Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications

Pau Climent-Pérez · Francisco Florez-Revuelta

Received: date / Accepted: date

Abstract Active and assisted living technologies are much needed, but some aspects of them cause user rejection due to concerns on privacy. This is even more concerning to users when visual information is used, processed, and transmitted. To respond to these concerns, and maximise user acceptance, visual privacy protection measures have to be put in place. In the past, human detection and object segmentation in video were constrained by technological limitations, and could only run with specific hardware and sensors. This paper introduces a proposal for an RGB-only based visual privacy preservation filter, which capitalises on 'deep learning'-based segmentation and pose detectors. A background update scheme is presented, which limits leakage of sensitive information when detection fails. Dilation of the mask can further prevent information leakage, but a trade-off is necessary to correctly update background information. This is achieved via a specific study which is also presented. A comparative study is performed to determine the best configuration for privacy preservation. Results show that union of dilated masks from different deep networks achieves the best overall result.

Keywords computer vision · visual privacy · data protection · video surveillance · active and assisted living

1 Introduction

Active and assisted living (AAL) technologies aim at ameliorating the increasing social and economic costs of ageing populations in developed countries. AAL supports independent, healthy living of older and frail people by using information and communication technologies (ICTs) in the home, the workplace, and in public spaces. Sensors embedded in the environment obtain meaningful data about

P. Climent-Pérez · F. Florez Revuelta
Department of Computing Technology,
University of Alicante, P.O. Box 99,
E-03080 Alicante, Spain
Tel.: +34 96 590 34 00 Ext 2515
E-mail: pcliment@dtic.ua.es

its dwellers and their interaction with the environment, which is useful to provide users of these spaces with advanced and personalised healthcare services. The European Union and other governmental bodies have recognised the importance of this field in light of present and near-future social challenges by funding specific calls for research into the development of related technologies, as noted by Calvaresi et al. [6].

Advances both in intelligent systems and computer vision algorithms as well as in used sensors have led to a pervasive use of cameras in AAL and other fields, as they provide richer sensory information, as stated by Nguyen et al. [19]. Additionally, with cameras a single information *stream* can be used for multiple purposes once installed, creating synergies among video-based services. A recent review by Climent-Pérez et al. [10] explores the latest developments in this field, and shows continued interest in this topic to present day.

Different studies have analysed the use of cameras for AAL from different approaches. One way of looking at a division of computer vision methods for AAL is that of Planinc et al. [22], that is, to think of services or applications, and the technologies which can be used to provide them. Some example applications mentioned by Sathyanarayana et al. [25] are fall detection and prevention, human action/activity recognition (HAR), sleep quality assessment (including apnoea), and the monitoring and detection of various physiological metrics, such as vital signs, facial expressions, or epilepsy. Regarding supporting technologies, there are several important landmark components that enable the provision services for AAL based on computer vision, specifically for some of the more complex applications such as human action recognition and human behaviour understanding: namely, machine learning, which encompasses the current trend of pervasive use of neural networks (and 'deep learning'); as well as the development and availability of new camera sensors, modalities (IR, Depth, stereo), and increasing affordability.

However, this pervasive use of cameras does come with public resistance. Increasing concern in society due to aggressive data retrieval, processing and storage by companies and governmental agencies has led to legislators in some areas, like the EU, to put forward regulations to limit these practices and safeguard citizens' rights to privacy. This has led to the creation of the general data protection regulation (GDPR), which in its article 25 includes the concept of *privacy by design and by default*[1]. That is, privacy has to be considered *from the inception* for any new technology or method (algorithm) that is to use data of private nature. It is worth noting that, although cameras for video surveillance in public spaces can be tolerated in exchange of a perceived increase in security, their release in private spaces for the provision of medical assistance is not as welcome, as noted by Arning and Ziefle [2]. Furthermore, Offermann-van Heek and Ziefle [17] show that end users (or patients) are more receptive to the technology than caregivers themselves. These conclusions have fostered the creation of a new study by the same authors [16] regarding user acceptance of camera-based technologies depending on different factors, such as camera placement (location), technologies used, additional privacy-preservation measures taken, as well as additional guarantees necessary for users to trust the technology.

[1] from: `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679` *(Visited: Feb 2020)*

The aim of this paper is to take these considerations into account and to explore privacy preserving measures that would improve user and other stakeholders' acceptance, by minimising users' perceived barriers that current implementations of the technology can pose. Along with [16], this work is part of the ongoing efforts in the EU-funded project on privacy-aware and acceptable lifelogging services for older and frail people (the PAAL project), which:

> "aims to understand users' requirements in terms of perceived benefits and barriers and to increase the awareness of issues associated with lifelogging, establishing guidelines for responsible research in these technologies". — *PAAL project aim and objectives*[2]

The remainder of this paper is organised as follows: next, in Section 2 a review on recent developments in visual privacy preservation techniques is presented; Section 3 introduces a proposal for privacy preservation based only on RGB data without additional inputs, removing previous constraints due to recent advances in neural networks. Finally, Section 4 discusses the methods proposed in the paper and presents some future work.

