Temporally smooth privacy-protected airborne videos

Omair Sarwar[1,2], Andrea Cavallaro[2] and Bernhard Rinner[1]

***Abstract*— Recreational videography from small drones can capture bystanders who may be uncomfortable about appearing in those videos. Existing privacy filters, such as scrambling and hopping blur, address this issue through de-identification but generate temporal distortions that manifest themselves as flicker. To address this problem, we present a robust spatio-temporal hopping blur filter that protects privacy through de-identification of face regions. The proposed filter is meant for on-board installation and produces temporally smooth and pleasant videos. We apply hopping blur to protect each frame against identification attacks, and minimise artefacts and flicker introduced by the hopping blur. We evaluate the proposed filter against different identification attacks and by assessing the quality of the resulting videos using a subjective test and objective measures.**

I. INTRODUCTION

Recreational videography may capture faces, licence plates, windows of private houses and may therefore lead to discomfort or privacy concerns. To address this problem, privacy filters are used to modify the appearance of privacy-sensitive image regions [1]–[6]. For example, the appearance of a captured face can be modified in order to conceal the identity of the person (see Fig. 1).

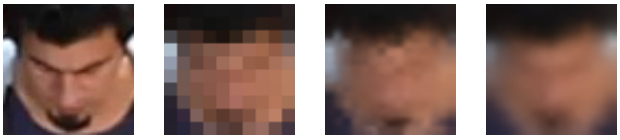
A privacy filter should cause only a minimal spatio-temporal distortion. However, filters such as scrambling [3] and hopping blur may generate abrupt changes in the intensity values of consecutive frames thus resulting in unpleasant flicker. A privacy-protected video should also prevent person identification under different attacks, such as naïve and parrot attacks. Naïve attacks compare unprotected gallery faces against privacy-filtered probe faces, whereas parrot attacks de-identify both gallery and probe faces. In addition to the above, naïve-SR attacks first restore (e.g. with super-resolution (SR) [7]) filtered probe faces and then compare them against unprotected gallery faces, and parrot-SR attacks filter both gallery and probe faces and restore them with super-resolution before comparing them against each other.

O. Sarwar was supported by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environment, funded by the Education, Audiovisual & Culture Executive Agency under the FPA no 2010-0015; and in part by Intelligent Vision Austria. A. Cavallaro also acknowledges the support of the UK EPSRC project NCNR (EP/R02572X/1).

[1,2]Omair Sarwar is with the Institute of Networked and Embedded Systems, Alpen-Adria-Universität Klagenfurt, Austria, and the Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom omair.sarwar@aau.at

[2]Andrea Cavallaro is with the Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom a.cavallaro@qmul.ac.uk

[1]Bernhard Rinner is with the Institute of Networked and Embedded Systems, Alpen-Adria-Universität Klagenfurt, Austria bernhard.rinner@aau.at



(a) (b) (c) (d)

Fig. 1: A face image de-identified (protected) with different privacy filters. (a) Original image (crop). Image protected with (b) pixelation, (c) hopping blur; and (d) proposed filter.

A privacy filter can be static or dynamic. Static filters keep their parameters, such as the standard deviation of a Gaussian blur, spatially and temporally constant [1], [6]. Static filters protect against naïve attacks, but are prone to parrot [2] and reconstruction attacks, such as naïve-SR and parrot-SR. Dynamic filters change their parameters spatially and/or temporally [3], [8] and protect faces against parrot and reconstruction attacks. However, they may introduce flicker.

Flicker-reduction approaches were developed for video compression [9]–[14] and can be applied prior to, during or after encoding [11], [12]. Approaches to be applied prior to [15] and during encoding [10]–[12] are coder-specific. Approaches to be applied after encoding measure the spatio-temporal correlation between frames [9], [13], [14] and are generic. However, these approaches cannot be applied to our scenario as correlation is compromised by privacy filters that use scrambling [3] and warping [8]. Therefore, an alternative solution for minimising flicker is needed.

