DE-IDENTIFICATION FOR PRIVACY PROTECTION IN SURVEILLANCE VIDEOS

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Electronics and Communication

by

Prachi Agrawal
200431006
prachi@research.iiit.ac.in

Center for Visual Information Technology
International Institute of Information Technology
Hyderabad, India
June 2010

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "De-identification for Privacy Protection in Surveillance Videos" by Prachi Agrawal, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. P. J. Narayanan

To My Parents

Acknowledgments

I would like to thank my guide, Dr. P. J. Narayanan, for his support and guidance during my research over the past few years. He is actively involved in the work of all his students and his enthusiasm for research is infectious. I was fortunate enough to have him as my advisor.

I am grateful to all the people who gave their time for the numerous dataset collection sessions and the 'preliminary' user studies. I thank Chandrika, Dileep, Maneesh, Mihir, Naveen, Pavan, Poornima, Pranav, Pratyush, Rishabh, Sheetal, Suhail, Supreeth, and Vidhya for their extreme patience and support. I also appreciate their valuable comments during the early user studies that helped me in improving the overall system. Special thanks to Himanshu and Sheetal for helping me with the figures, and Naveen for his help with the last minute video making during ACCV submission. I would also like to thank Mrityunjai for his help in putting up a demo for the R&D showcase. I am grateful to all my labmates at CVIT for their stimulating company. The time I spent in this lab is cherishable.

I am also thankful to my friends - Varun, Samba, Pinky, Sheetal, Supreeth - for being patient with me and giving me all the push I needed. I thank Chhaya, my best friend of over 9 years, who was with me all these years during my BTech and MS. Without her, life would not have been the same in college. Above all, I am thankful to my family for their unconditional support and undying encouragement in every sphere of my life. Whenever I was overwhelmed by an obstacle, academic and otherwise, it was their absolute confidence in me that kept me going. Without their endless support and patience, this thesis would not have been a reality.

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. Also, an increasing number of video cameras observe public spaces like airports, train stations, shopping malls, streets, and other public places today. Video surveillance is an important tool to prevent, detect and reduce crime and disorder. It is also widely used to monitor people's conduct, detect their presence in an area and/or possibly study their actions too. However, many civil rights and privacy groups have expressed their concern that by allowing continual increases in surveillance of citizens, we will end up in a mass surveillance society, with extremely limited, or non-existent political and/or personal freedoms. With the widespread use of surveillance today, we are becoming used to "being watched". As time goes on, the more accustomed we become to video surveillance, we are not as likely to fight to maintain our right to privacy. Critics believe that in addition to its obvious function of identifying and capturing individuals who are committing undesirable acts, surveillance also functions to create in everyone a feeling of always being watched, so that they become self-policing. At the same time prices of hardware have fallen, and the capabilities of systems have grown dramatically. The day is not far when it would be easy to convert even the low-cost video surveillance units into "intelligent" human recognition systems. These raise concerns on the unintentional and unwarranted invasion of the privacy of individuals caught in the videos.

The data collected during video surveillance consist mainly of images and sounds which allow identification of people captured, whether directly or indirectly, in addition to monitoring their conduct. While there may be a possible security need to identify the individuals in these videos, identifying the action suffices in most cases. The actor needs to be identified only rarely and only to authorized personnel. The privacy concerns related to the processing and distribution of the collected videos are genuine and will grow with wider adaptation of video technology. 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 ideal goals are to obscure the identity of the actor without obscuring the action. There is a natural trade-off between protecting privacy and providing sufficient detail. The goal is to protect the privacy of the individuals while providing sufficient feel for the human activities in the space being imaged.

The policies on the privacy issues are still unclear and are fast evolving. The need of the hour is to be cautious as digital data has the potential to be replicated with little control, especially when placed on the cloud. This thesis outlines the scenarios in which de-identification is required and the issues

brought out by those. We also present an approach to de-identify individuals from videos. Our approach involves tracking and segmenting individuals in a conservative voxel space involving x, y and time. A de-identification transformation is applied per frame using these voxels to obscure the identity. A robust de-identification scheme with a randomization module was designed in order to make reconstruction and comparison based identification attacks impossible. Face, silhouette, gait, and other characteristics need to be obscured, ideally. We show results of our scheme on a number of videos and for several variations of the transformations. We present the results of applying algorithmic identification on the transformed videos. We also present the results of a user-study to evaluate how well humans can identify individuals from the transformed videos. The results showed that as the parameter controlling the amount of de-identification increased, the actors became less identifiable (more privacy), while the video started losing the context and detail. The studies also suggest that an action can be recognized with more accuracy than the actor in a de-identified video, which is the guiding principle of de-identification. We also created a suitable dataset for the purpose of user studies, as the standard datasets available online are not challenging enough for such an evaluation.

Contents

Chapter Page

1 Privacy Protection: Need and Requirements . . . 1
1.1 Different Scenarios: Importance of Privacy Protection . . . 3
1.1.1 Casual Videos . . . 3
1.1.2 Public Surveillance Videos . . . 4
1.1.3 Private Surveillance Videos . . . 5
1.2 De-identification: Our Framework . . . 5
1.2.1 Focus on Obfuscating Features . . . 6
1.2.2 Focus on Preserving Context . . . 7
1.2.3 Usability of Videos . . . 8
1.3 Contributions . . . 8
1.4 Organization of thesis . . . 9

2 Previous Work . . . 11

3 De-identification: Related Issues . . . 17
3.1 Identification and Verification . . . 21
3.2 Subverting De-identification . . . 22
3.3 Storage of Videos . . . 23

4 Person De-identification in Videos . . . 24
4.1 Detect and Track . . . 24
4.2 Segmentation . . . 25
4.3 De-identification . . . 27
4.4 Randomization . . . 28
4.5 Experimental Results . . . 29
4.5.1 Algorithmic Evaluation . . . 30

5 User Study . . . 35
5.1 Limitations . . . 41
5.2 Discussion . . . 42

6 Conclusions and Future Work . . . 43

Bibliography . . . 47

List of Figures

Figure Page

1.1 Many surveillance systems are widely used today for the purpose of monitoring and ensuring peace by combating crime. However, such systems are also a threat to a common man's privacy. 2
1.2 Images taken from "Face Swapping" [4]. While the faces are cleanly replaced by someone else's, still manual identification is possible to a large extent. 6
1.3 Silhouette can help in identification. 6
1.4 Even a silhouette can help in identifying the gender. 7
1.5 Different possible de-identification operators. Images taken from PriSurv [42]. 10

2.1 Original image (a) de-identified using several ad-hoc de-identification techniques ((b) to (n)). Other than blocking out the entire image, as in (k), the experiments show that ad hoc attempts do not thwart face recognition software. Image taken from "Preserving Privacy by De-identifying Facial Images" [29]. The example shows what looks convincing to the human eye might not be good against automatic recognition. 12
2.2 Some methods may work well against algorithmic identification, but might not be good against manual identification. The images show original images and their corresponding de-identified versions. 13
2.3 Example images to emphasize on the trade-off between privacy protection and information content. High privacy protection results in low or zero information content. The images on the left are original while images on the right are their de-identified versions. 14
2.4 Some data hiding techniques preserve privacy as well as action. The left most images are the original images, while the corresponding images on the right are de-identified using different parameters to control the level of de-identification. 15

3.1 Example images of people from the LFW data set. 18
3.2 Human Face Verification Results on LFW. Image taken from "Attribute and Simile Classifiers for Face Verification" [24] 19

