Person De-identification in Videos

Prachi Agrawal and P. J. Narayanan

Center for Visual Information Technology,
IIIT - Hyderabad, India
prachi@research.iiit.ac.in,pjn@iiit.ac.in

Abstract. Advances in cameras and web technology have made it easy to capture and share large amounts of video data over to a large number of people through services like Google Street View, EveryScape, etc. A large number of cameras oversee public and semi-public spaces today. These raise concerns on the unintentional and unwarranted invasion of the privacy of individuals caught in the videos. To address these concerns, automated methods to *de-identify* individuals in these videos are necessary. De-identification does not aim at destroying all information involving the individuals. Its goals are to obscure the identity of the actor without obscuring the action. This paper outlines the scenarios in which de-identification is required and the issues brought out by those. We also present a preliminary approach to de-identify individuals from videos. A bounding box around each individual present in a video is tracked through the video. An outline of the individuals is approximated by carrying out segmentation on a 3-D Graph of space-time voxels. We explore two de-identification transformations: exponential space-time blur and line integral convolution. We show results on a number of public videos and videos collected in a plausible setting. We also present the preliminary results of a user-study to validate the effectiveness of the de-identification schemes.

1 Introduction

Advances in cameras and web technology have made it easy to capture and share large amounts of video data over the internet. This has raised concerns regarding the privacy of individuals. For example, when photographs of a monument are taken to create a panoramic view of the scene, people present are not aware of it and their consent is not taken before making them public. Technologies like Google Street View, EveryScape, etc., have a high chance of invading into one's private life without meaning to do so. Parents have also expressed concern on the possible compromise of the security of their children. The recent furore over Street View in Japan and the UK underscores the need to address the privacy issue directly. An increasing number of video cameras observe public spaces like airports, train stations, shops, and streets. While there may be a possible security need to see the individuals in them, identifying the action suffices in most cases. The actor need be identified only rarely and only to authorized personnel.

There is, thus, a need to *de-identify* individuals from such videos. De-identification aims to remove all identification information of the person from an image or video, while maintaining as much information on the action and its context. Recognition and de-identification are opposites with the former making use of all possible features to identify an object while the latter trying to obfuscate the features to thwart recognition. De-identification should be resistant to recognition by humans and algorithms. Identifying information captured on video can include face, silhouette, posture, gait, etc. Three types of videos need de-identification to not compromise the privacy of individuals. *Casual videos* that are captured for other purposes and get shared. Examples include images used by projects like Google StreetView, the net-cameras fitted in public spaces that can be viewed over the internet, videos or photos on sharing sites, etc. Individuals appear in these videos purely unintentionally and there is no need to know their identities. All individuals should therefore be de-identified irrevocably and early, perhaps at the camera itself. *Public surveillance videos* come from cameras watching spaces such as airports, streets, stores, etc. There is no intention to capture any specific set of persons, but there is an explicit intention to capture people occupying the space. These videos may be viewed at a monitoring station to look for anomalies but also to judge how users react to situations or products. These may be displayed on public monitors and a recorded version may be accessible to many people. The types of actions performed by individuals in these videos may be important, but not their identities. Hence de-identification is necessary. *Private surveillance videos* come from cameras placed at the entrances of semi-private spaces like offices. Individuals entering them have a purpose and access is often limited to authorized persons only. The videos may be of higher quality and are likely to have a more detailed view of the individuals. De-identification may not be essential, but could be recommended to take care of potential viewing by non-authorized people.

The privacy issues are genuine and will grow with wider adaptation of technology. Automated methods to de-identify individuals without affecting the context of the action in the video are needed to address them. It may be necessary to control the level of de-identification to cater to different situations. Some work directed towards face de-identification has been reported before. In this paper, we discuss the different issues relating to de-identification of individuals in videos. We strive to guard against algorithmic and manual identification using face, silhouette, gait, and other aspects. We also present the design of a de-identification scheme and present results on several standard and relevant videos. We also present preliminary results from a user study conducted to gauge the effectiveness of the strategy.

2 De-identification: General Framework

De-identification involves the detection and a transformation of images of individuals to make them unrecognizable. It is easy to hide the identity of individuals by replacing a conservative area around them by, say, black pixels. However, this hides most information on what sort of human activity is going on in that space, which may be important for various studies. The goal is to protect the privacy of the individuals while providing sufficient feel for the human activities in the space being imaged. There is a natural trade-off between protecting privacy and providing sufficient detail. Privacy protection provided should be immune to recognition using computer vision as well as using human vision.