In this paper, we present a privacy-preserving filter for drone videos that addresses the trade-off between privacy, fidelity and temporal smoothness. To the best of our knowledge, this is first time that flicker reduction is considered for a privacy filter. The proposed filter minimises spatio-temporal distortions and is robust against naïve, parrot, naïve-SR and parrot-SR attacks. Depending on the resolution of the captured face, the parameters of an Adaptive Hopping Gaussian Mixture Model (AHGMM) filter are adjusted according to the target spatial distortion and are then mixed with decaying weights to minimise flicker.

II. PROBLEM DEFINITION

We aim to robustly protect a face with minimal spatial and temporal distortions, and to prevent various identification attacks.

Let R_m be a privacy-sensitive region, such as a face, in frame I_m. Let $\bar{R}_m$ be the corresponding privacy-protected region generated with filter $F_{\Omega_i^*}$, which uses as parameters $\Omega_i^* \in \{\Omega_0, \Omega_1, \Omega_2, ...\}$. The larger the index, the stronger the distortion introduced in the privacy-protected region.

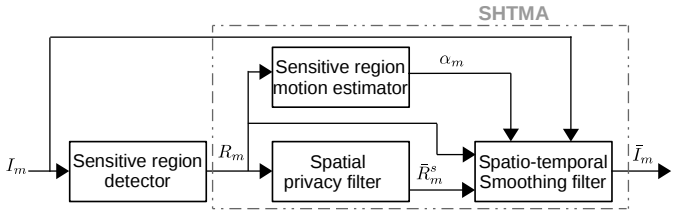


Fig. 2: Block diagram of the proposed Spatial Hopping Temporal Moving Average (SHTMA) filter that reduces flicker in privacy-preserving videography.

We can split our overall aim for the privacy filter into two concurrent and competing objectives. The first objective is that the privacy-protection filtering

$$\bar{R}_m = F_{\Omega_i^*}(R_m) \tag{1}$$

should ensure a minimal spatio-temporal distortion. The second objective is that $\bar{R}_m$ should be protected against de-identification attacks: when the region $\bar{R}_m$ corresponds to a face, the probability $P(.)$ of recognising the identity of the person should be no better, under various attacks, than that of a random classifier. Therefore, if $\mathcal{G}$ is an unprotected, filtered or reconstructed gallery data set of K subjects, then

$$P(\bar{R}_m|\mathcal{G}) \to 1/K, \tag{2}$$

where $1/K$ is the identification accuracy of a random classifier.

III. Flicker-Free Seamless Protection

To achieve the aim defined in Sec. II, we design the Spatial Hoping Temporal Moving-Average (SHTMA) filter (Fig. 2). Each frame I_m is processed by a sensitive-region detector (a face detector in our case), which returns one or more R_m. We assume the bounding boxes of faces to be available[1] and that navigation sensors of the airborne camera measure its height, h_m, and tilt angle, θ_m. These parameters are used to estimate pixel densities ρ_j (px/cm) of the captured face [6], where $j \in \{h, v\}$ indicates horizontal and vertical direction.

R_m is first protected using a spatial privacy filter, which generates a protected region $\bar{R}^s_m$. We choose *hopping blur* as the spatial privacy filter because of its robustness to attacks. This robustness is achieved through pseudo-random switching of the Gaussian kernels for different sub-regions of a face. The hopping hinders the estimation of the Gaussian kernel parameters from the filtered sub-regions, thus making face recognition difficult even under parrot and parrot-SR attacks.

The selected filter parameter for *hopping blur* is $\Omega_i^* = (0, \sigma_j^*)$, where σ_j^* is the standard deviation, estimated as [6]:

$$\sigma_j^* = \frac{3\rho_j}{\pi\rho_j^*}, \tag{3}$$

[1]The filter relies on a face detector (and tracker) whose design is beyond the scope of this paper.