4.1 Overview of the method. 24
4.2 (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. 28

4.3 The first row shows the clear frames. The next five rows show the output of Blur-2, Blur-4, LIC-10, LIC-20, and Blur-2 followed by an intensity space compression, in that order. 32
4.4 Results on two different videos. The clear frames are shown in the odd columns while corresponding de-identified frames in the even columns. 33
4.5 De-identified frames showing people performing different activities; the activity is recognizable but the person is not. 34

5.1 Screenshot of the portal used for the user study for Identification 36
5.2 Screenshot of the portal used for the user study for Search 37
5.3 Segmentation result on a video with dynamic background. 41

List of Tables

Table Page

4.1 Percentage of correct answers for the face and human detectors. 30

5.1 Number of correct identifications for *search* and *identification* experiments in the user study. Sets A to H had 9, 8, 11, 9, 9, 8, 10, and 10 users respectively. 38
5.2 Number of correct identifications in the user study on familiar people. 40
5.3 Human experience scores on a scale of 1 (low) to 7 (high). 40

Chapter 1

Privacy Protection: Need and Requirements

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 new 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[1], EveryScape[2], Mapjack[3], 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 furore over Street View in Japan and the UK underscores the need to address the privacy issue directly. Online photo sharing is growing more and more around the world. Many social networking websites and online image galleries like Facebook, Orkut, MySpace, Picasa, Flickr, etc., allow the users to share their images and videos with others. Due to the nature of these sites, if a user has not opted for a protected privacy setting, then all the existing users of the website have access to his or her image gallery. When Bob adds a third-party application on these sites, the application is given the ability to see anything that Bob can see. This means that the application is unknowingly granted access to the online photo albums of Bob and his fellow network members along with other information. The owner of the application is free to collect, look at, and potentially misuse this information. In most cases, these photo albums are meant for friends, family and acquaintances only. Moreover, in Bob's photo album, people other than Bob himself are captured, and these people have no control over the privacy setting of their images. If a certain person wishes to not have his or her pictures put online, there must be a way to remove his or her identity from these images, without affecting the context and picture quality adversely.

Video surveillance is a part of everyday life today. An increasing number of video cameras observe public spaces like airports, train stations, shopping malls, streets, and other public places. The data collected during video surveillance consist mainly of images and sounds which allow identification of people captured, whether directly or indirectly, in addition to monitoring their conduct. While there may

[1] http://maps.google.com
[2] http://www.everyscape.com
[3] http://www.mapjack.com

be a possible security need to identify the individuals in these videos, identifying the action suffices in most cases. The actor needs to be identified only rarely and only to authorized personnel. Thus, video surveillance raises privacy and data protection issues. With the growing use of video surveillance techniques by an increasing number of entities, there is a need for striking a balance between security and privacy requirements.

Video surveillance was initially used for traffic purposes. With time, CCTV cameras were especially refined for workplace surveillance. It became possible to monitor the actions of the employees and improve the regularity of labour performance as well as productivity. Employers make several arguments to justify their use of workplace surveillance, ranging from ownership rights to providing a safe work environment to their employee. What makes matters worse is that these employers are not required by law to disclose to their employees what kind of monitoring is being conducted. The employees are increasingly worried about a potential privacy abuse. The terror of unknown surveillance techniques and unclear laws about a privacy breach is so high that makes people think that even though bathrooms and locker rooms can be expected to be free of surveillance, even this protection is not absolute.

Figure 1.1 Many surveillance systems are widely used today for the purpose of monitoring and ensuring peace by combating crime. However, such systems are also a threat to a common man's privacy.

The privacy issues are genuine and will grow with wider adaptation of video technology. The policies on these issues are still unclear and are fast evolving. The need to *de-identify* individuals from such images and videos is obvious. The need of the hour is to be cautious as digital data has the potential to be replicated with little control, especially when placed on the cloud.

De-identification is a process which aims to remove all identification information of the person from an image or video. Recognition and de-identification are opposites with the former making use of all possible features to identify and the latter trying to obfuscate the features to thwart recognition. 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 kind of scenarios we described earlier need the information on the action and context to be preserved for the video to be of any use. Hence, we refine the definition of de-identification according to which de-identification is a process which 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*. There is a natural trade-off between protecting privacy and providing sufficient detail. The goal is to protect the privacy of the individuals while providing sufficient feel for the human activities in the space being imaged.

Face plays a dominant role in automatic and manual identification. De-identification aims to obfuscate faces in images to make them unrecognizable. However, videos present more challenges as they capture more information (silhouette, posture, gait, peculiar habits, etc.) as compared to images which can also be used for identification. Humans exploit this information effectively and algorithmic identification schemes using body silhouette and gait have been developed with some success in the past [13, 41]. Hence, an ideal de-identification transformation should aim to obfuscate all these features too, along with the face, that help in identification.

The de-identification should be resistant to recognition by both, humans and algorithms. Automated methods to de-identify individuals without affecting the context or the action in the video are needed to address them. User-assisted or computationally intensive de-identification methods would be difficult to deploy in a surveillance scenario where the output from a camera is constantly being watched on a monitor. It also may be necessary to control the level of de-identification to cater to different situations. We now discuss the different scenarios where de-identification might be necessary.

1.1 Different Scenarios: Importance of Privacy Protection

Three types of videos compromise with the privacy of individuals. The threat to privacy, seemingly small in some cases, can evoke serious concerns of those involved and need to be addressed.

1.1.1 Casual Videos

These are videos that are captured for other purposes and get shared. In such videos, either people are captured unintentionally, or they get shared over the internet without their consent, or both. Examples include the net-cameras fitted in public spaces that can be viewed over the internet, videos or photos on sharing sites, images used by projects like Google Street View, etc. Privacy advocates have objected to the Google Street View feature, pointing to views which show people engaging in activities visible from public property in which they do not wish to be seen publicly. The seriousness of the issue and the need to de-identify individuals in such videos can be gauged by the fact that some nations have ordered the firm to halt plans to photograph their streets until more privacy safeguards were put in place. Residents

of a place in England formed a human chain to stop a Google camera van from entering the village. A man from a city in Austria threatened the driver of a Google camera van with a garden pick.

It is obvious that these photographs were taken from public property and it was not Google's intention to capture these individuals in these images. Google allows users to flag inappropriate or sensitive imagery for it to review and remove. However, the privacy advocates argue that if the onus is placed on the individual after an offending image has been published to take steps to remedy the situation, the core purpose of the Data Protection Act will have been defeated. 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.

1.1.2 Public Surveillance Videos

These videos come from cameras watching spaces such as airports, streets, stores, etc. and people are aware that they are continuously being watched. Arguably, a Closed Circuit Television (CCTV) system is an important tool to assist with efforts to combat crime and disorder whilst enhancing community safety. However, it may also be regarded as the most potent form of privacy breach. It also compromises with people's right to freedom of movement - freedom to move without leaving continued traces of one's movement. The fear of always being under the watch may influence people's conduct and activities in places known to be under surveillance.

To judge the quality of images necessary for a CCTV system, it is important to identify the purposes for which CCTV is used. A CCTV system can be employed for one of the below listed purposes:

- **Monitoring:** To watch the flow of traffic, or movement of people in a shop, etc. For example, CCTV at street corners.
- **Detecting:** To detect the presence of a person in an image or video. For example, camera controlled automatic doors.
- **Recognizing:** To recognize someone you know, or determine that somebody is not known to you. For example, CCTV used at the entrances of semi-private spaces and homes.
- **Identifying:** To identify someone beyond reasonable doubt. For example, as proof in robbery cases in banks.