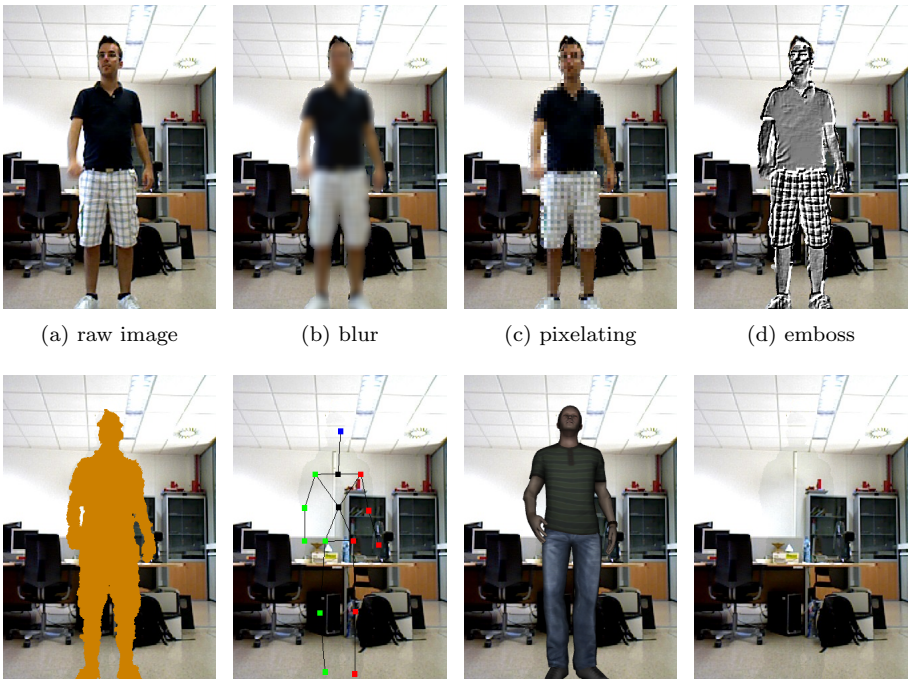
2 Visual privacy preservation

The 'privacy by design' concept mentioned in the GDPR includes, of course, any new technology aimed at an ageing society, and specially affects the development of new technologies based on computer vision, as faces and any other (visually) identifying or biometric cues are specifically regarded by the regulation as highly sensitive data. This has led to an increase in research aimed at visual privacy preservation methods in different contexts: from the social media *advisor* proposed by Orekondy et al. [20], which advises users on the suitability for upload of certain pictures based on their contents (faces, credit cards being shown, passersby, etc); to systems on board unmanned aircraft systems or vehicles (UAS/UAVs), such as the one by Babiceanu et al. [3], to limit the capture of video to a certain area of interest, as well as blurring non-relevant faces, car plates, or other sensitive information that might be recorded accidentally.

Although several methods in the literature focus on faces as the sole cue used to determine identity [3, 11, 26, 27], it is important to note that there are other features of a person that might reveal their identity such as those called soft- and non-biometric markers as stated in the survey by Ribaric et al. [24]: these include voice and gait (soft-biometric), as well as clothing, tattoos, and hairstyle (non-biometric).

There are several ways in which privacy can be preserved. This is thoroughly reviewed by Padilla-López et al. [21], where visual privacy protection methods are introduced, namely: intervention, blind vision, secure processing, redaction methods, and data hiding methods. Each consists on different approaches to the problem: in intervention, cameras and devices are prevented from capturing images; in blind vision and secure processing, algorithms are unaware of the nature of the images being processed; redaction methods destroy or alter the information in sensitive areas of the image; finally, data hiding methods conceal the sensitive

[2] For more information visit: `www.paal-project.eu` *(last access: Feb 2020)*



(a) raw image (b) blur (c) pixelating (d) emboss

(e) silhouette (f) skeleton (g) 3*D* avatar (h) invisibility

Fig. 1: Examples of visualisation models implemented in [21] for a sample frame, ordered from lower to higher level of protection offered. Their implementation relies on RGB-D data. Also, user silhouette and skeleton detection need to be correctly performed (reproduced from [21]).

information in the image, in a reversible way by using pixel scrambling, cryptography, etc.

One of the most common approaches is data redaction, that is, the elimination or non-reversible distortion of sensitive information from the acquired footage. This can be done via different *classical* filters as shown in Figure 1, namely: blurring, pixelation, embossing, a solid colour box, or other *destructive* means. Similarly, Hasan et al. [14] propose *cartooning* as a means for data redaction. Their proposed object detection replaces sensitive information in videos and lifelogging imagery with fitted or scaled cartoons (images of the objects being replaced (i.e. TV and phone screens, PC monitors, credit cards, faces, etc.). The idea is rooted in the need to preserve semantics in the video while preserving privacy, which is tested by means of a user study.

As mentioned, non-destructive means also exist (i.e. using *data hiding* methods). These methods usually alter the whole image or sensitive parts of it in a way that is fully reversible, but only to authorised viewers. One way this can be done is by using cryptography (e.g. scrambling of image pixels based on a secret key). An example of this means of privacy preservation is shown by Çiftçi et al. [9]. In their work they apply a false colour replacement scheme within the JPEG archi-

tecture itself. A difference image is encrypted and kept along the obscured image for reversal.

In [11, 18] several implementations of privacy enhancing technologies (PETs) are introduced. Different solutions to avoid unintentionally being photographed and recognised are presented (i.e. *intervention* methods). These PET proposals are classified according to their current availability. The first type, 'destructive' PETs is implemented via a pair of glasses that impairs processing/recognition of facial cues when worn by the users, and uses readily available technology. Another body of research [30, 32], specially in the egovision area, has looked into *bystanders* and their lack of relation to the user (and lack of consent). Visual Bubble, by Wang et al. [32] is a system for the protection of bystanders based on distance. Stereo cameras are used to determine distance and only people which are physically close to the user have their identities revealed. As opposed to distance-only, Steil et al. [30] propose a system not only for bystander protection, but also other sensitive situations, based on an end-to-end deep learning model the output of which is used to mechanically close a shutter in front of the camera (for reassurance). Eye tracking is then used to determine when the user has moved their head to point at a less sensitive area.