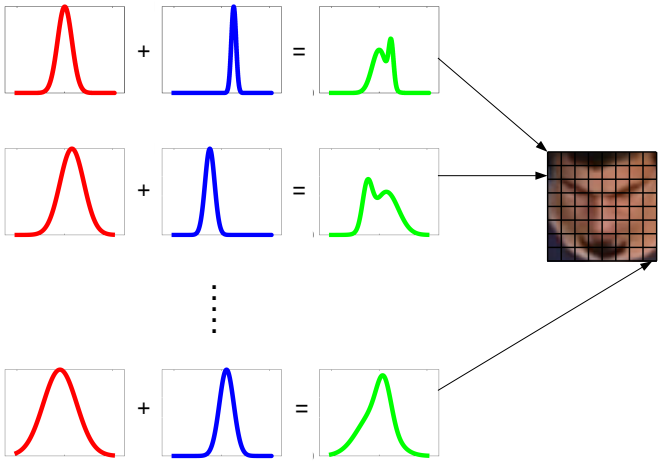


Fig. 3: One dimensional representation of the hopping blur composed by superimposing a selected (red) and a supplementary Gaussian function (blue). Both functions pseudo-randomly change their parameters among the sub-regions thus resulting in a hopping Gaussian mixture model (green).

and ρ_j^* is the threshold of pixel density at which a state-of-the-art classifier starts recognising faces better than a random classifier.

We define a Gaussian mixture model (GMM) by combining the selected Gaussian function with another Gaussian and pseudo-randomly change the GMM parameters (mean, standard deviation and weight) in different sub-regions of R_m. This hopping GMM (see Fig. 3) is convolved with R_m, and the filtered sub-regions are then globally smoothed in order to reduce blocking artefacts.

The hopping GMM for different sub-regions of a frame introduces flicker as the models change independently from frame to frame. Moreover, directly replacing R_m in I_m with the protected face region $\bar{R}^s_m$ may introduce strong boundary effects.

To mitigate these problems and to generate temporally smooth (and protected) face regions, we blend the internal boundary of $\bar{R}^s_m$ and low-pass filter it as:

$$\bar{R}_m = \alpha_m[\alpha_s\bar{R}^s_m + (1-\alpha_s)R_m] + (1-\alpha_m)[\alpha_s\bar{R}^s_{m-1} + (1-\alpha_s)R_{m-1}], \tag{4}$$

where $\alpha_s \in [0,1]$ and $\alpha_m \in [0,1]$ are a spatial and temporal weight, respectively. As a constant value of α_s blends $\bar{R}^s_m$ and R_m, but does not remove the sharp boundary between $\bar{R}_m$ and $\bar{I}_m$, we decrease α_s moving away from the boundary of the region. Moreover, a lower value of α_m increases smoothness but may introduce unpleasant delays in the video when the person moves. For this reason, to balance smoothness and delay, we adaptively select α_m depending on the motion of the face region, which is measured as displacement of the centres of R_m and R_{m-1}.

IV. Experiments

Methods. We compare the proposed SHTMA filter with (i) AHGMM, a non-space-variant Gaussian blur that uses

hopping Gaussian kernels; (ii) Adaptive Gaussian Blur (AGB) [6], a space-invariant Gaussian blur that uses a single Gaussian kernel[2]; and (iii) Space Variant Gaussian Blur (SVGB) [4], a linear space-variant Gaussian blur that linearly reduces the kernel size while filtering a face region.

Classifier and attacks. We use the OpenFace [16] face recognizer to evaluate the privacy protection performance of a probe face video. OpenFace extracts a 128-dimensional feature vector for each frame using a deep Convolutional Neural Network (CNN) and then uses a Support Vector Machine (SVM) classifier [17].

We evaluate all filters under a naïve, parrot, naïve-SR and parrot-SR attacks. We use the SRCNN [7] super-resolution algorithm for naïve-SR and parrot-SR attacks.

Performance measures. As privacy measure, we use the cumulative rank-n identification accuracy, η, defined as

$$\eta = \sum_{n=1}^{N} \left(\frac{1}{KM} \sum_{k=1}^{K} \sum_{p=1}^{P} x_{kp} \right)_n, \tag{5}$$

where N is the identification rank, K is the number of subjects, P is the number of frames in a video, and

$$x_{kp} = \begin{cases} 1 & \text{if } l = \hat{l} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where l and $\hat{l}$ are the true and predicted labels, respectively.