In two out of four purposes listed above, 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 and 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 irrevocable de-identification is necessary. A revocable de-identification should suffice for other purposes.

1.1.3 Private Surveillance Videos

These come from cameras placed at the entrances of semi-private spaces like offices and homes. 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. Video surveillance systems are the most pervasive and common workplace surveillance systems employed in large corporations today. Private information including identities of individuals, activities, routes, etc., are routinely monitored. What makes matters worse is that the employers are not required by law to disclose to their employees what kind of monitoring is being conducted. While the surveillance might be justified by the employers on many grounds, ranging from ownership rights to providing a safe work environment to their employees, misuse of private information about trusted employees can severely hamper their morale and may even lead to unnecessary litigation. Since access is often limited to authorized persons only, de-identification may not be essential, but could be recommended to take care of potential viewing by unauthorized people.

1.2 De-identification: Our Framework

De-identification involves the detection and a transformation of images or videos of individuals to make them unrecognizable, without compromising on the action and other contextual content. There is a natural trade-off between protecting privacy and providing sufficient detail. The goal is to protect the privacy of the individuals while providing sufficient feel for the human activities in the space being imaged. Privacy protection provided should also be immune to recognition using computer vision as well as using human vision.

Ideally, it must be possible to improve the control of provided privacy, the quality and the clarity of videos based on the requirement, context and action. A variable de-identification parameter could control the amount of de-identification provided, as required. This parameter could either be controlled by a manual feedback system, or an automatic action-context recognition system. The context recognition system can ensure the usability of the videos stored, as identification of action is necessary in de-identified videos. The feedback from such a system can control the amount of privacy provided based on the action and context before storing the videos. A reversible de-identification algorithm with an encryption key is another possible approach towards balancing privacy and security requirements. Videos once stored in a particular de-identified format can be re-identified using the key in case of a security breach. Most importantly, an ideal de-identification system when implemented, needs to give reassurance to those whose images are being captured that their private information is safe.

1.2.1 Focus on Obfuscating Features

Identifying information captured on video can include direct features such as face, silhouette, posture, gait, etc., and discernible features such as race, gender, etc. 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 a 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.

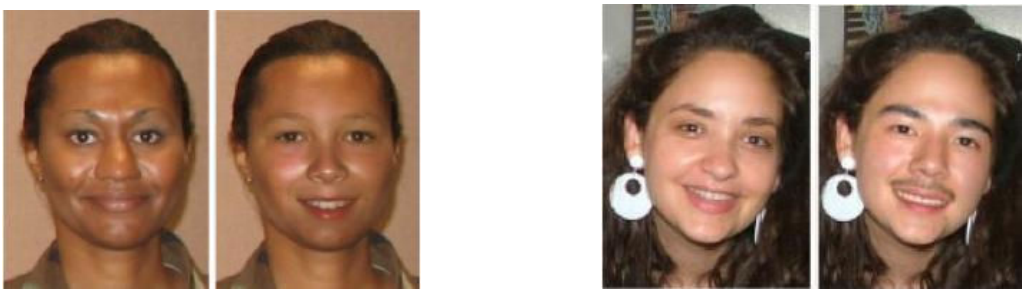


Figure 1.2 Images taken from "Face Swapping" [4]. While the faces are cleanly replaced by someone else's, still manual identification is possible to a large extent.

2. The body silhouette and the gait are important clues available in videos which need to be obfuscated. Humans exploit them effectively and algorithmic identification schemes using them have been developed with some success [13, 41]. While obfuscating the silhouette with a small loss of detail is easy, gait is hard to hide. The silhouette can be dilated or expanded to remove its information content while a tight segmentation of the individuals may preserve the silhouette. 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.

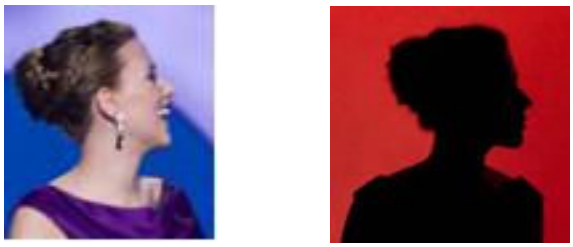


Figure 1.3 Silhouette can help in identification.

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, they 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. We do not address gender or race hiding in this work, though they may pose critical privacy issues.

Figure 1.4 Even a silhouette can help in identifying the gender.

1.2.2 Focus on Preserving Context

Privacy protection is achieved by hiding the subjects' faces, body, silhouette, clothes, etc. However, there is an obvious trade-off between privacy and security. Excess obscuration makes video surveillance meaningless. Context preservation is important for the usability of the videos captured. Steganographic or cryptographic methods involve detecting the private information in the video, and hiding it within the video [11, 43]. This information can be recovered at a later stage, if needed. The system is designed such that only authorized people have access to the private information, while others can not see the subject in the video at all. Such methods are very effective at providing privacy and also preserve the usability of the data. The private information can be encrypted and hidden using a secure key. However, such videos lose all context information and make the video useless for viewing in real-time. In many surveillance scenarios, the captured videos are monitored by an observer in real-time. In such scenarios, analyzing the action being performed in the video might be of importance for security and other purposes. De-identification should aim to remove only the identity information from a video, while retain as much visual information on the action or context.

In another work on privacy protection [42], an interesting analysis was conducted on personal sense of privacy from the viewpoint of the relationship between a viewer and a subject. Their results suggest that different people have different privacy requirements: some people may want their privacy to be protected from certain viewers, while others may not care about their privacy at all. Keeping in mind

the above analysis, different levels of de-identification must be provided to the subjects appearing in videos. The authors employed twelve different abstraction (de-identification) operators to control visual information in the videos. The privacy provided by these operators varied gradually (Figure 1.5: (a) through (l)). The first eight operators preserved the context in the video, but operators in (i) to (l) completely remove all context information from the processed video along with the identity. An alternative to such operators is to either replace the subject with a stick figure or tag the video with information that does not reveal the identity but conveys the action or context nonetheless (Figure 1.5: (m) to (o)). However, such techniques either require precise position and pose tracking for a stick figure, or accurate real-time gender, age and action detection. Such high level and computationally intensive vision technologies are beyond the reach of current surveillance technologies. An ideal de-identification operator should preserve both, privacy and context, in a video in real-time. With growing computational power everyday, we might be able to attain such a system in future.

1.2.3 Usability of Videos

A protected video should be legitimate for further computer vision tasks as well as manual monitoring. The de-identification transformation should not make the video meaningless by removing all context from it, or render it unviewable by distorting the naturalness. Preliminary operations like motion detection, object tracking, action analysis, etc., should be possible in a de-identified video. A person watching the live streaming of such a video should be able to analyze the subject's behaviour. Actions such as the subject entering a certain area of a camera's field of view, or picking up an object, or abandoning the luggage, or even acts of aggression should be identifiable. An unusual behaviour may be indicative of a security threat and an alarm could be raised well in time to avoid a mishap. Finally, the de-identification process should also support untransformed video to be viewed if the situation demands. One approach is to store the original video, with sufficiently hard encryption, along with the de-identified video. The required keys for decryption is available only with authorized persons. This needs additional storage space, which can be reduced by saving only the portions that contain humans.