Another approach to privacy preservation is related with retaining the capability to recognise the activity while the video is degraded or transformed heavily in such way that is either fully unrecognisable by humans, or alternatively, the identities of the people appearing in the video are [8, 34] (i.e. related to the concepts of *secure processing* and *blind vision*). One example of this is shown by Wu et al. [34] where a degradation transform is learnt in a way that maximises action recognition on the degraded video. Similarly, Chen et al. [8] propose a privacy-preserving representation-learning variational generative adversarial network (PPRL-VGAN) to learn an image representation that is independent of user identity, while retaining discriminative power for facial expression recognition.

Privacy preservation is also sometimes seen through the lens of *user consent* and *context*. What this means is that sometimes users do not mind appearing on footage (e.g. being in a group with people they know, or recording an important event in their lives, or as a means to safeguard security), whereas others they might not want to appear on the video. This dynamic nature of context and consent calls for technology to be aware of user preferences at all times, with a default set per user. Several researchers have gone in this direction, such as 'respectful cameras' by Schiff et al. [26], where high-visibility vests and hard hats are used as a means to mark user preference. Similar to this idea, Shu et al. [27] propose *Cardea*, a system where users' default preferences are registered for different contexts, and can be overridden by a preset set of gestures to bypass the default preference in a specific context.

Another approach at integrating *context* in visual privacy preservation is by considering not just the context of the physical environment, but also the *social* context, that is, the relationships between the end user of the technology and the people having access to the user's information. This establishes a circle of trust, with concentric layers, with people closer inside in the circle having access to more information, or to less redacted data, whereas people further away in the social circle of the user might have less access to the *raw* (unprocessed) user data. This is thoroughly discussed by Chaaraoui et al. [7], where a privacy-by-context approach is introduced: first, the elements that constitute the identity of a user are

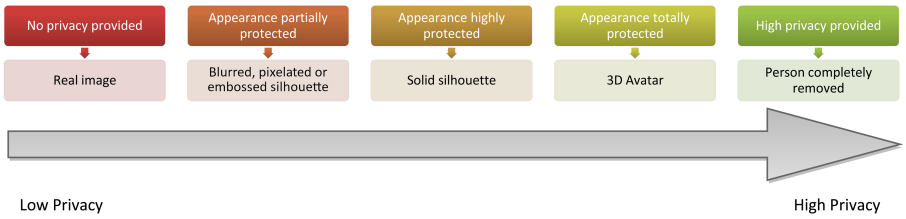


Fig. 2: Different levels of privacy according to the observer and their relationship to the observee. The left-most level offers the view of the full unprocessed image, aimed for the users themselves or very close relatives. As levels lay more to the right, visual privacy is increased, e.g. changing the person for a $3D$ avatar, which still retains semantics of the scene, but better preserves identity.

identified; with these, it is then possible to adapt the privacy level (i.e. using the different visualisations provided, as shown in Figure 2) based on the relationship of the user and the observer, as well as other cues of what variables make up the context, which the authors propose to be:

1. **Identity**: of the user, to retrieve their preferences.
2. **Appearance:** e.g. clothing, partial or full nudity.
3. **Location:** kitchen, bathroom, bedroom, etc.
4. **Ongoing activity:** cooking, watching TV, etc.
5. **Event:** what happened during this (i.e. fall, loss of consciousness, alarm button pressed).
6. **Observer:** to determine whether they have access rights.
7. **Relationship:** i.e. relative, health processional, caregiver, friend, etc.
8. **Response:** by the subject (if requested).

With such a context-aware scheme with different levels of data protection, it is possible to obtain tailored visualisations for different stakeholders. Privacy can be preserved, while maintaining the necessary intelligibility required for each application and observer. Indeed, there must be a trade-off between those two components of a privacy filter (i.e. intelligibility v. usefulness of the data). It is important that the user and other stakeholders be aware of the limitations of certain visualisation schemes for specific application contexts: for instance, a healthcare provider might need to have cues on the visual aspect of the user (for instance, face skin coloration, to check for blocked airways in a choking event) that might be concealed by a stringent privacy filter, according to other variables of the context, or the access rights granted to them.

This paper builds upon the study by Padilla-López et al. [21], but takes into consideration the appearance and current widespread use of *deep-learning* based neural networks which were unavailable in the time of publication of their study. In [21], RGB-D sensors were used to obtain user pose information inferred from skeletal and depth information provided by these sensors. This was a hard constraint, as RGB-only based methods were immature for the application of visual privacy preservation, as their accuracy was not at par with those using depth information. False negatives, that is, missing a detection for several frames, and thus not applying the selected data obfuscation method, would imply revealing the

users' identities and would defeat the purpose of a privacy preservation method. Furthermore, most RGB-D sensors were targeted at the gaming market, meaning the users were expected to be facing the device and therefore detection would fail when the user was facing away, or in poses not expected during a video game (e.g. sitting on a couch, since the user might be merged to the background, lying on the floor or standing against a wall, etc). Furthermore, most implementations of user detection for RGB-D based devices requires a predetermined starting pose (so called, Y-pose, i.e. standing with raised arms and flexed elbows) to start tracking a person's movements. Yet, properly detecting all movements from the start, and any body pose, is decisive in many AAL scenarios. That is, the methods for privacy preservation presented in [21] rely too heavily on the depth information, and would not work when using RGB information only.

