly on the size of the image. We, therefore, scaled the glass-section to 711 × 300 pixels to achieve stable yet satisfying results.

3.4.2 *Recording of and training with synthetic data*

To improve the performance of RAGNet we additionally trained it with synthesized training data that specifically resembles our use case more closely than the original training data, i.e. persons with glasses. The synthesizing process to create the training data was similar as proposed in [FYH+17].

To train the RAGNet three images are needed, one that remarks the ground truth and has no reflections in it, one that represents the reflections in the image (reflection mask) and the last one that is the original image with the reflections in it (see Figure 4). The unprocessed training frames are extracted from a reflection free video. The glass-sections are then extracted as described in Section 3.3 and rescaled to the demanded size. The reflections are randomly selected from a set of handcrafted prerecorded reflection templates. Within reasonable limits, these templates are randomly scaled, rotated, and intensity-adjusted. In the last step the reflections are added to the ground truth and the edges are smoothed with a Gaussian filter.

With this algorithm a dataset holding 3000 entries was created and RAGNet was trained for 15 epochs with the setup recommended by Li et al. [LLY+23]. The validation was done based on the original loss function and the mean peak signal-to-noise ratio (PSNR) on 20 validation data set entries. The training after 15 epochs resulted in improved results, as depicted in Figure 5.

The results of RAGNet trained on the synthesized data sets did not perform well on real test data. This behavior was expected as it was observed by the authors of RAGNet too when they used synthetic data originally [FYH+17]. This could be due to overfitting or too unrealistic artifacts. To resolve this problem the training set was extended with real reflection images as follows.

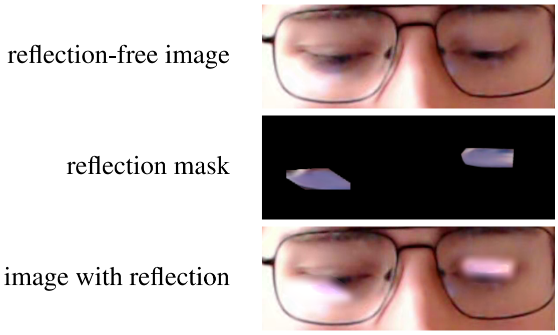

Figure 4: Example of the synthetic training data

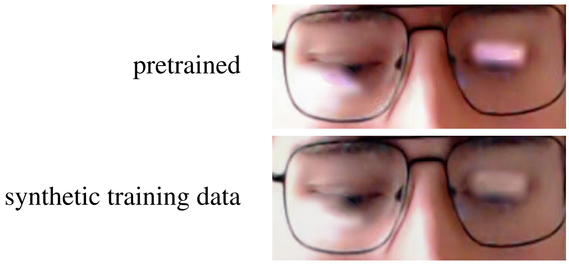

Figure 5: RAGNet original and retrained (synthetic)

3.4.3 *Recording of stroboscope data*

To generate real test data including images with and without reflection for the same head pose, we used the following setup. A person in front of a PC screen watched a program that alternated a full screen output between plain white and black. During each state, an image of the person was acquired using a webcam. The time between the state changes was chosen so that it enables the light to set and the camera to produce a stable picture but also short enough so that the person's head won't move significantly. As previously mentioned the training needs a third image per data set. The image representing the reflection is calculated by subtracting the reflection free image from the image with reflections. A set of the stroboscope images is displayed in Figure 6.

To create viable test data it has to be assured, that the room where the images are recorded doesn't contain additional reflection sources. The person the images are taken of is ideally illuminated from above and no background light is disturbing the image. Otherwise the whole face would be brighter if the white light of the screen is turned on. If the light from above is too bright the reflections on the glasses would not be significant enough to be seen.

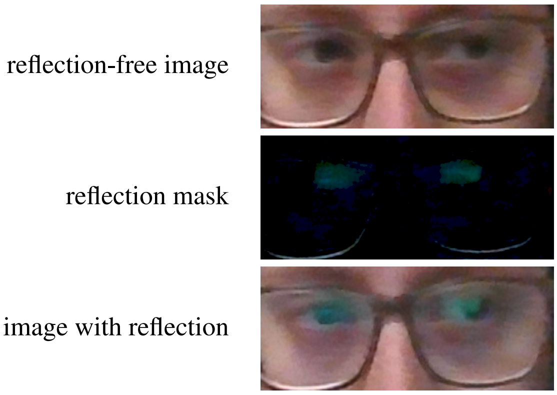

Figure 6: Example of the stroboscope training data

3.4.4 *Training with synthetic and stroboscope data*

To improve the results of RAGNet the data set was extended by 1095 stroboscope entries and additionally 903 synthetic entries. The data was divided into 60% training, 20% validation and 20% test data, resulting in a training set with 2997 entries.