2.1 Criteria for De-identification

The characteristics or features used to recognize humans in videos is the focus of a de-identification transformation, such as the following.

1. Face plays the dominant role in automatic and manual identification. Thus, the de-identification transformation should pay more attention to detect and obfuscate faces in the video more than other aspects.
2. The body silhouette or the gait are important clues available in videos which need to be obfuscated. Humans exploit them effectively and algorithmic identification using them have been developed with some success [1, 2]. The silhouette can be dilated or expanded to remove its information content. Gait relates to the temporal variation of a person's arms and silhouette. Masking it needs the temporal silhouettes to be changed in a non-predictable way.
3. Other information about individuals may be critical to specific aspects of privacy, such as the race and gender. Both are hard to mask completely. Though race may relate closely to skin colour and can be masked by RGB or hue-space transformations, these destroy the naturalness of the videos in our experience. Gender is more subtle and no clearly defined manifestation has been agreed on, which makes obfuscation of gender hard.

2.2 Subverting De-identification

We now discuss ways by which the de-identification can be subverted or "attacked" to reveal the identity of individuals involved. The de-identification process has to be satisfactorily robust to these methods.

1. Reversing the transformation used for de-identification is the most obvious line of attack. The transformation should, thus, be irreversible. Blurring using convolution is susceptible to reversal by deconvolution. Frames of the de-identified video may be treated as multiple low-resolution observations when a form of blurring is used. Techniques similar to those used in super-resolution may facilitate the reversal of the blurring partially or completely. We use a blurring involving several neighbouring blocks in space and time to prevent reversal.
2. Recognizing persons from face, silhouette, gait, etc., is being pursued actively in Computer Vision. The problem may be set as a series of verification problems, given a list of people. The de-identification transformation has to be robust to the common computer vision algorithms.
3. Manual identification is another way to subvert de-identification, though it is considerably more expensive. It is not clearly known what properties humans use to identify and recognize individuals. However, general blurring and colour manipulation makes recognition highly unlikely even by humans. User study is an effective way to judge the effectiveness of the de-identification approach and to compare between multiple approaches.
4. Brute-force verification is a way to attack a de-identified video. Such attacks are possible if some knowledge of the de-identification algorithm and its parameters

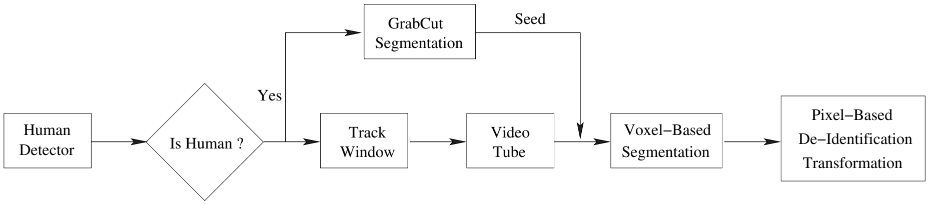


Fig. 1. Overview of the method.

are available. Different combinations of algorithms and their parameters can be applied on target individuals, with comparison performed in the de-identified space. A match in the transformed space can strongly indicate a match in the original space. This way of attack cannot be prevented easily; they can only be made arbitrarily hard by the underlying combinatorics.

It should be noted that only transformations that ignore the input video can theoretically be totally safe. Brute-force attack is possible on others. Such a transformation will replace individuals in the video with a constant (say, black or white) or random colour. We rule out such methods as they destroy all information on the action performed.

3 De-identification: Proposed Approach

An overview of our method is outlined in Figure 1. The system comprises of three modules: Detect and Track, Segmentation, and De-identification.

3.1 Detect and Track

The first step is to detect the presence of a person in the scene. HOG [3] based human detector gives good results with a low miss rate. Other human detectors [4] can also be employed. To track the person in the subsequent frames, a motion compensation based segmentation is useful, which assumes that the foreground objects are small compared to the background. Hence, the dominant motion in each frame is due to the camera. Motion vectors for each pixel can be calculated using optical flow. The dominant motion is the average of motion vectors of all pixels in a frame. The foreground pixels vary significantly from the average motion.

We study the effectiveness of the de-identification process in this paper. We, therefore, concentrate on the de-identification step that follows human detection. Standard databases such as CAVIAR have the necessary ground truth information to study de-identification alone. We created such ground truth on the additional videos we used for experiments.