3 Privacy preservation in RGB videos

Similarly to Padilla-López et al. [21], we propose a series of different filters or masks of sensitive information on the image: transparency (invisibility), replacement by an avatar, pixelation, blurring and embossing; but this time with RGB information only. Contrary to [21], in this case, no additional depth information is used. Loosening the constraint of having to use RGB-D devices has several advantages. For one, any RGB camera can be used, which are ubiquitous; furthermore, person detectors embedded in RGB-D devices were mostly trained for video gaming purposes, required an initialisation pose, and did not detect many *natural* human poses, as said. However, removing this constraint has only been made possible recently, thanks to the advancement of 'neural network'-based human mask and pose detectors, such as mask region convolutional neural networks (Mask R-CNN, [15]), or more recently *DensePose* [13].

Figure 3 shows the overall view of the visual privacy filters implemented in this paper. First, each RGB image is fed to DensePose and Mask R-CNN networks. The masks generated from these networks are then merged (via union) to create the user mask. Background (BG) modelling is then applied. Dilation is applied to the mask during this process, and then its negative is used to determine the area which will be updated during background update. This process achieves invisibility of the user in the video. The reason behind applying the invisibility filter first is to avoid revealing any identifying piece of clothing or extremities not detected as foreground (mask errors, i.e. false negatives), which would defeat the purpose of the system. Blurring, embossing and pixelation use the original mask to redact the sensitive area of the image, and lay it on top of the invisibility mask of the previous step. Similarly, the avatar is laid on top of the background (invisibility) using pose information from DensePose, as well as texture information from a bitmap (as explained later, in Fig. 7).

All these steps are explained in detail in the following subsections. We start by reasoning why the union of masks is used (Sec. 3.1), rather than a single segmentation network. Next, the invisibility filter via background modelling is explained (Sec. 3.2). Finally, Sec. 3.3 explains how the other privacy filters work: embossing, pixelation, and blurring; as well as the avatar superimposition.

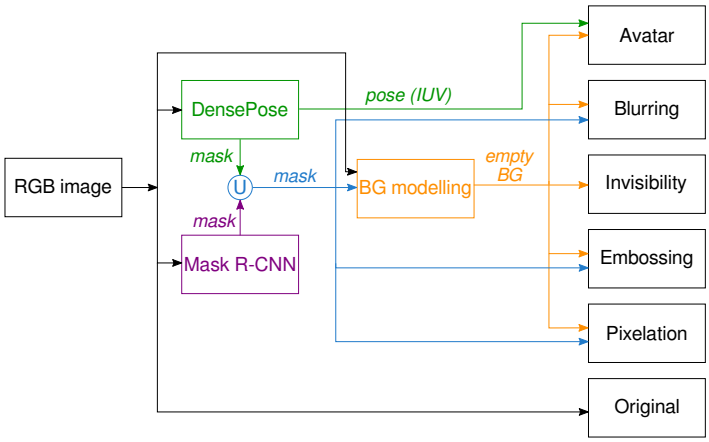


Fig. 3: Overview of all privacy filters implemented. DensePose and Mask R-CNN masks are merged (via union) to create the user mask, background (BG) modelling uses this information to keep an empty background (user invisibility). Blurring, embossing and pixelation modify the pixels in the original mask to show them on the background. Similarly, the avatar is laid on top of the background using pose information from DensePose.

3.1 Person detection mask comparison

As an initial step, it would be important to determine which of several methods provide the most adequate mask for segmentation of people present in the scene for the task of visual privacy preservation. In this task, it is important to have a minimal amount of false negatives, i.e. recall must be high, ideally as close to 1.0 as possible. This is so, because undetected areas of the person might reveal their identity and would not be protected by the redaction applied in later stages.

DensePose and Mask R-CNN differ fundamentally in their understanding of what constitutes a person. Mask R-CNN by He et al. [15] is a general framework for object instance segmentation. That is, it aims at segmenting different parts of an image as either background, or other objects, including people. For this, it extends on Faster R-CNN [23] by adding a branch for predicting an object mask in parallel with the branch in the original architecture used for bounding box recognition. As opposed to that, DensePose, by Güler et al. [13], is a network aimed at finding a surface-based representation of of all people present in a scene (taking into account clothing, and inferring the surface beneath). It is worth noting here, however, that these are not $3D$ coordinates (i.e. real world coordinates). DensePose is a fully convolutional network (FCN), which performs both classification of pixel-level segmentation into either background or the different body parts, and regression to find the $2D$ coordinates (UV space) within the part. The final result is a IUV map: that is a map where each pixel is assigned an identity (body part it belongs to); a U-axis coordinate and a V-axis coordinate within the body part. As a result of how Mask R-CNN is trained, it takes clothing as part of the detection, thus, the masks provided tend to comprise slightly larger areas than DensePose. This

(a) (b)

Fig. 4: Example comparing masks generated by Mask R-CNN and DensePose. (a) is the original frame; (b) shows resulting masks, Mask R-CNN is depicted in green, and DensePose in red. Yellow is the intersection of both. It can be noted that DensePose is more anatomically correct, whereas Mask R-CNN leaves out some details of the feet and hands. Yet, the green mask (Mask R-CNN) covers loosely fitting clothes better (best seen in colour).

effect can be seen in Figure 4, which shows a comparison of the generated masks for both networks.

To demonstrate the current possibilities for visual privacy preservation using person detectors directly on RGB data (instead of RGB-D), available implementations[3] of DensePose and Mask R-CNN were used. The DensePose network implementation works on top of Detectron 2, by Wu et al. [33]. For Mask R-CNN, the implementation of Abdulla [1] was used.