With this dataset the pretrained RAGNet was refined in the following three steps.

First, the network was further trained for 55 epochs until the loss started to converge. Second, to prevent training towards a local minimum, for which the first step of the RAGNet produced a empty reflection masks, we trained 30 epochs using a modified loss function that included the reflection mask only, until the RAGNet produced plausible reflection masks. Third, to mitigate errors in the reflection free images, the network was trained until convergence for additional 70 epochs with the original loss function, to finish the joint optimization of reflection mask and reflection reduced image generation.

After retraining, RAGNet produced plausible reflection masks and eliminated reflections better than the original version of the network when applied to images of faces with glasses. The inference for a single entry of the validation set is shown in Figure 7.

Since the reflection reduced images show a shift in their color distribution, we extend their processing with an appropriate color correction [RAGS01].

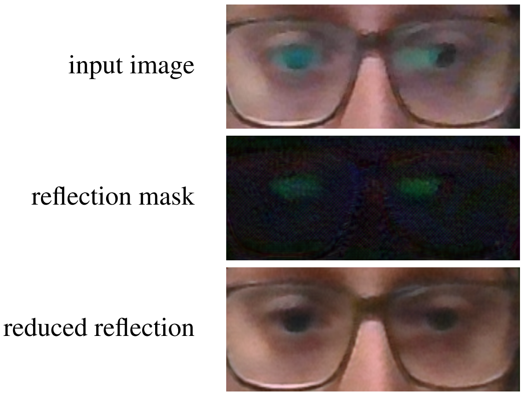

Figure 7: RAGNet retrained (synthetic + stroboscope)

3.5 Motion detection

To reduce temporal artifacts of the cached reflection reduction, that become most noticeable during small head motions, we perform a motion detection, to directly reuse the previous detection in these cases. Furthermore, the motion detection reduces the frequency of cache searches.

To detect motion, i.e., changes between successive images, we compare the current frame with its predecessor via the structural similarity index measure (SSIM) [WBSS04]. SSIM was chosen because it offers to compute similarity only for the brightness of two images, and reflections almost always affect image brightness.

A threshold of 0.95 for the SSIM score was empirically identified to give reliable results.

We further improve performance by computing the motion detection only for the eye region (see Figure 8). The position and size of the region-of-interest around the eyes is again computed based on the facial landmarks, including a certain margin around the eyes to allow keeping the bounds unchanged for the SSIM computation during slight head movements. It is automatically updated only if the eyes reach the current bounds.

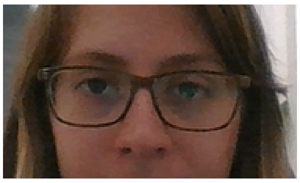

Figure 8: Section around eyes for motion detection

3.6 Asynchronous processing

Even for the size-reduced glass-section the RAGNet takes 190 ms to process a single frame. Therefore, we decided to move the RAGNet processing to a separate thread. The main thread supplies the RAGNet thread with the current frame and respective landmarks, as shown in Figure 1. The RAGNet side thread then processes the frame asynchronously and generates the reflection reduced image and the reflection mask. The reflection reduced image is then color corrected. The input images, output images and facial landmarks are stored in the cache using a ring buffer scheme.

3.7 Pose-based cache search

For each input frame the main thread searches for a fitting similar frame in the cache. The selection is based on similarity of the facial landmarks of the current and the cached frames. The search can be executed in 5 ms for 50 cache elements using this approach.