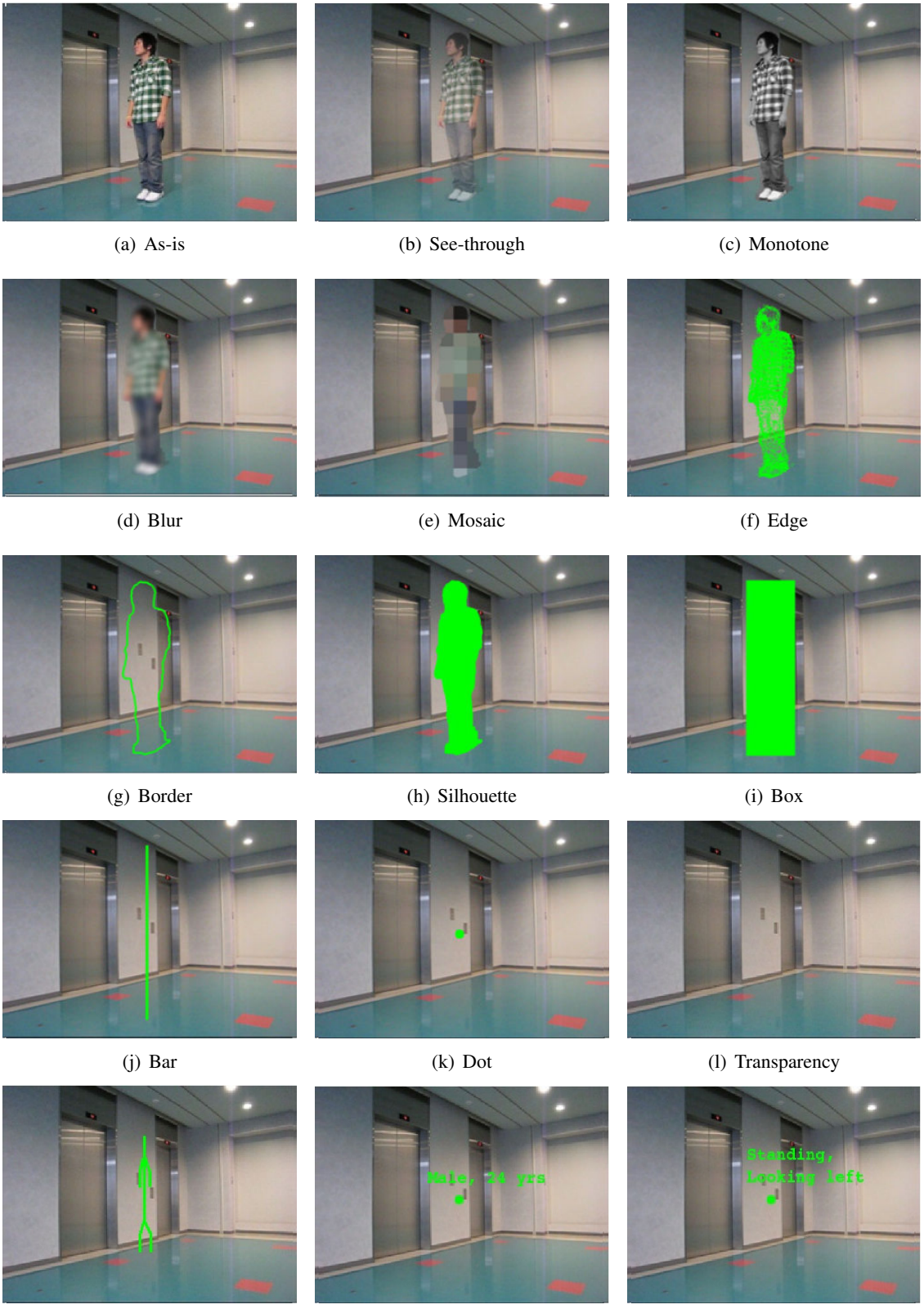
1.3 Contributions

In this thesis, we discuss issues relating to de-identification of individuals in videos. The goals of an ideal de-identification algorithm are to hide identity while preserving action and context. We analyzed the issues relating to de-identification of individuals in videos to protect their privacy by going beyond face recognition. We strive to guard against both, algorithmic and manual identification. We evaluated the performance of current computer vision algorithms on the de-identified videos to test the proposed schemes against algorithmic identification. We also present results from a user study conducted to gauge the effectiveness of the strategy and its robustness against manual identification. Following are the key contributions of this work:

1. Definition and elucidation of the concept of de-identification as a privacy protection tool that does not compromise on the action and other context. Different scenarios where de-identification is necessary and the issues brought out by those were outlined.
2. Designed a robust de-identification scheme with a randomization module in order to make reconstruction and comparison based identification attacks impossible. The scheme was tested on several standard and relevant videos.
3. Conducted elaborate user studies in order to check the effectiveness of the system and usability of the de-identified videos. Also, an algorithmic evaluation was conducted to check the robustness of our system against common computer vision algorithms.
4. Created a suitable dataset to test the effectiveness of the system as the standard datasets available online are not challenging enough for such an evaluation.

1.4 Organization of thesis

Previous work on de-identification is given in Chapter 2. Chapter 3 outlines the various aspects related to de-identification. Chapter 4 explains our method and presents the experimental results on several standard and relevant videos. Chapter 5 present a detailed user-study while the concluding remarks are in Chapter 6.



(a) As-is (b) See-through (c) Monotone

(d) Blur (e) Mosaic (f) Edge

(g) Border (h) Silhouette (i) Box

(j) Bar (k) Dot (l) Transparency

(m) Stick (n) Annotation-1 (o) Annotation-2

Figure 1.5 Different possible de-identification operators. Images taken from PriSurv [42].

Chapter 2

Previous Work

In the past, outlines of privacy preserving systems have been presented to highlight the underlying issues [35, 36, 42]. These were only sketches and not reports of an implemented de-identification system. However, they provide a useful insight into the design of an ideal de-identification system. Such works help us understand the human psyche and the requirements of a de-identification system. They outline, for example, how it is better to be biased towards false positives and err on the side of caution than have the privacy information of a person revealed. The different surveys conducted on people show that personal sense of privacy differs in different scenarios for different people, and it is largely dependent on who is watching the video. Koshimizu et al. evaluated people's reactions to different types of visual abstractions based on their relationship with the viewer [23], while Yu et al. suggested to integrate the idea of personal sense of privacy in a de-identification system [42].

Most implementations of privacy protection schemes focus on faces [18, 20, 29, 31]. However, face is only one out of a long list of identifiable features of an individual: body structure, silhouette, gait, gender, race, etc., also aid recognition and hence should be masked adequately. Although face de-identification is not enough when it comes to providing privacy (especially in a video), the motivation behind all these schemes was similar to ours: protecting privacy of an individual. Hence, we provide a brief description of the privacy protection schemes implemented in the past.