As an initial study, we compare the masks provided by these two state-of-the-art neural networks since their approach and understanding of what constitutes a person differs, as explained. Table 1 shows the average values for the precision and recall (with standard deviations) for the manually annotated subset[4] (MAS) of the MuHAVi dataset [28]. This labelled subset consists of 7,940 manually annotated ground truth silhouette images, coming from 136 video sequences of 14 different actions (including: walk, run, collapse, guard, kick, punch, stand-up, turn around, and variations) performed by two actors (Actors 1 and 4 of the original full dataset) and from two cameras (Cameras 3 and 4 from the original), with up to 4 samples or repetitions of each performance. The F_1 score is also provided for easy interpretation of the best overall set of masks generated. As can be seen, the highest F_1 score is obtained when using the union of the masks generated by both networks. The most likely explanation is that Mask R-CNN seemingly oversimplifies the shape of the mask, causing convex shapes (head, feet) to be smoothed out, whereas DensePose finds a better fit in these areas (closer to the

[3] Both are available online on GitHub. The reader is referred to the references section for details.

[4] Available under request from: `http://velastin.dynu.com/MuHAVi-MAS/` *(last access: Feb 2020)*

Table 1: Average precision, recall and F_1 score for each generated set of masks on the MuHAVi-MAS dataset. Best values are shown in **bold**; second best are underlined.

Mask used	Precision	Recall	F_1 score
DensePose (D)	$\mathbf{0.86} \pm 0.24$	0.67 ± 0.23	0.75 ± 0.23
Mask R-CNN (M)	0.74 ± 0.25	$\underline{0.84} \pm 0.27$	$\underline{0.79} \pm 0.25$
Union ($D \cup M$)	$\underline{0.77} \pm 0.19$	$\mathbf{0.87} \pm 0.23$	$\mathbf{0.81} \pm 0.20$

body shape), however, DensePose leaves out most of the pixels showing just loose clothing, which Mask R-CNN does a better job at detecting (given their different design goals). It is worth mentioning that the union of both masks gets also the best recall value (i.e. detected pixels over expected pixels or ground truth). Recall is a good measure of *safety* when it comes to *not revealing* identifying information, and therefore it is a main goal of visual privacy approaches to have as higher a value as possible.

For this reason, we have studied the effect that enlarging the obtained mask produces on the recall value. This enlargement can be done via dilation, and while it reduces precision past a certain point, this has not a contrary effect in the overall result as it is only applied in the background update scheme. Figure 5 shows the precision and recall scores according to the radius (in pixels, px) of the structuring element (circle) used for dilation (tries from zero, no dilation, to 24 pixels). It can be observed that with a value of around 15 px, the recall increase is not substantial with larger radii, and is close to 1. Please note that for the generation of Figure 5, images in the MAS subset for which there was no detection from Mask R-CNN and DensePose, were removed, in order to get the recall score as close to one as possible. This allows to determine the right value for the dilation without having to take into account the real-world upper limit of the detection methods used. See Figure 6a-6c for a visual example of this dilation process.

3.2 Background modelling: invisibility filter

With the given mask and dilation radius values, background modelling can be then applied. To properly model the background a dynamic update mechanism is necessary. Please note that in background subtraction, as in any field related to computer vision, there has been a vast amount of recent research. The reader is referred to the reviews by Bouwmans et al. [4, 5], which cover, respectively, *classical* (engineered) as well as neural network based solutions. A good resource is also the *BGS library* (with BGS standing for background subtraction) by Sobral [29]. However, having already a foreground mask given by a segmentation algorithm, background subtraction is unnecessary, and therefore, only modelling (an update mechanism) is needed. This eases the burden of frame subtraction, thresholding, pixel-level models, etc. making it possible to use simple means to keep an updated background model. A temporal mean using a frame queue is sufficient. This is similar to some of the basic models listed in [29] (i.e. temporal mean or median).

In the proposed solution, background modelling (background update) is implemented by keeping a queue F of recent frames $f_0, f_t, f_{2t} ... f_{Nt}$. There are two