To measure fidelity, we use the Peak Signal to Noise Ratio (PSNR) that calculates the power ratio of the original frame with respect to the filtered frame in a video.

We measure flicker through a subjective as well as an objective evaluation. The objective evaluation uses the maximum of absolute difference, ψ, of pixel intensities defined as [18]

$$\psi = \sum_{h=1}^{H} \sum_{v=1}^{V} \varphi(h, v), \tag{7}$$

where H and V are the horizontal and vertical dimensions of R_m, respectively; h and v indicate the pixel position; and

$$\varphi(h,v) = max\Big(0, |\bar{R}_m(h,v) - \bar{R}_{m-1}(h,v)| - |R_m(h,v) - R_{m-1}(h,v)|\Big), \tag{8}$$

where $R_m(h,v)$ ($\bar{R}_m(h,v)$) and $R_{m-1}(h,v)$ ($\bar{R}_{m-1}(h,v)$) are the unprotected (filtered) pixel intensity values from the current and previous frame, respectively.

Datasets. We captured an Ultra-HD video probe data set with a GoPro5 camera mounted with a custom lens (25 mm) using the set up shown in Fig. 4. For training, we captured an HD video gallery indoor data set with the built-in camera of a Lenovo K5 smart phone ensuring pitch angle variation of $10° - 90°$ degrees. There were 11 subjects in both datasets with only frontal faces. We extracted 7944 and 399 key-frames from the probe and gallery video data sets, respectively, using the algorithm in [19], followed by manual post-processing to remove frames affected by motion blur. We pre-processed all key-frames by equalizing illumination, smoothing noise with a bilateral filter, aligning by an affine transformation using eye centres and finally applying elliptical masking to remove non-facial parts.

[2]AGB is regarded as a flicker-free filter.

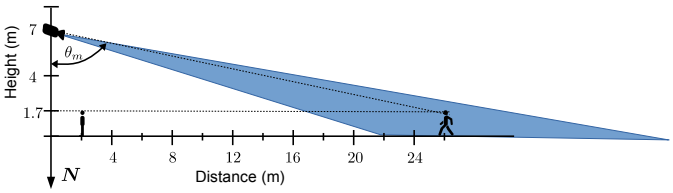
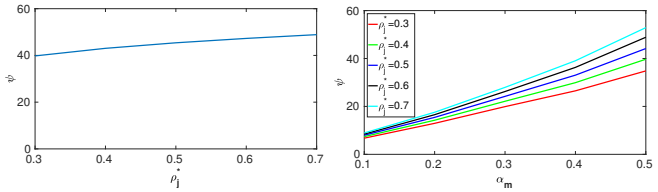


Fig. 4: Setup for the collection of a probe video dataset. The subject moves from a distance of 26 m to 2 m towards the camera, which is positioned at a height of 7 m or 4 m. The variation of the pitch angle, θ_m, is about $20.6° - 78.5°$ from the Nadir direction N.



(a) (b)

Fig. 5: Flicker ψ of (a) AGB [6] at different thresholds ρ_j^* and of (b) SHTMA at different values of the smoothing factor α_m and ρ_j^*. ψ of AGB increases with ρ_j^*. In contrast, ψ of SHTMA depends on both α_m and ρ_j^*, and it negligibly increases with ρ_j^*, especially for low value of α_m (however, it significantly increases for high values of α_m).

Impact of the parameters on flicker. The value of α_m depends on the threshold ρ_j^* and the face movement. We evaluated the effect of α_m on the resulting flicker with different values of ρ_j^* on the detected faces in a video (Fig. 5). The flicker of SHTMA depends on α_m and on ρ_j^*, with a higher variation at larger values of α_m. To achieve a flicker equal to or less than AGB, α_m needs to be selected adaptively depending on ρ_j^*, e.g. at $\rho_j^* = 0.6$ px/cm, an $\alpha_m \in [0, 0.5]$ is selected depending on the face motion.