Commonly used face de-identification schemes rely on methods that work well against human vision such as pixelation and blurring. However, it has been shown that simple blurring techniques which might look good to the eye provide little or no protection from face recognition software [29]. In some cases, the recognition rate might be even better on the 'de-identified' images than on the original ones [19]. More recent methods such as the k-Same [29] and k-Same-Select [18] implement the k-anonymity protection model which provide provable privacy and preserve data utility. In the k-Same [29] approach, a distance metric is computed between images, and k closest images are averaged to create a new face. This method provably limits the ability of face recognition softwares to recognize faces by $\frac{1}{k}$. The k-Same-Select [18] method builds on this approach. The factors which constitute data utility are recognized (expressions, gender, etc.). The input face images are divided into subsets based on these

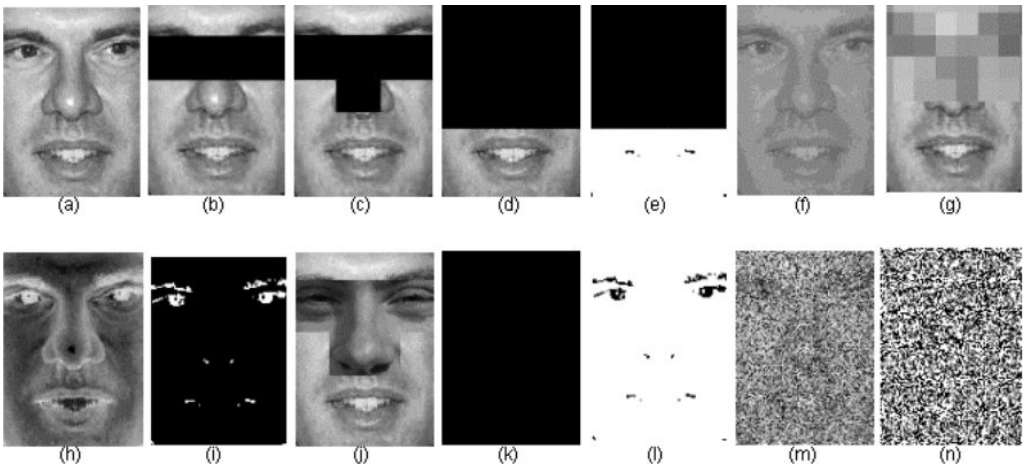


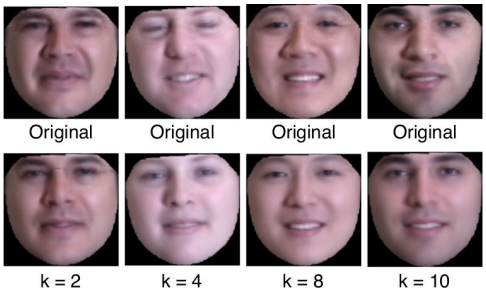
Figure 2.1 Original image (a) de-identified using several ad-hoc de-identification techniques ((b) to (n)). Other than blocking out the entire image, as in (k), the experiments show that ad hoc attempts do not thwart face recognition software. Image taken from "Preserving Privacy by De-identifying Facial Images" [29]. The example shows what looks convincing to the human eye might not be good against automatic recognition.

factors and the *k*-Same method is applied on these subsets individually. This approach preserves utility of data along with the privacy in the de-identified images. Gross et al. later combined a model-based face image parametrization with the formal privacy protection model [20]. They proposed a semi-supervised learning based approach for multi-factor models for face de-identification.

Phillips proposed an algorithm for privacy protection through the reduction of the number of eigenvectors used in reconstructing images from basis vectors [31]. However, it faces the same problem like any other PCA based algorithm: with every new addition in the face database, the eigenvectors need to be calculated again which is not a computationally trivial task. In other work [17], the need for automatic techniques for protecting the privacy of people captured in images by Google Street View was recognized and addressed by a method to obscure the faces and number plates of cars in these images. However, the primary focus of this work is on handling a large scale data and reducing the number of false positives in order to maintain the visual quality of images, while keeping recall as high as possible. However, for a privacy protection scheme to work, it needs to be susceptible to false positives, and not otherwise.

Face modification has also been attempted as a way of image manipulation [2, 4, 5]. Bitouk et al. replace faces from one image into another, by aligning the faces in the two images automatically to a common coordinate system [4]. Blanz et al. estimate the shape, pose and direction of illumination in the target and source faces, and fit a morphable 3D model to each face optimizing all the parameters [5]. They render the new face by transferring the scene parameters of the target image to the source 3D model. However, face modification is different from de-identification in yet another way. The focus

(a) Image taken from "Face Swapping" [4].



(b) Images taken from "Semi-Supervised Learning of Multi-Factor Models for Face De-Identification" [19]. *k* corresponds to the number of neighbours used for averaging.

Figure 2.2 Some methods may work well against algorithmic identification, but might not be good against manual identification. The images show original images and their corresponding de-identified versions.

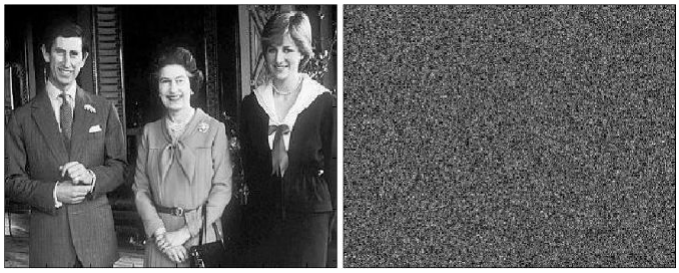
of face modification methods is seamless transfer of information from one or more input images to the target image. De-identification is a very different problem; the focus in de-identification is on destroying all identifiable features from the image, which requires less effort than a seamless face substitution algorithm.

There has been little work in the past dealing with entire human body for de-identification. Chen et al. presented a system to protect the privacy of pre-specified individuals in a video taken in a hospital [10]. They used an automatic people identification system that learned from limited labeled data. They also 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. introduced the concept of personal boundary and incorporated it in a context adaptive human movement analysis system [30]. Foreground pixels are divided into coherent blobs based on colour similarity. Multiple blobs constitute a human body and are tracked across the frames. These blobs are used to block 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 the information to hide it from the end user. Different frameworks have been proposed to hide the private information in the video itself, e.g., as a watermark [43] or as encrypted information in DCT blocks [11]. This information can be retrieved later on request. The main challenge with such object removal techniques lies in recreating occluded objects and motion after the removal of private information. This can be achieved by inpainting the holes created in a seamless manner. These schemes work on the entire human body instead of just faces. They preserve a natural-looking video without any visual artifacts. However, the main problem with these approaches is that they also remove all the visual information content related to action in the de-identified video. Such videos are rendered useless in scenarios where real-time viewing is a requirement. They could benefit from a framework

(a) Image taken from "Tools for Protecting the Privacy of Specific Individuals in Video" [10].

(b) Image taken from "Hiding Privacy Information In Video Surveillance System" [43].

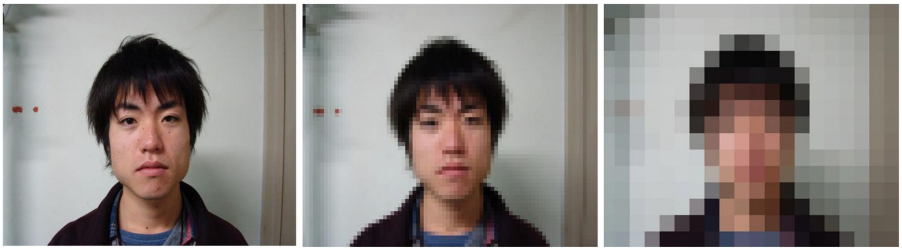


(c) Image taken from "Blind Vision" [3].

Figure 2.3 Example images to emphasize on the trade-off between privacy protection and information content. High privacy protection results in low or zero information content. The images on the left are original while images on the right are their de-identified versions.

that provides a variable amount of control to the users over the information viewed in a video [43]. Neustaedter et al. also suggested a prototype design of a smart system which learns from the visual feedback it receives and provides a varying level of privacy based on the feedback [28].