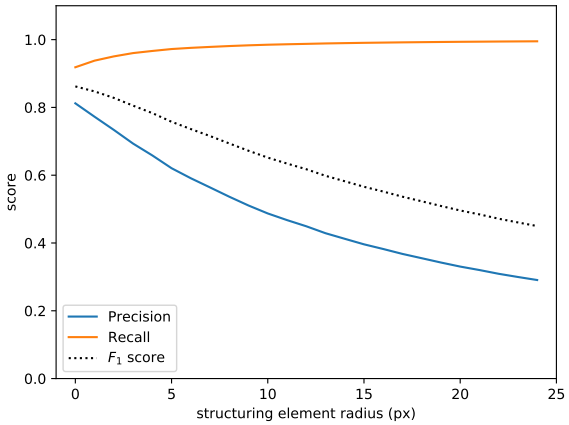
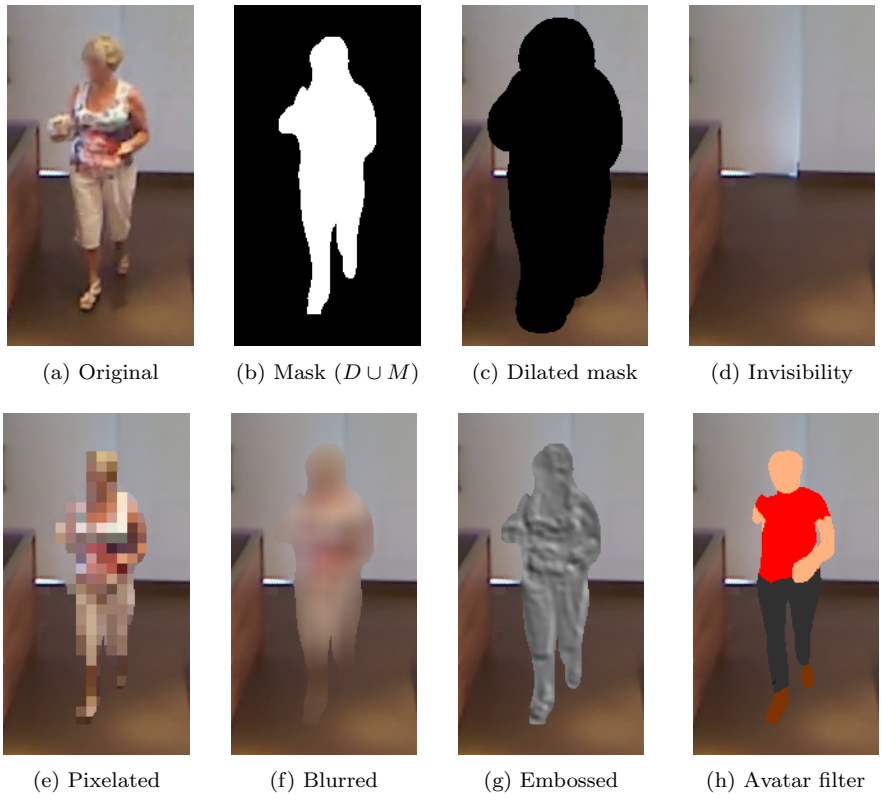


Fig. 5: Results for the dilation study. Recall achieves a near 1.0 value at around a radius size of the circle-shaped structuring element of 15 pixels (px).

parameters to this algorithm: N which is the number of frames kept, and t which is the update rate (in frames). To initialise the algorithm, the first frame is copied over N times in the queue. In subsequent updates (every t frames), the oldest frame is dropped, and the most recent frame is incorporated to the queue. The background model is then updated, as the average image of the frames in the queue. This mechanism by itself prevents the background model from incorporating pixels corresponding to foreground objects if they remain in movement. To further protect the model from non-moving foreground objects (e.g. a sitting person), the background model is only updated with the information outside the detection mask. With this scheme an invisibility visual privacy filter can be achieved. False negatives in the person detector, however, cause the model to incorporate pixels from the person's appearance. Nonetheless, this is a rare case (e.g. only 0.05 % of frames in MuHAVi-MAS generated no mask at all), and the background quickly recovers, as usually only one frame in the whole queue is affected, and the average value is closer to the rest, therefore mitigating the leakage of sensitive information on the person's identity.

3.3 Other visual privacy filters

As Figure 1 shows, each filter offers a different level of protection. For instance, faces can still be barely recognised in embossing, whereas blurring and pixelation offer more protection, depending on the radius of the applied distortion. Replacement of the person by an avatar conceals more visual cues, while the invisibility filter protects completely the appearance of the user. Padilla-López et al. [21] presents an empirical study to determine users' acceptance and perceived level of protection for different filters. The reader is referred to their work for further details.



(a) Original (b) Mask ($D \cup M$) (c) Dilated mask (d) Invisibility

(e) Pixelated (f) Blurred (g) Embossed (h) Avatar filter

Fig. 6: Example frame from the Toyota Smarthome dataset [12], within the workflow of the proposed method. (a) shows the original frame at time t; (b) shows the union mask obtained for this frame; (c) shows the background image fed to the background updating scheme with a dilated negative mask; (d) shows the resulting background model which is also the invisibility filter; (e) through (h) show different filters applied to the foreground mask (pixelation, blurring, embossing, and replacement with an avatar).

The background model, as explained, serves as one of the safest visual privacy preservation filters. It only allows the observer to get certain contextual cues of the environment, such as presence or absence of objects, lighting conditions, movement of furniture (e.g. chairs), without any representation of the observed subjects. This high level of data redaction is suitable for stakeholders for which the observed user does not have a close relationship. Other filters presented here are for closer members of the user's social circle (e.g. healthcare providers or caregivers, members of the family), since they allow a better understanding of the scene and reveal more of the subject's appearance on the screen.

Figure 6 shows one example for all the filters presented, on a sample video from the Toyota Smarthome dataset[5], by Das et al. [12]. The top row shows the original frame, obtained mask (union of DensePose and Mask R-CNN), the

[5] Available from: `https://project.inria.fr/toyotasmarthome/` *(Last access: Feb 2020)*

image fed to the background model (i.e. ignoring the area around the foreground detection to avoid its incorporation to the model), as well as the invisibility filter (background model) achieved (Fig. 6d). On the bottom row, additional visual privacy preservation filters are shown. From left to right: pixelation, blurring, embossing, and replacement with an avatar, respectively.