Excluding mouth and eyes, 35 out of the 68 facial landmarks are used for per-frame head pose encoding using a 2-column matrix, storing one position (x,y) per row (Equation 1). The dissimilarity d_i between the ith cached element's facial landmark matrix M_i and the current landmark matrix $M_{current}$ is determined by the Frobenius norm F of their difference (Equation 2). The cached element with the smallest d_i is selected, if it is below the threshold t_{norm} (Equation 3), which is an image size-normalized threshold with user-defined parameter t. A value of $t = 15$ was empirically found to yield a good trade-off between cache hit rate and visually pleasing output.

$$M = \begin{pmatrix} y_1 & x_1 \\ \vdots & \vdots \\ y_{35} & x_{35} \end{pmatrix} \tag{1}$$

$$d_i = F(M_{current} - M_i) \tag{2}$$

$$t_{norm} = t \cdot \frac{\text{image width}}{1000} \tag{3}$$

3.8 Cache-based reflection reduction

Mitigating the expensive, thus slow execution of RAGNet, cached results of preceding frames that were computed by RAGNet already, are now employed to reduce reflections on the current input image. As described above, we retrieve the data of the cached frame that is most similar to the current frame in terms of the detected head pose, assuming no significant changes in the background of the videoconference feed, i.e., the user's surrounding. To reduce the reflections in the current input image, we use the cached reflection mask, i.e., the difference between the cached input image and the cached reflection reduced image.

To account for slight differences between the head pose of the current and the cached frame, the reflection mask needs to be adjusted accordingly. To this end, the following three approaches for reflection mask adjustment were tested.

1. Homography transformation based on four facial landmarks: the outermost points along the eyebrows (17, 26) and the lower left and right parts of the chin (6, 10).
2. Affine transformation based on three facial landmarks: the outermost points along the eyebrows (17, 26) and the lowest point along the contour of the face at the middle of the chin (8).
3. Correlation: Application of the cached reflection mask at the location of highest correlation (normalized mean shifted cross correlation) between the glass-section of the cached frame and the current input image.

Since reflections are no fixed parts of glasses, but may instead change their position on the surface as the head moves, they do not necessarily move uniformly with the glasses. Thus, head motions may still result in some artifacts in the form of brightness mismatches along the edges of the cached reflection mask.

From the three tested approaches, correlation leads the fewest artifacts and is therefore suggested to be used by default. An exemplary result of the reflection reduction using correlation is shown in Figure 9.

4 EVALUATION

In the following we evaluate the different components of our proposed method.

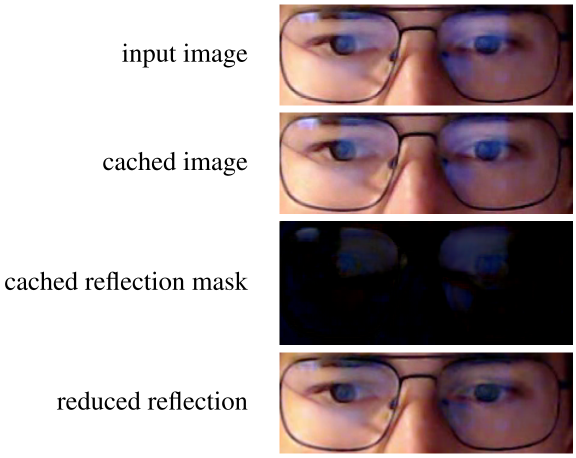

Figure 9: Cache-based reflection reduction

4.1 Glass-section detection

The face detection is evaluated on a test dataset of 2400 images. The test images originate from the recorded webcam streams of four different web meeting participants. The quality of the face detection is evaluated for scaled and unscaled input images in Table 1. The accuracy does hardly degrade for scaled images and is sufficiently high.

	Unscaled	**Scaled**
Scale factor	1	0,3125
TP[1]	2372 / 2400	2369 / 2400
Accuracy	98.83%	98.71%

Table 1: Face detection accuracy on scaled images

The glasses detection was optimized on a subset of the dataset. The quality of the algorithm was evaluated on the rest of the dataset. As shown in Table 2, the proposed glasses detection method performs similar to the reference method by Fernández et al. [FGUC15]. It should be noted, however, that different data sets were used for validation, because there is no reference implementation available for the comparison method.

	Training set	**Test set**	**Reference method** [FGUC15]
TP	392 / 397	1935 / 1971	2959 / 3000
FP[1]	0	2	-
FN[1]	5	34	-
Accuracy	98.74%	98.17%	98,65%

Table 2: Glasses detection accuracy

Since there is no reference glass-section cropping algorithm, our approach is validated against a previously manually selected image region. We use the excess image area and the missing image area relative to the manually selected area as metrics to determine the quality of the automatic glass-section cropping algorithm. The results are listed in Table 3.

	Training set	**Test set**
Images	400	2000
Excess area	14,21%	13,79%
Missing area	8,56%	9,55%

Table 3: Automatic eye region cropping

The overall relative error area is sufficiently small for the subsequent processing stages. The sum of errors is only sligthly increasing from the training to the test set. This implies that the algorithm shows a good generalization and should be applicable to new previously unseen images.