Cryptographical techniques such as secure multi-party computation have also been proposed to protect privacy of multimedia data [3, 21]. Sensitive information is encrypted or transformed in a different domain such that the data is no longer recognizable but certain image processing operations can still be performed. While these techniques provide strong security guarantee, they are computationally intensive and at the current stage, they support only a limited set of image processing operations. Also, the de-identified videos do not have any information content and are not fit for real-time viewing. An alternative to such schemes which preserve identity but remove all the information content is to replace



(a) Image taken from "Recoverable Privacy Protection for Video Content Distribution" [25].

(b) Image taken from "Scrambling for Privacy Protection in Video Surveillance Systems" [16].

Figure 2.4 Some data hiding techniques preserve privacy as well as action. The left most images are the original images, while the corresponding images on the right are de-identified using different parameters to control the level of de-identification.

the subject with a stick figure [39] or tag the video with information that does not reveal the identity but conveys the action or context nonetheless. However, such techniques either require precise position and pose tracking for a stick figure, or accurate real-time gender, age and action detection. Such high level and computationally intensive vision technologies are beyond the reach of current surveillance technologies.

There has been some work in the past on recoverable privacy protection for video content, where action is not completely lost. The framework is based upon data hiding techniques. For example, the privacy information can be embedded within the privacy protected low resolution image as wavelet coefficients [25], which can be retrieved later. The level of de-identification can be controlled by the scaling parameters which result in the low resolution image. Another approach is based on scrambling the privacy information in transform-domain or codestream-domain [16]. A region-based scrambling can be performed by pseudo randomly flipping the sign of transform co-efficients (or inverting some bits of the codestream) while encoding. This scrambling depends upon a private encryption key and is completely reversible. The level of scrambling can be adjusted by restricting it to fewer coefficients, allowing from mild fuzziness to complete noise. In both these methods, the information can be later retrieved using a key which must be available with entrusted parties.

There have been studies in the past to evaluate the performance of simple privacy protection techniques for day-to-day home-office situations and people's perception of de-identification in general. Boyle et al. showed that blur filtration balances privacy and awareness for day-to-day office situations [8]. Neustaedter et al. showed that blur filtration is insufficient to provide an adequate level of

privacy for risky home situations [28]. They concluded from a survey that people will be suspicious of any system initially but could learn to trust it after a period of usage, like Active Badge System [40].

The first important step in most privacy protection systems is to identify individuals whose privacy needs to be to protected. While face recognition is the most obvious technique, the miss rate of such recognition systems might be high as typical surveillance systems have low-resolution cameras. An alternative is to use specialized visual markers to enhance recognition. For example, Schiff et al. identify individuals wearing yellow hats to address privacy concerns in real-time [34]. However, we maintain that all individuals should be provided with sufficient privacy protection by default, unless otherwise stated. Hence, we do not employ privacy information identification technique in our de-identification system.

Detecting and segmenting humans in images and videos is a very active area of research today which may help a complete de-identification system [27, 32]; success in those would mean less misses in detection of a human by a de-identification algorithm. 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. A de-identification system needs to keep its pace with the ongoing progress in these fields.

Chapter 3

De-identification: Related Issues

What is privacy? Certainly it is much more than one's name, address, and passport number. It includes one's daily routine, shopping habits, web surfing habits, medical history, work history, credit details and much more. Everyone has a right to privacy. This right ensures that they have total control of their personal information, control over who has it and what can be done with it. However, with the growth of digital technology, people are worried that their personal information is compromised everyday even without their consent. Our computers store cookies when some sites are visited. This enables these sites to track our web surfing habits. Many advertising websites use these cookies to track the users' interests and show appropriate ads in their browser which might interest them. Credit card companies, in order to spot potential customers, keep track of people's shopping habits and mortgage history. For the same reasons, insurance companies keep track of people's medical history. While it might not sound like a big deal to some of us, people have been victims of identity theft or harassment or even stalking as a result of one or all of these privacy breaches. In short, we are facing a privacy (and potentially security) breach in every sphere of life these days.

Advances in cameras and web technology have made it easy to capture and share large amounts of video data over the internet. Another such technology which makes it even easier to access photographs of far off regions in the world is Google Street View. Google Street View is a technology featured in Google Maps and Google Earth that provides panoramic views from various positions along many streets in the world. Ever since its launch, many people all around the world have raised privacy concerns, which include being spotted in a public place where they do not wish to be seen publicly, easy access for a potential break-in into a private property, etc. Some parents have expressed concern over Street View compromising the security of their children. In the UK, when images were found of a man leaving an adult bookstore, a man vomiting and another man being arrested, the service drew criticism. However, Google maintained that the photos were taken from public property. Before launching the service, Google removed photos of domestic violence shelters, and it allowed users to flag inappropriate or sensitive imagery for Google to review and remove. When the service was first launched, the process for requesting that an image be removed was not trivial, but Google has changed its policy to make

Figure 3.1 Example images of people from the LFW data set.

removal more straightforward. Another of Google's responses to concerns about privacy laws has been a pledge to blur the faces of people and vehicle license plates filmed on Street View photos.

However, Privacy International, a UK-based human rights "watchdog", sent a formal complaint about the service to the UK Information Commissioner's Office (ICO), which cited more than 200 reports from members of the public who were identifiable on Street View images. The report said that even though Google had assured that privacy would be protected by blurring faces and license plates, the system failed to do so on many occasions. According to the report, in many cases even when the system worked and blurred the faces, it was not enough and people were still recognizable. A few months later, Switzerland's Federal Data Protection and Information Commissioner also announced that his agency would sue Google because in Street View "numerous faces and vehicle number plates are not made sufficiently unrecognizable from the point of view of data protection". Apart from the face, images and videos capture directly identifiable features such as silhouette, posture, gait, etc., and certain discernible features such as race, gender, etc., which can not be masked by blurring faces.