3.2 Segmentation

The bounding boxes of the human in every frame, provided by the ground truth, are stacked across time to generate a *video tube* of the person. Multiple video tubes are formed if there are multiple people in the video. Segmentation of the person is performed on the video tube as follows. The video space is first divided into fixed voxels of size ($x \times y \times t$) in the spatial (x, y) and temporal (t) domains. Dividing the video space into voxels has two advantages. Firstly, it reduces the computation required in the large video space. Secondly, a block-based segmentation removes fine silhouette information while preserving gross outlines. Fine boundaries of a person reveal a lot about the body shape and gait [1, 2] and can aid recognition.

Segmentation is a labeling where each voxel *v* is assigned a label $\alpha_v \in \{0, 1\}$, where 1 is for foreground and 0 for background. For segmentation, the video tube is divided into blocks of *B* voxel-planes in time. One voxel-plane overlap is used between consecutive blocks to enforce continuity across the blocks. A 3D graph is constructed on these blocks in the voxel space and a mincut is performed on this graph. A rigid but blocky (because of voxelation) outline of the human is extracted by this. The energy term *E* associated with this graph is of the form

$$E(\underline{\alpha}, \underline{\theta}, \underline{v}) = U(\underline{\alpha}, \underline{\theta}, \underline{v}) + \lambda_1 V_1(\underline{v}) + \lambda_2 V_2(\underline{v}), \qquad (1)$$

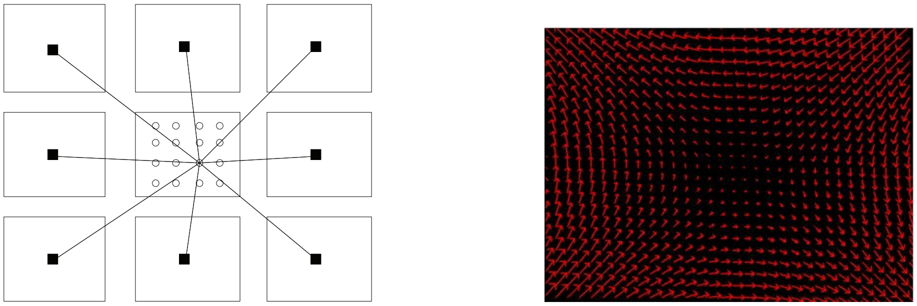
where *U* is the data term and V_1, V_2 are the smoothness terms corresponding to the intra-frame and inter-frame connections between two voxels respectively. $\underline{\theta} = \{\theta^0, \theta^1\}$ are two full-covariance Gaussian colour mixtures, one each for foreground and background, with K clusters each. Hence, $k \in [1, K]$, $\alpha = \{0, 1\}$ and $\theta^\alpha = \{w_k^\alpha, \mu_k^\alpha, \Sigma_k^\alpha\}$. We used $K = 6$ for the results presented here. The Gaussian Mixture Models (GMMs) are used for adequately modeling data points in the colour space.

The energy *E* is defined in such a way that a minimization of this energy provides us with a segmentation that is coherent across time and space. A mincut on the graph minimizes this energy function efficiently [5]. Initialization of foreground and background seeds is done by performing GrabCut [6] on the first frame with the human. The foreground and background GMMs are also initialized in this process. These GMMs later provide seeds to the graph, as well as help in defining the energy terms.

The data term *U* is similar to the one used by GrabCut [6], defined as $U(\underline{\alpha}, \underline{\theta}, \underline{v}) = \sum_n D(\alpha_n, \theta_k, v_n)$ where n is the number of voxels and

$$D(\alpha_n, \theta_k, v_n) = \min_{k=1\cdots K} [- \log w_k^{\alpha_n} + \frac{1}{2} \log \det \Sigma_k^{\alpha_n} + \frac{1}{2} \bar{v}_n^T \Sigma_k^{\alpha_n -1} \bar{v}_n] \qquad (2)$$

where $\bar{v}_n = v_n - \mu_k^{\alpha_n}$. The representative colour v_n for a voxel should be chosen carefully. We first compute the distance D_0 and D_1 to the background and foreground respectively for each pixel in a voxel, using pixel colour instead of v_n in Equation (2). The pixels are sorted on the ratio $\frac{D_0}{D_1}$ in the decreasing order. We choose the colour of m^{th} pixel after sorting as the representative colour v_n. The value of *m* is kept low so that voxels with even a few foreground pixels are biased towards the foreground. This is important for de-identification as the foreground needs to be segmented conservatively. We also identify seed voxels for the graphcut segmentation based on D_0 and D_1. If the distance to foreground, D_1, is very low for the m^{th} pixel, the voxel is a seed foreground. However, if the distance to background, D_0, is very low for the $(N - m)^{th}$ pixel (where N is the number of pixels in the voxel), the voxel is a seed background.