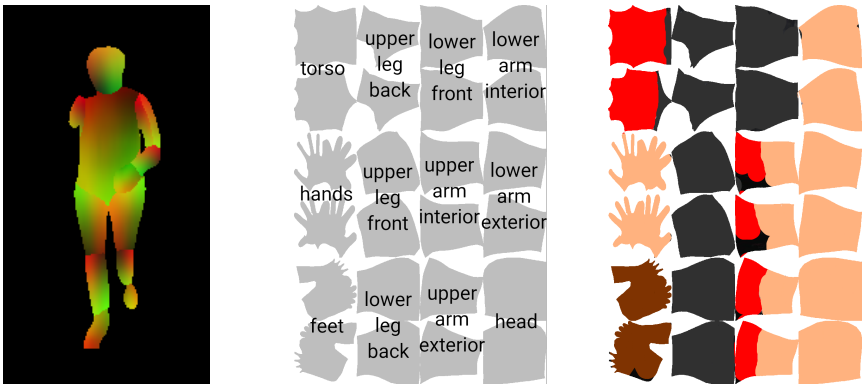
Pixelation, blurring and embossing all work in a similar fashion. First, the invisibility filter is applied, and the foreground mask undergoes a transformation to reduce visual information to an extent in which the person's pose is still recognisable, but their identity is not. Embossing might require some dilation in the foreground mask to include the border of the person. However, precision has to remain almost intact when it comes to dilating the foreground mask to include some of the detection's border, since otherwise the score falls quite rapidly with increasing dilation of the original mask (as seen in Fig. 5). A value of just 1 px around the border is added.

Superimposition of an anonymising avatar is achieved via DensePose. The top layer of this network has two heads: one results in an "individuals" map (indices representing different detections, i.e. people), and another called the IUV map (individual body part indices, and location of these in a UV coordinate space). This IUV map can then be used to apply an avatar texture to replace the original user's appearance and create a completely anonymous appearance of the person. This filter allows an observer to determine the user's body pose, and actions being performed, while keeping an almost complete abstraction on the user's identity. This is useful for stakeholders needing to see the range of movement of the user, but might not necessarily be interested in identifying details. Figure 7 shows the IUV map obtained for the same frame shown in Figure 6. A texture atlas using 24 body surfaces with their UV coordinates is used to map each pixel from the texture onto the IUV map, obtaining a textured avatar as shown in Figure 6h.

4 Conclusion

In this paper, it has been shown how to use off-the-shelf deep neural network implementations for the purpose of visual privacy preservation. This is in contrast to approaches prior to 'deep learning' based ones, which had to rely on RGB-D data to facilitate person detection and segmentation. A comparison study of two existing networks is performed. Results show the masks provided are complementary to one another, thus union of the masks appears to be beneficial for the task at hand. Furthermore, it has been shown that moderate dilation around the person detection is useful to further preserve the privacy of imperfectly detected limbs or identifying pieces of clothing. Finding a trade-off between too much or too little dilation helps prevent two problems: keeping a background where objects moved by the users in the scene are not shown in their new locations, which is contextual information very useful to observers (e.g. caregivers, relatives, etc.), when too much dilation is applied; and accidentally revealing too much of the observee's identities.

From a user perspective, replacement by an avatar might be seen as a safe way to produce semantically understandable video outputs, which preserve the privacy of the observed user. This type of replacement de-identifies or conceals gender, ethnicity, clothing or nudity, etc. while retaining body pose and actions being performed. We are starting further studies in user acceptance in collaboration with



(a) (b) (c)

Fig. 7: The IUV map inferred by DensePose (a) codes body part indices, and UV coordinates within each body part (shown as red–green colour variations when rendered as *visible* in the RGB colourspace). The texture atlas (b) is used to map the UV coordinates of each body part to XY coordinates of the atlas plane. This mapping technique follows [31]. Finally, the texture used in our study (c) shows the colours chosen for each body part, to form the final result of Fig. 6h.

social scientists and psychologists to analyse whether these and other filters could help users accept these otherwise much needed technologies in the near future.

As part of future work, we intend to use omnidirectional cameras mounted on the ceiling to detect and recognise human actions. However, as shown in Figure 8, the networks utilised in this paper fail in this type of images. Most training datasets where humans are present show humans from side views, and mostly in upright positions. To solve this issue, we plan to use side view cameras and existing detectors to train a human pose detector which works directly on the top view, taking advantage of camera topology and calibration to automatically collect ground truth from the side views which would then be used for training on the top view. With this type of solution, it would be possible to monitor larger areas from a single, affordable ceiling-mounted camera. It is also important to gather a rather large dataset for omnidirectional cameras, as it has been observed there is a lack of publicly available footage of this type.

Acknowledgements This work is part of the PAAL – "Privacy-Aware and Acceptable Lifelogging services for older and frail people" project: The support of the Joint Programme Initiative "More Years, Better Lives" (JPI MYBL, award number: PAAL_JTC2017) and the Spanish Agencia Estatal de Investigación (grant no: PCIN-2017-114) is gratefully acknowledged.

Conflict of interest

The authors declare that they have no conflict of interest. Additionally, the funding bodies had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

(a) Original image

(b) Textured avatar output (DensePose)

(c) Segmentation provided by Mask R-CNN

Fig. 8: Example frame of an omnidirectional camera with failure cases. Original frame is shown in (a); malformed DensePose result is overlaid in (b); human detection missing some limbs and the head, from Mask R-CNN is shown in (c). Please note frames with some output from the networks were selected to compose this figure, and a majority did not provide one (false negatives).