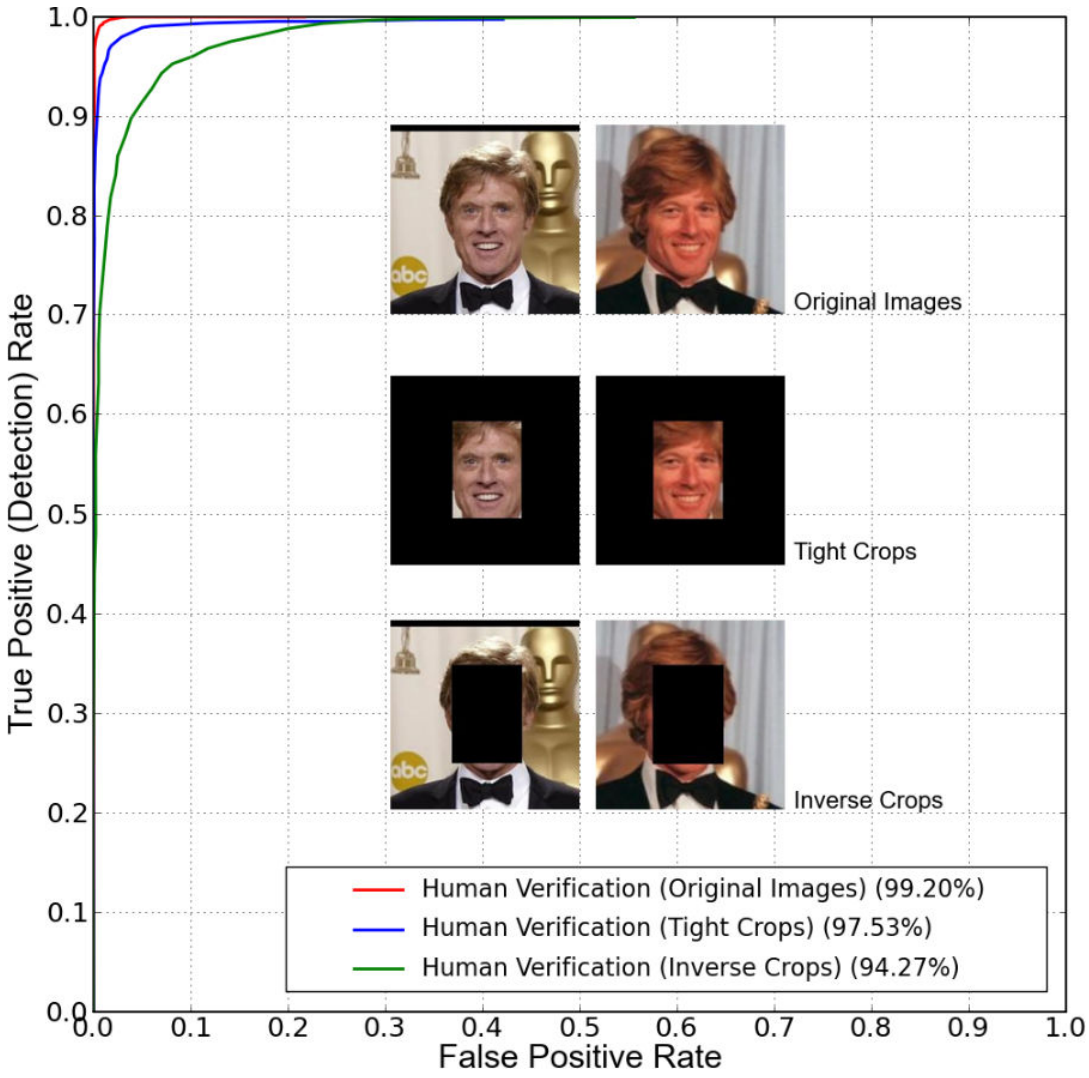


Figure 3.2 Human Face Verification Results on LFW. Image taken from "Attribute and Simile Classifiers for Face Verification" [24]

In fact, in a recent work, Kumar et al. [24] conducted a study where they measure human performance on the LFW data set. They conducted three different tests and concluded that context plays an important role in recognition by humans, apart from the face. For the 6,000 image pairs in LFW, they gathered replies from 10 different users for each pair, for each of the three tests (a total of 240,000 user decisions). For each of these tests, the users were asked to label the pair of images shown as belonging to the same person or not, and rate their confidence in doing so. Example images of people from the LFW data set used in the performance evaluation are shown in Figure 3.1. The images are of well known public figures.

First the test was performed on the original LFW images. The result is shown by the red line in Figure 3.2. Because of the nature of these images (some images of individuals were taken with the same

background) and as face verification is a fairly easier task than recognition, the users achieved a 99.2% accuracy. For the second test, the images were tightly cropped around the face (eyes, nose, mouth, and sometimes ears, hair, and neck), and the rest of the image was blacked out. This was done to remove the context, background, and hair from the images that could help in recognition. The results are shown in blue in Figure 3.2. The recognition accuracy dropped to 97.53%, showing that people are affected by the context and background while making a judgement. In order to confirm that the region outside the face was indeed helping people with identification, they conducted a third test. For this test, which was the opposite of the second one, the region inside the face was blacked out while the remaining part of the image showed. People still obtained 94.27% accuracy, as shown by the green line in Figure 3.2. This study showed that context, background, hair, etc., help humans in recognition. The aim of their study was to assert that automatic face verification algorithms should not use the region outside the face, as it could affect the accuracy in a manner not applicable on real data. However, it also asserts the underlying principle on which this work is based: face de-identification is not enough when it comes to privacy protection of people.

Need for de-identification increases with the dramatically increasing use of surveillance cameras throughout the public and private sectors and the decreasing cost of putting up such a system. Widespread implementation of low-cost video surveillance is harmful for several reasons, apart from the obvious privacy breach. We are becoming used to being watched, and at earlier and earlier ages. Many schools have installed video monitoring throughout their campuses. An increasing number of day care centers are connected to the Internet so parents can check in on their children. As time goes on, the more accustomed we become to video surveillance, we are not as likely to fight to maintain our right to privacy. Moreover, with the growth in digital technologies, the cost of biometrics systems is also bound to decrease. The day is not far when it would be easy to convert even the low-cost video surveillance units into human recognition systems.

There are ongoing efforts to minimize the privacy and security risks of video surveillance to more acceptable levels. Some of these efforts include strengthening legal oversight mechanisms, improving awareness, and educating the masses. However, the addition of privacy-enhancing technologies could enable individuals to manage their own personally identifiable information and minimize privacy risks at an earlier level. There exist techniques in various domains that provide tools to protect one's private information. Knowledgeable individuals might be able to control the amount of private information in images before they share them online. But, realistically, few people have the requisite knowledge, awareness, or patience to take advantage of such privacy-enhancing strategies. It is important to educate the masses about the potential threat of privacy and security breach from sharing their images online and making privacy-protection tools available to them.

However, in surveillance scenarios, where one has no control over the video being captured, it is best to not leave these technologies to be employed (or requested) by the consumers at their will, but incorporate them into the video surveillance systems from the beginning based on their implied use. It is better to err on the side of caution than to compromise with the privacy of people. For instance, a

surveillance system which relies on the streaming video for taking an immediate action can be encrypted using a visual domain de-identification technique. However, if it suffices to store the video, and later view the contents in case a mishap has occurred, a data-encryption technique can be employed at the camera itself and the data can be stored in a safe place. The actual data can be retrieved later when needed.

However, the basic difficulty in designing a de-identification system is the lack of a standard method to measure the effectiveness of the system. What works in one situation and context might not work in another. Even worse, a method that looks good to the eye might be not be good against automatic methods. For example, Newton et al. proposed to defeat de-identification provided by simple ad hoc techniques by applying the same transformation on both the training images as well as the gallery images [29]. They showed that many such techniques may look convincing to human eyes, but in general, they provide little or no protection from face recognition software. Following a similar approach, Gross et al. showed that in some cases the resulting recognition rates in the modified images are even higher than the rates achieved on the original, unaltered images [19].

A robust de-identification system should

- preserve action, remove identity.
- work on full body and not just faces.
- work well against both, human and algorithmic evaluation.
- be robust against subversion attacks.
- be automatic and have a variable control parameter.

We have designed a de-identification system keeping all this in mind. The rest of the chapter outlines other privacy issues related with automated video surveillance systems, to provide a general introduction to the foundations on which our de-identification system is designed.

3.1 Identification and Verification

Verification is an easy problem where a yes-no question needs to be answered. Whether it involves a computer or a human, the problem is always the same. Given an image pair, do the images belong to the same person? A human brain can do it with ease. Algorithmically, it involves a series of transformations on both the images, and comparing the images in the transformed domain. If a one-to-one correspondence can be established between the two images after a finite number of transformations, the image pair is said to be of the same person. Recognition is a more difficult problem as compared to verification, for humans as well as computers. Although, recognition can also be treated as a series of verification problems between all possible image pairs. However, in a huge dataset with thousands of images, a recognition problem might take a lot of time, if treated as a series of verifications. Hence, it

is treated as a different problem altogether where the computer is trained to extract and learn the discernible features of a face. When a computer is presented a photograph to recognize a person, it extracts the recognizable features of the face, compares them with the previously trained examples and returns a match. This method, though less time consuming, is less accurate. The difference between humans and computer at recognition problem is comparable, with humans performing much better than computers. Though algorithmic recognition is less accurate, a determined party can spend a tremendous amount of time on it, given cheap computational power of today. A de-identification system should aim to thwart algorithmic as well as human attacks towards recognition and verification.