(a)

(b)

Fig. 2. (a) Distances for pixel $(3,3)$ of a voxel from each neighbouring voxel. The distances to the neighbouring voxels in the adjacent voxel plane are calculated in a similar manner. (b) Saddle shaped vector field used for LIC.

The smoothness terms V_1 and V_2 are defined as $V(\underline{v}) = \sum_{v_p,v_q \in \underline{v}} \delta_{pq} \cdot V_{pq}$, where δ_{pq} is 1 when v_p and v_q are neighbours and 0 otherwise, and $V_{pq} = \exp^{-\beta \|v_p - v_q\|^2}$, where v_p is the mean colour of a voxel. β is the expected value calculated as $\beta = (2\mathcal{E}(\|v_p - v_q\|^2))^{-1}$, where $\mathcal{E}$ is the expectation operator [6].

3.3 De-identification

After the segmentation of the person, the next step is to apply the de-identification transformation on the human being present. Two de-identification approaches were explored. One of them is an exponential blur of pixels in a voxel.

In exponential blur, the output colour for each pixel in a foreground voxel is a weighted combination of its neighbouring voxels' average colours. All voxels within distance a participate in the computation of this colour. The weight corresponding to each voxel decreases exponentially with the distance from the voxel center to the pixel. The weights for the $(l,m,n)^{th}$ pixel of voxel v_i can be calculated $\forall v_p \in \mathit{\Gamma}_i$ as:

$$\gamma(l,m,n) = e^{-\frac{d^2_{(l,m,n),v_p}}{8a^2}}, \tag{3}$$

where $\mathit{\Gamma}_i$ is the set of voxels which lie within distance a from v_i, and $d_{(l,m,n),v_p}$ is the distance of the $(l,m,n)^{th}$ pixel from the voxel center v_p. Figure 2(a) shows the distances $d_{(l,m,n)}$ in one voxel plane for the pixel $(3,3)$. This simple blurring function ensures that there is no abrupt change in colour at the voxel boundaries. The temporal blurring of the space-time boundaries aims to remove the gait information of the individual. The amount of de-identification is controlled by varying the parameter a; more the value of a, more is the de-identification.

The second de-identification transformation is based on line integral convolution (LIC). LIC is generally used for imaging vector fields [7] on a texture. We use LIC to distort the boundaries of the person. Different vector fields can be used for achieving different effects. The gradient vector field of the texture image rotated by $90°$ gives a