References

1. Abdulla, W.: Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN` (2017)
2. Arning, K., Ziefle, M.: "get that camera out of my house!" conjoint measurement of preferences for video-based healthcare monitoring systems in private and public places. In: Geissbühler, A., Demongeot, J., Mokhtari, M., Abdulrazak, B., Aloulou, H. (eds.) Inclusive Smart Cities and e-Health, pp. 152–164. Springer International Publishing, Cham (2015)
3. Babiceanu, R.F., Bojda, P., Seker, R., Alghumgham, M.A.: An onboard uas visual privacy guard system. In: 2015 Integrated Communication, Navigation and Surveillance Conference (ICNS), pp. J1–1–J1–8 (2015). DOI 10.1109/ICNSURV.2015.7121232
4. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: An overview. Computer Science Review **11-12**, 31 – 66 (2014). DOI https://doi.org/10.1016/j.cosrev.2014.04.001
5. Bouwmans, T., Javed, S., Sultana, M., Jung, S.K.: Deep neural network concepts for background subtraction:a systematic review and comparative evalu-

ation. Neural Networks **117**, 8 – 66 (2019). DOI https://doi.org/10.1016/j.neunet.2019.04.024
6. Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A.F., Sturm, A.: Exploring the ambient assisted living domain: a systematic review. Journal of Ambient Intelligence and Humanized Computing **8**(2), 239–257 (2017)
7. Chaaraoui, A.A., Padilla-López, J.R., Ferrández-Pastor, F.J., Nieto-Hidalgo, M., Flórez-Revuelta, F.: A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. Sensors **14**(5), 8895–8925 (2014). DOI 10.3390/s140508895
8. Chen, J., Konrad, J., Ishwar, P.: Vgan-based image representation learning for privacy-preserving facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1570–1579 (2018)
9. Çiftçi, S., Akyüz, A.O., Ebrahimi, T.: A reliable and reversible image privacy protection based on false colors. IEEE transactions on Multimedia **20**(1), 68–81 (2017)
10. Climent-Pérez, P., Spinsante, S., Mihailidis, A., Florez-Revuelta, F.: A review on video-based active and assisted living technologies for automated lifelogging. Expert Systems with Applications **139**, 112,847 (2020). DOI https://doi.org/10.1016/j.eswa.2019.112847
11. Dabrowski, A., Krombholz, K., Weippl, E.R., Echizen, I.: Smart privacy visor: Bridging the privacy gap. In: Abramowicz, W. (ed.) Business Information Systems Workshops, pp. 235–247. Springer International Publishing, Cham (2015)
12. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7297–7306 (2018)
14. Hasan, R., Shaffer, P., Crandall, D., Apu Kapadia, E.T., et al.: Cartooning for enhanced privacy in lifelogging and streaming videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 29–38 (2017)
15. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
16. Offermann-van Heek, J., Arning, K., Ziefle, M.: All eyes on you! impact of location, camera type, and privacy-security-trade-off on the acceptance of surveillance technologies. In: Donnellan, B., Klein, C., Helfert, M., Gusikhin, O., Pascoal, A. (eds.) Smart Cities, Green Technologies, and Intelligent Transport Systems, pp. 131–149. Springer International Publishing, Cham (2019)
17. Offermann-van Heek, J., Ziefle, M.: They don't care about us! care personnel's perspectives on ambient assisted living technology usage: Scenario-based survey study. JMIR Rehabil Assist Technol **5**(2), e10,424 (2018). DOI 10.2196/10424
18. Krombholz, K., Dabrowski, A., Smith, M., Weippl, E.: Exploring design directions for wearable privacy. In: USEC'17 (2017)
19. Nguyen, T.H.C., Nebel, J.C., Florez-Revuelta, F.: Recognition of activities of daily living with egocentric vision: A review. Sensors **16**(1) (2016). DOI

10.3390/s16010072
20. Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3686–3695 (2017)
21. Padilla-López, J., Chaaraoui, A., Gu, F., Flórez-Revuelta, F.: Visual privacy by context: proposal and evaluation of a level-based visualisation scheme. Sensors **15**(6), 12,959–12,982 (2015)
22. Planinc, R., Chaaraoui, A.A., Kampel, M., Florez-Revuelta, F.: Computer vision for active and assisted living. In: Active and Assisted Living: Technologies and Applications, Healthcare Technologies, pp. 57–79. Institution of Engineering and Technology (2016). DOI {10.1049/PBHE006Ech4}
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
24. Ribaric, S., Ariyaeeinia, A., Pavesic, N.: De-identification for privacy protection in multimedia content: A survey. Signal Processing: Image Communication **47**, 131–151 (2016)
25. Sathyanarayana, S., Satzoda, R.K., Sathyanarayana, S., Thambipillai, S.: Vision-based patient monitoring: a comprehensive review of algorithms and technologies. Journal of Ambient Intelligence and Humanized Computing **9**(2), 225–251 (2018). DOI 10.1007/s12652-015-0328-1
26. Schiff, J., Meingast, M., Mulligan, D.K., Sastry, S., Goldberg, K.: Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In: Senior, A. (ed.) Protecting Privacy in Video Surveillance, pp. 65–89. Springer London, London (2009). DOI 10.1007/978-1-84882-301-3_5
27. Shu, J., Zheng, R., Hui, P.: Cardea: Context-aware visual privacy protection for photo taking and sharing. In: Proceedings of the 9th ACM Multimedia Systems Conference, MMSys '18, p. 304–315. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3204949.3204973
28. Singh, S., Velastin, S.A., Ragheb, H.: Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, pp. 48–55. IEEE (2010)
29. Sobral, A.: BGSLibrary: An opencv c++ background subtraction library. In: IX Workshop de Visão Computacional (WVC'2013). Rio de Janeiro, Brazil (2013). URL `https://github.com/andrewssobral/bgslibrary`
30. Steil, J., Koelle, M., Heuten, W., Boll, S., Bulling, A.: Privaceye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA '19. Association for Computing Machinery, New York, NY, USA (2019). DOI 10.1145/3314111.3319913
31. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
32. Wang, S., Cheung, S.C.S., Sajid, H.: Visual bubble: Protecting privacy in wearable cameras. IEEE Consumer Electronics Magazine **7**(1), 95–105 (2017)
33. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)

34. Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 606–624 (2018)