4.2 RAGNet performance

The different retrained instances of the network are evaluated quantitatively by comparing their average PSNR and SSIM [HZ10]. The disjoint test data set consists of 1000 mixed stroboscopic and synthetic images. Table 4 shows both metrics for the generated reflection reduced output images.

Training set	**Epochs**	**PSNR**	**SSIM**
RAGNet original	150	15.24	0.731
Synthetic data	15	23.80	0.880
Synthetic + stroboscope data	55	28.86	0.935
Reflection-only + joint training	30/70	27.39	0.937

Table 4: Performance per training strategy

The network instance with reflection-only pre-training followed by full training, has the best average SSIM score and reaches the second highest average PSNR value. It is the only network computing a meaningful reflection mask for our scenario.

The network robustly detects reflections on the constrained input data and convincingly reduces reflections on single images. Even though very bright (clipped to white) reflections in input images result in visible artifacts, their appearance is still reduced noticeably.

Given the RAGNet (in synchronous mode) would run fast enough, it is only evaluated on individual frames without incorporating temporal consistency, resulting in a noticeable flickering. This directs towards future

[1] TP: True Positives, FP: False Positives, FN: False Negatives

research on, e.g., temporal low-pass filtering the reflection mask output or extending the RAGNet architecture to include recurrent layers for temporal context.

The reflection reduction works good on the constrained data set of similar test data. The model has problems generalizing on unseen footage. This limitation could clearly be overcome with a more diverse training set.

4.3 Motion detection

The motion detection was subjectively tested for plausibility. The estimated SSIM index correlates well with the present amount of motion, i.e., the SSIM index reaches values near one for non-moving persons.

4.4 Asynchronous processing

The temporal coherence and overall reflection removal quality was verified subjectively. The asynchronous processing introduces some additional flickering to the resulting video stream, caused by remaining differences between cached and current frames. While mismatches due to peoples motion are limited to small offsets via motion detection, changes in lighting are implicitly compensated over time due to the ring-buffered cache.

Regarding the reflection mask adjustment, the homography approach yields mediocre results. Even with careful selected landmarks there were some clearly visible remaining artifacts when applying the transformed cached reflection mask. Restricting the degrees of freedom by using affine transformations resulted in more consistent and therefore more pleasing results. Best results were achieved using the correlation approach, restricting the applied transformation even more, yielding the temporally most consistent results. This resulted in an overall visually more pleasing perception.

4.5 Execution time

The real time requirements require a strict optimization of the different components. All performance tests were performed using input videos with a resolution of 1920 × 1080 pixels on a system with an NVIDIA GTX 960 and an AMD RADEON VEGA 56. The glasses detection is running on the former while the RAGNet is running on the latter.

The largest performance improvement could be achieved by executing the RAGNet asynchronously. The reduction of the input resolution for the glasses detection and the movement detection resulted in further performance improvements. The mean execution times for processing a single frame, averaged over 200 frames, is displayed in Table 5. Finally, applying some common optimizations throughout the pipeline, such as reducing the number of image copy operations, yielded a final frame rate of 31.25 Hz.

Optimizations	**Execution time [ms]**
Synchronous	410
Asynchronous	199
Asynchronous & scaled	38
Further optimization	32

Table 5: Average processing time per frame

The composition of the frame processing time for a single exemplary frame is shown in Table 6.

Processing Step	**Execution time [ms]**
Read Frame	1
Glasses detection	20
Motion detection	6
Cache search	4
Transfer to current frame	2

Table 6: Processing time per system component

5 DISCUSSION

The model generally strongly depends on the input video stream. The best results are achieved under good lighting conditions and for reflections in the upper half of the glasses.

Generalizability. While using comparatively small data sets, like in this work, typically implies little generalizability, building upon the far more diversely pretrained RAGNet mitigates this weak point for our approach. It should therefore also be possible to also reduce reflections from light sources other than screens, e.g. ceiling lamps, and even under different lighting situations.

Moreover, the restriction of the processing to the glass-section should further increase generalizability, as the network does not need to learn (to ignore) arbitrary environments.

Limitations. Reflections which directly occlude the eyes sometimes result in worse reflection removal performance with stronger artifacts, as shown in Figure 10.