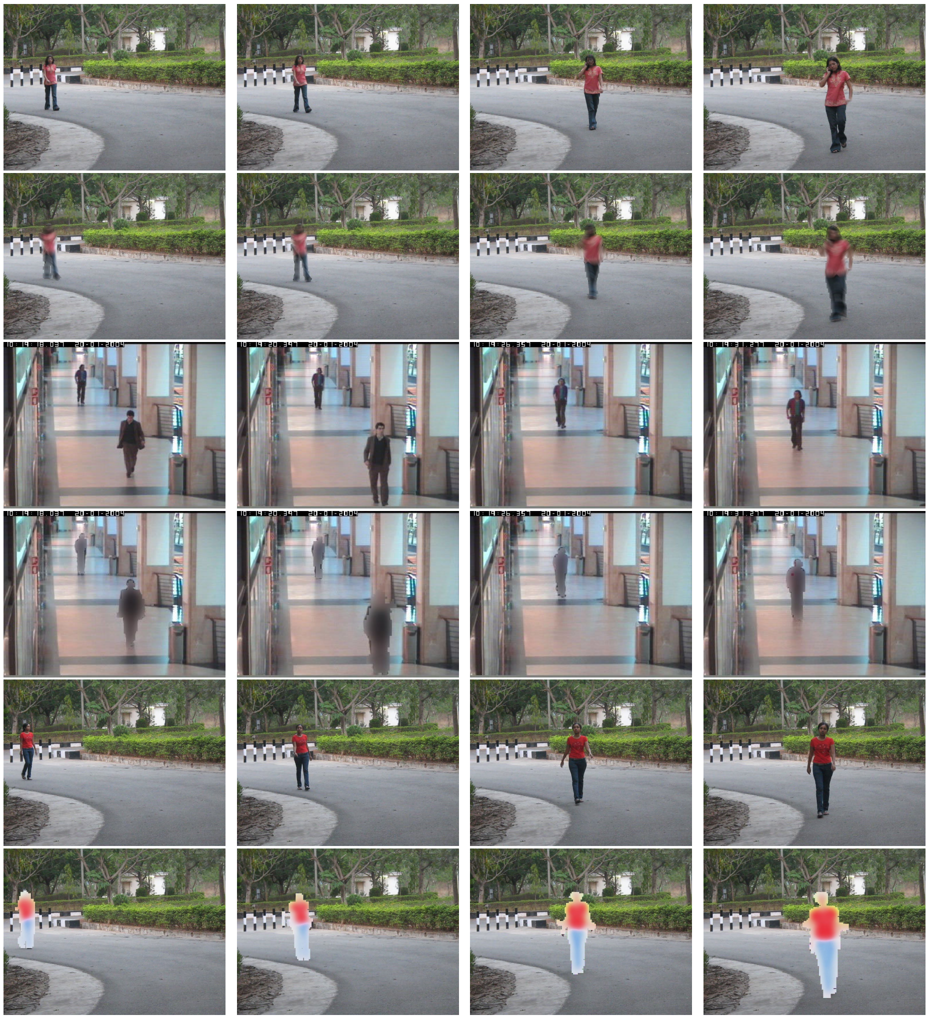


Fig. 3. Results of LIC-10, Blur-4, and Blur-4 followed by an intensity space compression on three different videos, with clear frames in odd rows and de-identified frames in even rows.

painterly effect to the image. We used a saddle shaped vector field (Figure 2(b)) for our experiments. LIC is applied to the foreground pixels obtained after segmentation. The amount of de-identification acquired can be controlled by the line length L, of the convolution filter.

Intensity space compression was additionally tried. The intensity values of the foreground pixels are compressed after an exponential blur or LIC. The result is boosted up by a fixed value after the compression. It provides greater de-identification, but the video loses more information. This simple technique hides the race of a person successfully. The results are presented in Figures 3 and 4.

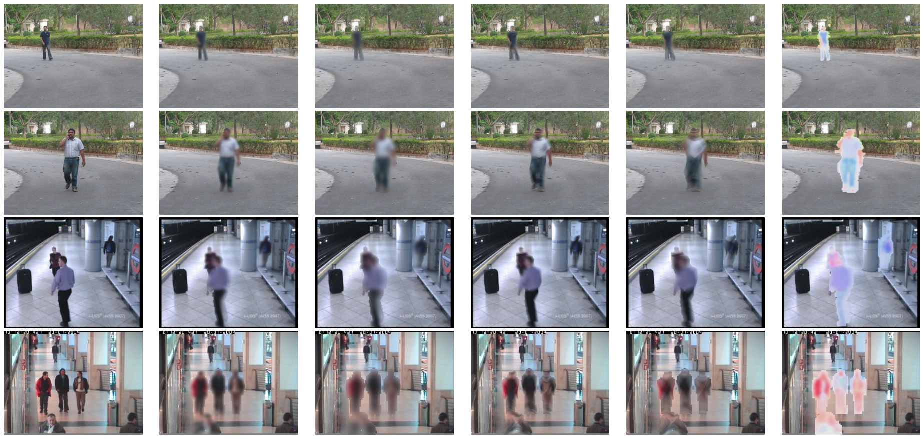


Fig. 4. The first column shows the clear frame. The next five columns show the output of Blur-2, Blur-4, LIC-10, LIC-20, and Blur-2 followed by an intensity space compression, in that order.

4 Experimental Results

We implemented the above system and conducted the experiments on standard datasets like CAVIAR, BEHAVE, etc., and on our own that provide more clearly visible individuals in videos. Some of the results are shown in Figures 3 and 4. Different parameters were tried for each of the de-identification transformations; $a = 2$ and 4 for exponential blur and $L = 10$ and 20 for LIC. We divide the video into $N = 4 \times 4 \times 2$ sized voxels. m was kept as 3 (10% of N) for our experiments. Increasing the voxel size across time domain increases the blockiness across the frames. If a person is moving fast enough, it can introduce jumps in the segmented output around the boundary. More results can be seen in the video submitted as supplemental material.