Attacks on privacy: analysis. Fig. 6 shows the results with the unprotected probe faces for the baseline and for a naïve, parrot, naïve-SR and parrot-SR attacks. Under the naïve attack, SHTMA and AHGMM maintain the highest privacy level (i.e. η comparable to a random classifier) even with $\rho_j^* = 0.6$ px/cm, where AGB and SVGB result in an η larger than that of a random classifier. SHTMA achieves almost the same robustness as AHGMM against a parrot, naïve-SR and parrot-SR attack and it is unaffected by temporal smoothing. In contrast, faces filtered with AGB and SVGB have lower privacy protection (i.e. larger η).

Fidelity analysis. Fig. 7 shows the relationship between η of a filter under different attacks and the corresponding fidelity. SHTMA has a slightly higher fidelity than AHGMM with almost similar values of η. This slightly increased

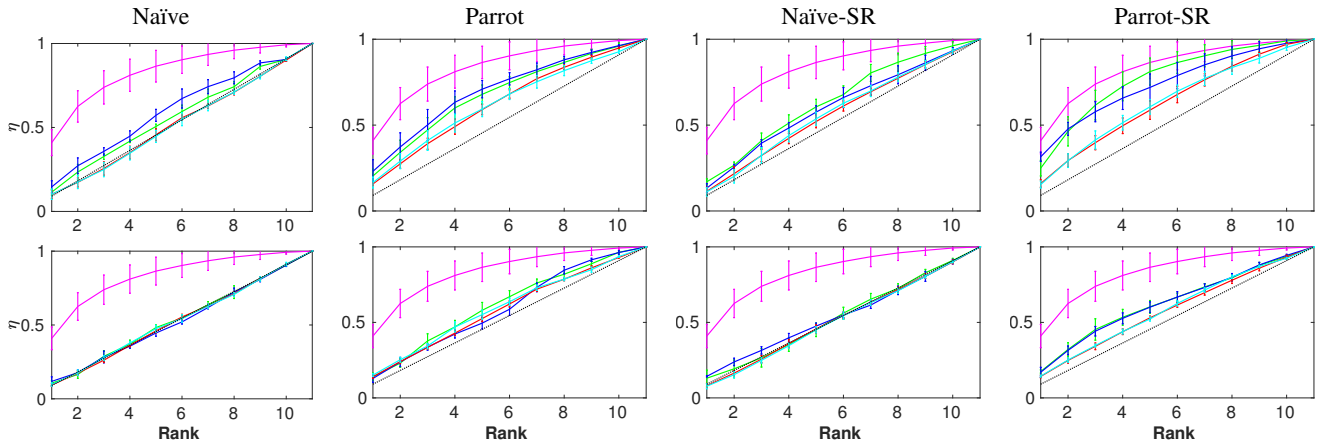


Fig. 6: Rank-n identification accuracy, η, for privacy filters under naïve, parrot, naïve-SR and parrot-SR attacks at threshold $\rho_j^* = 0.6$ px/cm (first row) and $\rho_j^* = 0.4$ px/cm (second row). The filled marker shows the mean and the vertical bar the standard deviation of η for the multi-resolution frames. Legend: Unprotected, SHTMA, AHGMM, AGB [6], SVGB [4]. SHTMA and AHGMM have the highest robustness against attacks (behaviour similar to a random classifier), especially under parrot and parrot-SR attacks.