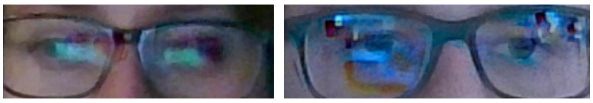

Figure 10: Artifacts for reflections covering the eyes

While strong variations in illumination will most likely not break the approach, they might reduce the effectiveness, resulting in, e.g., brightness or color mismatches.

Since both limitations arise from the limited data set, it is reasonable to assume that the proposed system can overcome them by extending the training to a larger and more diverse data set, which we leave for future work with a focus on robustness.

https://www.doi.org/10.24132/CSRN.3301.20

6 CONCLUSION

This paper presents an approach to reduce reflections on glasses in real-time. We showed that the RAGNet neural network can be arranged in an appropriate pipeline to convincingly reduce reflections on glasses. For application in live videoconference scenarios, we achieved real-time capability by reducing the network input size using the newly introduced glass-section detection and the proposed asynchronous processing scheme. Moreover, temporal consistency is strengthened via robust motion detection and color transfer.

While the goal of complete reflection removal was not achieved, the synchronous mode would result in visually more pleasing reflection removal on selected inputs, but is not real-time capable. The real-time capable asynchronous mode introduces some artifacts and flickering. Furthermore, some aspects of the implementation still offer potential for improvement, e.g., for multiple persons or handling of the remaining error cases, such as reflections largely occluding the eyes. The method is currently still very resource demanding, motivating further optimization, e.g., via motion compensation for cached frames.

Also beyond videoconferences, the proposed method could be a helpful tool for preprocessing videos in applications that use eye tracking or emotion analysis. The method could also be reduced in scope, to be used as a reflection detection.

7 ACKNOWLEDGMENTS

Partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491805996 / GR 5932/1-1.

8 REFERENCES

[AST⁺22] Amanlou, A., Suratgar, A. A., Tavoosi, J., Mohammadzadeh, A., and Mosavi, A. Single-image reflection removal using deep learning: A systematic review. *IEEE Access*, 2022.

[FGUC15] Fernández, A., García, R., Usamentiaga, R., and Casado, R. Glasses detection on real images based on robust alignment. *Machine Vision and Applications*, 26(4):519–531, May 2015.

[FYH⁺17] Fan, Q., Yang, J., Hua, G., Chen, B., and Wipf, D. A generic deep architecture for single image reflection removal and image smoothing. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3258–3267, 2017.

[HZ10] Hore, A. and Ziou, D. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[JBAB00] Jiang, X., Binkert, M., Achermann, B., and Bunke, H. Towards detection of glasses in facial images. *Pattern Analysis & Applications*, 3:9–18, 2000.

[Kin09] King, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. http://dlib.net.

[KS14] Kazemi, V. and Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In *IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[LCL21] Li, T., Chan, Y.-H., and Lun, D. P. K. Improved Multiple-Image-Based Reflection Removal Algorithm Using Deep Neural Networks. *IEEE Transactions on Image Processing*, 30:68–79, 2021.

[LLY⁺23] Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., and Zuo, W. Two-stage single image reflection removal with reflection-aware guidance. *Applied Intelligence*, pages 1–16, 2023. https://github.com/liyucs/RAGNet.

[NVI] NVIDIA Corporation. CUDA. https://developer.nvidia.com/cuda-toolkit. Visited on 19.01.2022.

[PSB⁺21] Prasad, B. H. P., S, G. R. K., Boregowda, L. R., Mitra, K., and Chowdhury, S. V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2390–2399, October 2021.

[RAGS01] Reinhard, E., Ashikhmin, M., Gooch, B., and Shirley, P. Color Transfer between Images. *IEEE Computer Graphics and Applications*, 21:34–41, October 2001.

[SAT⁺16] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[Sid21] Siddharth Mandgi. Real-time Glasses Detection. https://medium.com/mlearning-ai/glasses-detection-opencv-dlib-bf4cd50856da, September 2021. Visited on 19.01.2022.

[Tia19] Tianxing Wu. Real-time Glasses Detection. https://github.com/TianxingWu/realtime-glasses-detection, November 2019. Visited on 19.01.2022.

[WBSS04] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[WSL⁺21] Wan, R., Shi, B., Li, H., Duan, L.-Y., and Kot, A. C. Face Image Reflection Removal. *International Journal of Computer Vision*, 129(2):385–399, February 2021.

https://www.doi.org/10.24132/CSRN.3301.20