Figure 4 shows the output of different de-identification transformations on a single frame from different videos. The effect of changing the parameters of the transformations can be seen clearly. Increasing the value of a and L increases the de-identification achieved, but it results in more loss of information in a scene. In general, LIC-10 and Blur-2 are insufficient in masking the identity of people. Blur-4 and LIC-20 perform better. Body structure plays a huge role in identifying people when their faces are obfuscated beyond recognition. LIC distorts the outline of a person based on the vector field used because LIC tries to image the vector field using the person's image as a texture. However the output of LIC-20 sometimes looks unnatural and ghost-like. The intensity space compression, as shown in Figures 3 and 4, can claim to remove the race information. But it preserves the body structure of the person. This happens because the intensity values of the foreground pixels are boosted up and hence rendered visibly different from the background pixels. This trade-off can be avoided by dilating or eroding the foreground mask before applying the intensity space compression.

4.1 User Study

Recognition by humans is one of the ways to subvert de-identification. It is difficult to quantitatively state the effectiveness of the system as it is not known which features humans use to identify and recognize individuals. Hence, a user study was conducted to test the usefulness of the system. We showed 4 different sets of 6 videos each, processed with a different parameter value in each set, to 40 individuals. Half of them were quite familiar with the individuals appearing in the video. Others were only casually familiar. Users were asked to match the individuals appearing in the video against a palette of 30 photographs shown. They were also asked to state the factor that helped them in the recognition. The results are summarized in Table 1. Overall correct recognition was fairly high due to the familiarity of the users with the subjects. The gait or the walking style was also a big give-away for many subjects. The highest recognition was reported for individual 4; about 80% of the users got the correct answer. Almost everyone reported that recognition was possible because of the unique walking style of the person. For individual 2, only about 20% of the answers were correct because this person had no unique body shape or walking style. The correct answers were only from those sets in which low values of parameters for Blur and LIC were used.

	Familiar		*Unfamiliar*	
Algorithm, Parameter	*Correct*	*Incorrect*	*Correct*	*Incorrect*
Blur, $a = 2$	24	6	11	19
Blur, $a = 4$	21	9	10	20
LIC, $L = 10$	24	6	15	15
LIC, $L = 20$	23	7	13	17

Table 1. Statistics: User Study

4.2 Discussion

The preliminary results suggest that a high level of blurring should be used for effective de-identification. While the facial and other features can be masked adequately, the gait and other temporal characteristics are hard to mask. The amount of temporal blurring we tried was not sufficient, given some familiarity with the subjects. Our user study seems to suggest that de-identifying an individual to others familiar with him/her is a very challenging task. Without familiarity, gait and other characteristics are of low value and face plays the most important role.

5 Related Work

In the past, outlines of privacy preserving systems have been presented [8, 9] to highlight the issues. These were sketches and not reports of an implemented de-identification system. Most privacy protection schemes focus on faces [10–13]. Commonly used schemes rely on methods which work well against human vision such as pixelation and blurring. [14, 15] studied the effectiveness of blur filtration. Neustaedter et al. [15] concluded that blur filtration is insufficient to provide an adequate level of privacy. More recent methods such as the *k*-Same [10] and *k*-Same-Select [11] provide provable privacy and preserve data utility. Face modification has also been attempted as a way of image manipulation in [16–18]. The focus of these methods is seamless transfer of information from one or more input images to the target image. De-identification is a very different problem. The focus is on destroying all identifiable features from the image, which requires less effort than a seamless face substitution algorithm.

There has been very little work in the past dealing with entire human body for de-identification. Chen et al. [19] proposed a method for human body obscuring using motion history information of the edges. This method hides the identity of the actor, but it also removes all the information on the action. Park et al. [20] introduced the concept of *personal boundary* in a context adaptive human movement analysis system. Foreground pixels form multiple coherent colour blobs that constitute a human body. These blobs are used for blocking human identity. The problem with this approach is that it preserves the overall silhouette of the person which can aid recognition.

Another technique used for protecting privacy is based on segmenting the privacy information from a video and encrypting it to hide it from the end user. Different frameworks have been proposed to hide the data in the video itself, e.g., as a watermark [21] or as encrypted information in DCT blocks [22]. This information can be retrieved later on request. Prototype designs have also been proposed to provide a variable amount of control to the users over the information viewed in a video [15, 21] which is a requirement of an ideal de-identification scheme.