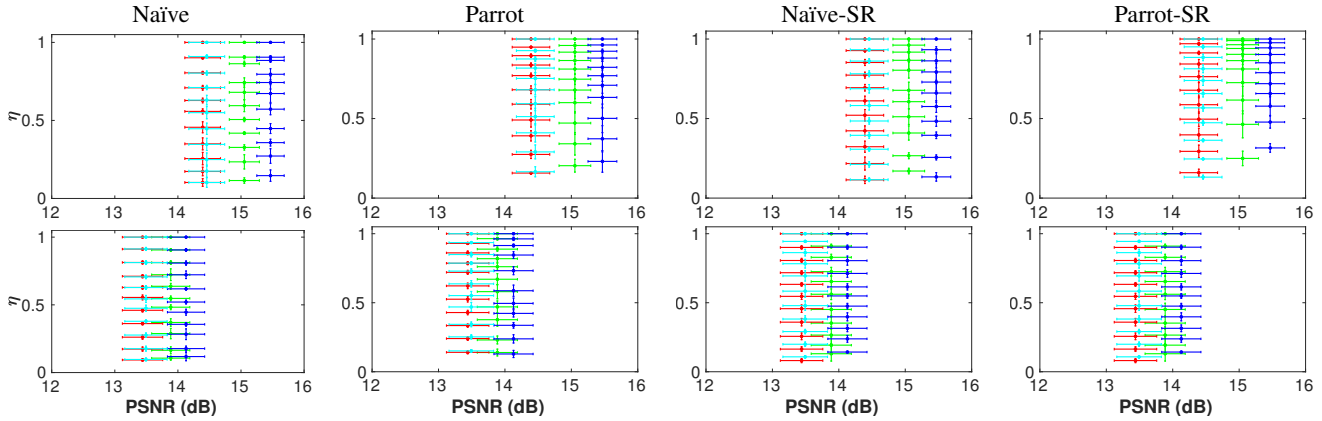


Fig. 7: Relationship between rank-n identification accuracy, η, and fidelity, PSNR, for privacy filters under a naïve, parrot, naïve-SR and parrot-SR attacks at threshold $\rho_j^* = 0.6$ px/cm (first row) and $\rho_j^* = 0.4$ px/cm (second row). The filled marker shows the mean of η and PSNR, the vertical bar indicates the standard deviation of η and the horizontal bar indicates the standard deviation of PSNR for the multi-resolution frames. Legend: SHTMA, AHGMM, AGB [6], SVGB [4]. SHTMA leads to a slightly higher fidelity than AHGMM, due to temporal smoothing. SVGB uses the smallest Gaussian kernels for the outer parts of a face and leads to the highest value, but with lower privacy protection.

fidelity is due to temporal smoothing which also minimises spatial distortion created by the switching kernels. In contrast, SVGB has the highest fidelity at the cost of obtaining the lowest privacy level (larger η), followed by AGB. The main reason for the higher fidelity and lower privacy level of SVGB, compared to AGB, is that the outer parts of a face are processed by smaller Gaussian kernels as SVGB linearly decreases the kernel size. In summary, SHTMA slightly improves fidelity while still robustly protecting the faces against attacks.

Flicker analysis: objective evaluation. Fig. 8 depicts the relationship between η of a filter under different attacks and the corresponding flicker measured using Eq. 7. The flicker generated by SHTMA is significantly lower than that of AHGMM, with almost the same values of η for any threshold ρ_j^*. This lower flicker is the result of averaging the frames with decaying weights: this temporally reduces the effect of the pseudo-random switching of the Gaussian kernels. In contrast, although the flicker of SVGB and AGB is similar to that of SHTMA, SVGB and AGB lead to larger η values (i.e. lower privacy). Comparatively, the flicker of SVGB is slightly greater than that of AGB, due to the linear variation of the Gaussian kernels. SHTMA lowers flicker while being robust against navïe, parrot, navïe-SR and parrot-SR attacks.

Flicker analysis: pair-wise subjective evaluation. We finally evaluate flicker with a set of 20 human observers: 14 males and 6 females, aged between 25 and 35 years old, and without any image or video processing experience. We selected three videos captured with different pitch angles/scales, and filtered them with AGB, SVGB,

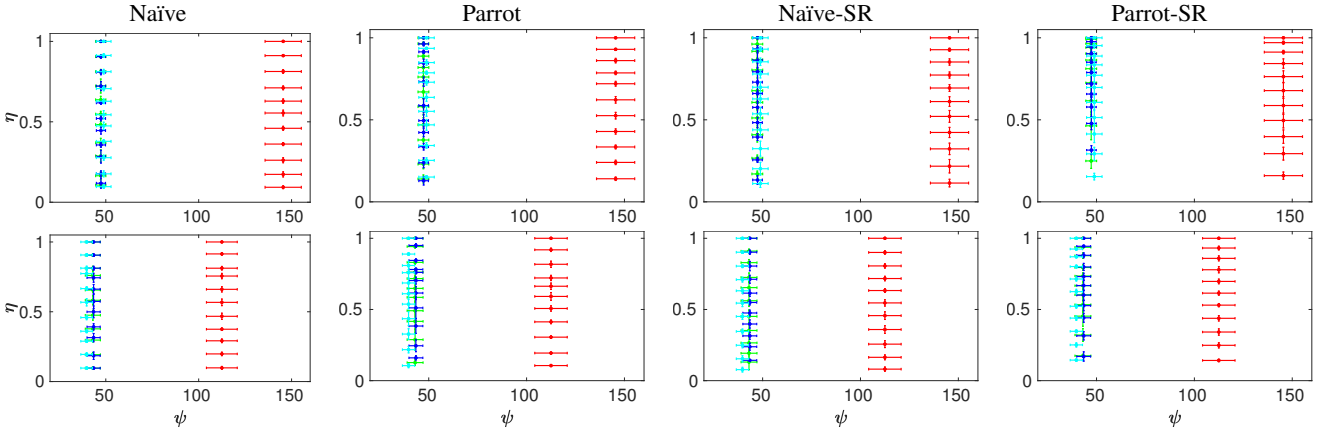


Fig. 8: Relationship between rank-n identification accuracy, η, and flicker, ψ, for privacy filters under naïve, parrot, naïve-SR and parrot-SR attacks at threshold $\rho_j^* = 0.6$ px/cm (first row) and $\rho_j^* = 0.4$ px/cm (second row). The filled marker shows the mean of η and ψ, and the vertical and the horizontal bar indicate the standard deviation of η and ψ, respectively, for the multi-resolution frames. The larger ψ (see Eq. 7), the stronger the flicker. Legend: SHTMA, AHGMM, AGB [6], SVGB [4]. SHTMA has a much lower flicker than AHGMM, and is similar to AGB and SVGB without any decrease in η, thus improving the trade-off between η and ψ.

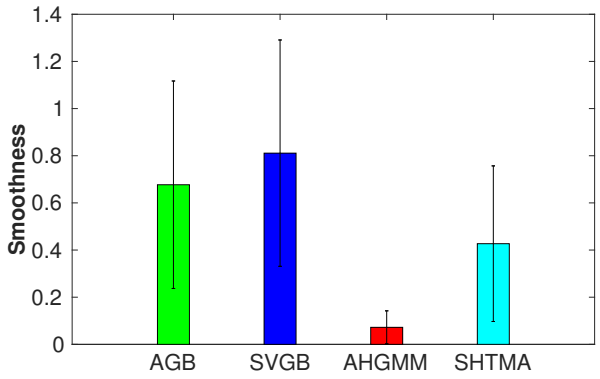


Fig. 9: Subjective evaluation results. The bars indicate the mean (the larger the value, the more frequently videos processed by a method were chosen because of a better smoothness); the vertical lines represent the standard deviation.

Fig. 10: Sample frames from the videos used for the subjective evaluation.

AHGMM, and SHTMA (see Fig. 10 and Fig. 11). We paired the four filtered versions of each video, thus generating six combinations. The observers were asked to select the smoother video for each pair. Fig. 9 shows the results of this subjective evaluation: SHTMA improves smoothness compared to AHGMM, but is less smooth than AGB and SVGB. This may be caused by small jerking in case of significant face movements and by the adaptation of α_m for maintaining the dynamics of the motion of the face.

V. Conclusion

We presented a visual privacy-preserving filter for drone videos that is based on spatio-temporal processing of face regions and that improves the trade-off between privacy, fidelity and temporal smoothness. The proposed filter combines a robust privacy filter based on hopping blur and temporal smoothing, which slightly increases fidelity and significantly decreases the flicker introduced by the hopping blur. Future work includes expanding our analysis to larger datasets and validating the proposed filter on other types of sensitive regions.