Detecting and segmenting humans in images and videos is a very active area of research today which may help a complete de-identification system [23, 24]. Recognizing humans from faces, silhouettes, gait, etc.is also an active area; success in those provides more methods a de-identification system should guard against.

6 Conclusions

In this paper, we analyzed the issues relating to de-identification of individuals in videos to protect their privacy by going beyond face recognition. We also presented a basic system to protect privacy against algorithmic and human recognition. We present results on a few standard videos as well as videos we collected that are more challenging to hide identity in. We also conducted a user study to evaluate the effectiveness of our system. Our studies indicate that gait and other temporal characteristics are difficult to hide if there is some familiarity with the subjects and the user. Blurring is a good way to hide the identity if gait is not involved. We propose to conduct further studies to evaluate the de-identification system against recognition by computer vision algorithms. That is likely to be easier than guarding against manual identification of individuals.

Acknowledgements: Partial funding was provided by the Naval Research Board of India.

References

1. Collins, R.T., Gross, R., Shi, J.: Silhouette-based human identification from body shape and gait. In: Proceedings of IEEE Conference on Face and Gesture Recognition. (2002) 351–356
2. Yoo, J.H., Hwang, D., Nixon, M.S.: Gender classification in human gait using support vector machine. In: ACIVS. (2005) 138–145
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (1), IEEE Computer Society (2005) 886–893
4. Tu, P., Sebastian, T., Doretto, G., Krahnstoever, N., Rittscher, J., Yu, T.: Unified crowd segmentation. In Forsyth, D.A., Torr, P.H.S., Zisserman, A., eds.: ECCV (4). Volume 5305 of Lecture Notes in Computer Science., Springer (2008) 691–704
5. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV. (2001) 105–112
6. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3) (2004) 309–314
7. Cabral, B., Leedom, L.C.: Imaging vector fields using line integral convolution. In: SIGGRAPH '93: Proc. on Computer graphics and interactive techniques, ACM (1993) 263–270
8. Senior, A.W.: Privacy enablement in a surveillance system. In: ICIP. (2008) 1680–1683
9. Yu, X., Chinomi, K., Koshimizu, T., Nitta, N., Ito, Y., Babaguchi, N.: Privacy protecting visual processing for secure video surveillance. In: ICIP. (2008) 1672–1675
10. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying facial images. IEEE Transactions on Knowledge and Data Engineering **17** (2003) 232–243
11. Gross, R., Airoldi, E., Malin, B., Sweeney, L.: Integrating utility into face de-identification. In: Privacy Enhancing Technologies. (2005) 227–242
12. Gross, R., Sweeney, L., la Torre, F.D., Baker, S.: Semi-supervised learning of multi-factor models for face de-identification. In: IEEE CVPR. (2008)
13. Phillips, P.: Privacy operating characteristic for privacy protection in surveillance applications. (2005) 869
14. Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: CSCW. (2000) 1–10
15. Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. ACM Trans. Comput.-Hum. Interact. **13**(1) (2006) 1–36
16. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S.M., Colburn, A., Curless, B., Salesin, D., Cohen, M.F.: Interactive digital photomontage. ACM Trans. Graph. **23**(3) (2004) 294–302
17. Blanz, V., Scherbaum, K., Vetter, T., Seidel, H.P.: Exchanging faces in images. Comput. Graph. Forum **23**(3) (2004) 669–676
18. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. ACM Trans. Graph. **27**(3) (2008) 1–8
19. Chen, D., Chang, Y., Yan, R., Yang, J.: Tools for protecting the privacy of specific individuals in video. EURASIP Journal on Advances in Signal Processing **2007**(1) (2007) 107–107
20. Park, S., Trivedi, M.: A track-based human movement analysis and privacy protection system adaptive to environmental contexts. (2005) 171–176
21. Zhang, W., Cheung, S.C.S., Chen, M.: Hiding privacy information in video surveillance system. In: ICIP (3). (2005) 868–871
22. Cheung, S.C.S., Paruchuri, J.K., Nguyen, T.P.: Managing privacy data in pervasive camera networks. In: ICIP. (2008) 1676–1679
23. Ren, X., Berg, A.C., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: ICCV, IEEE Computer Society (2005) 824–831
24. Mori, G., Malik, J.: Recovering 3d human body configurations using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **28**(7) (2006) 1052–1062