References

[1] J. Wickramasuriya, M. Datt, S. Mehrotra, and N. Venkatasubramanian, "Privacy protecting data collection in media spaces," in *Proc. ACM Int. Conf. on Multimedia*, New York, USA, Oct. 2004, pp. 48–55.

[2] E. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying facial images," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, pp. 232–243, Feb. 2005.

[3] F. Dufaux and T. Ebrahimi, "Scrambling for privacy protection in video surveillance systems," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1168–1174, Aug. 2008.

[4] M. Saini, P. K. Atrey, S. Mehrotra, and M. Kankanhalli, "Adaptive transformation for robust privacy protection in video surveillance." *Advances in Multimedia*, vol. 2012, pp. 1–14, Feb. 2012.

[5] P. Korshunov and T. Ebrahimi, "Using face morphing to protect privacy," in *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Kraków, Poland, Aug. 2013, pp. 208–213.

(a) (b) (c) (d) (e)

Fig. 11: Face regions used in the subjective evaluation. (a) Unprotected regions. Regions filtered with (b) AGB [6], (c) SVGB [4], (d) AHGMM and (e) SHTMA (proposed).

[6] O. Sarwar, B. Rinner, and A. Cavallaro, "Design space exploration for adaptive privacy protection in airborne images," in *Proc. IEEE Advanced Video and Signal-based Surveillance*, Colorado Springs, USA, Aug. 2016, pp. 159–165.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[8] P. Korshunov and T. Ebrahimi, "Using warping for privacy protection in video surveillance," in *Proc. Int. Conf. on Digital Signal Processing*, Fira, Santorini, Greece, July 2013, pp. 1–6.

[9] J. X. Yang and H. R. Wu, "A non-linear post filtering method for flicker reduction in H.264/AVC coded video sequences," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Oct. 2008, pp. 181–186.

[10] S. Qiao, Y. Zhang, and H. Wang, "PI-frames for flickering reduction in H.264/AVC video coding," in *Proc. Int. Conf. on Computer Science and Service System*, Aug. 2012, pp. 1551–1554.

[11] A. Jimnez-Moreno, E. Martnez-Enrquez, V. Kumar, and F. D. de Mara, "Standard-compliant low-pass temporal filter to reduce the perceived flicker artifact," *IEEE Trans. on Multimedia*, vol. 16, no. 7, pp. 1863–1873, Nov. 2014.

[12] Z. Wen, J. Li, J. Liu, Y. Zhao, and J. Wen, "Intra frame flicker reduction for parallelized HEVC encoding," in *Proc. Data Compression Conf.*, Snowbird, USA, March 2016, pp. 111–120.

[13] S. B. Yoo, K. Choi, and J. B. Ra, "Blind post-processing for ringing and mosquito artifact reduction in coded videos," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 721–732, May 2014.

[14] D. T. Vo, T. Q. Nguyen, S. Yea, and A. Vetro, "Adaptive fuzzy filtering for artifact reduction in compressed images and videos," *IEEE Trans. on Image Processing*, vol. 18, no. 6, pp. 1166–1178, June 2009.

[15] J. Yang, J. B. Park, and B. Jeon, "Flickering effect reduction for H.264/AVC intra frames," *SPIE 6391, Multimedia Systems and Applications*, vol. IX, pp. 6391 – 9, Oct. 2006.

[16] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., Jan. 2016.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, USA, June 2015, pp. 815–823.

[18] H. Yang, J. M. Boyce, and A. Stein, "Effective flicker removal from periodic intra frames and accurate flicker measurement," in *Proc. IEEE Int. Conf. on Image Processing*, San Diego, USA, Oct. 2008, pp. 2868–2871.

[19] J. L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proc. Int. Conf. on Pattern Recognition.*, vol. 3, Barcelona, Spain, Sep. 2000, pp. 314–317 vol.3.