3.2 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 de-identification transformation is the most obvious line of attack. The transformation should, thus, be irreversible. Our system is robust towards such kind of attacks. The technical details of such a system are presented in Chapter 4, Section 4.4.

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. We conducted experiments to validate our system's robustness against some common computer vision algorithms viz. face detection and person detection. The results are shown in Chapter 4, Section 4.5.1.

3. Manual identification is another way to subvert de-identification, though it is considerably more expensive. It is not clearly known what properties or features 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. A detailed user study was conducted to evaluate the effectiveness of our algorithm in different classic scenarios. The results are in accordance with our expectations and are presented in Chapter 5.

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 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. The evaluation of the effectiveness of these methods was left out as it is out of the scope of the present work.

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.

3.3 Storage of Videos

The de-identification process should support untransformed video to be viewed if the situation demands. This is essential to support the primary purpose of watching the space, whether for security or information. That is, the de-identification should be selectively reversed when needed. It is important that individuals do not appear in the clear at any time in the video otherwise. The safest approach is to de-identify the video at the capture-camera. Only the transformed video is transmitted or recorded. Clear video can be viewed only by reversing the transformation. This requires the de-identification to be reversible, which poses some risk of being attacked. The parameters needed for reversing the transformation should be saved along with the video using sufficiently strong encryption. Another approach is to store the original video, with sufficiently hard encryption, along with the de-identified video. The required keys for decryption is available only with authorized persons. This needs additional storage space, which can be reduced by saving only the portions that contain humans.

Another relevant issue is the computational costs of de-identification. Videos are bulky and their transformation requires serious computing power. The safest option of applying the transformation at the capture device requires the processing to be performed on an embedded device near the camera. Other option is to transmit the video securely to a server, which de-identifies the video. We do not address the computational issues in this thesis, though they are important.

Chapter 4

Person De-identification in Videos

An overview of our method is outlined in Figure 4.1. The system comprises of three modules: Detect and Track, Segmentation, and De-identification. We now explain each of these modules in detail.

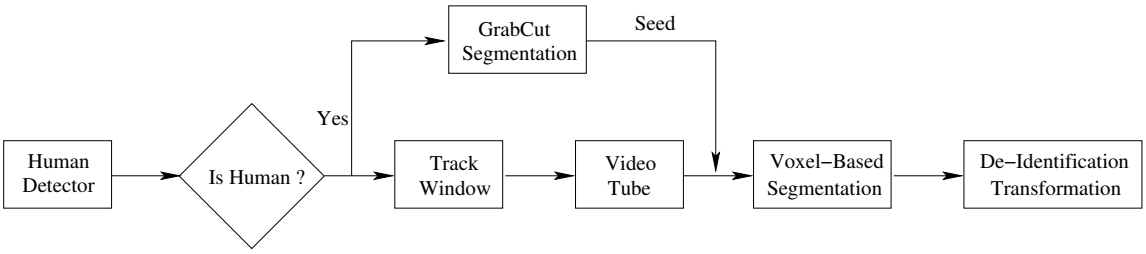


Figure 4.1 Overview of the method.

4.1 Detect and Track

The first step is to detect the presence of a person in the scene. HOG based human detector gives good results with a low miss rate [15]. Other human detectors may also be employed [6, 37]. A robust tracking algorithm is required, as any error in tracking will increase the chances of recognition. We use a patch-based recognition approach for object tracking [1]. The object is divided into multiple spatial patches or fragments, each of which is tracked in the next frame by a voting mechanism based on the histogram of the corresponding image patch. The voting score for different positions and scales from multiple patches is minimized in a robust manner to combine the vote maps and select the most appropriate candidate. This approach is robust to partial occlusions and pose variations. It also takes into account the relative spatial distributions of the pixels, unlike traditional histogram-based tracking methods [14, 44].

Although the algorithm allows for voting on different scales of the object, to avoid errors resulting from partial occlusions and fast changing scale, we apply the human detector every F frames. The

Apply HOG human detector [15].
if Bounding Box (BB) of human overlaps a TrackedWindow **then** {Same Person}
    Replace old TrackedWindow with BB.
else {New Person}
    Add BB as new TrackedWindow.
    Perform GrabCut [33] with BB as input. Build GMMs.
end if
For each new frame, update the existing TrackedWindows after patch-based tracking [1].
Form each person's video tube by stacking their TrackedWindows across time.
Divide the video into fixed $4 \times 4 \times 2$ voxels.
if Number of voxel planes in a person's video tube $= 4$ **then**
    Build a 3-D Graph on voxels of the video tube.
    Perform Graph Cut [7, 22].
    Retain the last voxel plane.
    Apply de-identification transformation on the segmented frames using one of the techniques mentioned in Section 4.3.
    Apply the randomization kernel.
end if

Algorithm 1: Pseudo code: Overview of the method

output of the human detector becomes the input to the tracking module. The value of *F* depends on the amount of movement in the video. If the scale of the human doesn't change much over the course of the video, then a high value of *F* can be chosen. If the scale changes every few frames, then *F* is small. We set the value of *F* to 40 for our experiments.

4.2 Segmentation

The bounding boxes of the human in every frame, provided by the tracking module, 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. This reduces the computation required in the large video space. Also, 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, and can aid recognition [13, 41]. The values of *x* and *y* are typically set to 4 each and *t* can be anything between 2 and 10, depending on the degree of movement in the frames.

Segmentation assigns each voxel ν a label, 1 for foreground and 0 for background. For this, the video tube is divided into blocks of *B* voxel-planes in time. A voxel-plane is a collection of voxels obtained

by combining $[F_n, F_{n+1}, \cdots F_{n+t-1}]$ frames in the video space, where *t* is the size of each voxel in the temporal domain. The voxels are treated as superpixels and a 3D graph is constructed per block, where each node corresponds to a voxel [26]. One voxel-plane overlap is used between consecutive blocks to enforce continuity across the blocks. *B* must be small (between 3 and 10) for good results, but not too small, as it would make the overall computation time high.

The energy term *E* associated with the graph is of the form

$$E(\underline{\alpha}, \underline{\theta}, \underline{\nu}) = U(\underline{\alpha}, \underline{\theta}, \underline{\nu}) + \lambda_1 V_1(\underline{\nu}) + \lambda_2 V_2(\underline{\nu}), \tag{4.1}$$

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. The Gaussian Mixture Models (GMMs) are used for adequately modeling data points in the colour space [12]. $\underline{\theta} = \{\theta^0, \theta^1\}$ 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. These GMMs provide seeds to the graph, as well as help in defining the energy terms. The energy *E* is defined such that a minimization of it provides a segmentation that is coherent across time and space.

The data term *U*, similar to the one used by GrabCut [33], is defined as $U(\underline{\alpha}, \underline{\theta}, \underline{\nu}) = \sum_n D(\alpha_n, \theta_k, v_n)$ where n is the number of voxels and

$$D(\alpha_n, \theta_k, v_n) = \min_{k=1\cdots K} [-\log w_k^{\alpha_n} + \frac{1}{2} \log \det \Sigma_k^{\alpha_n} + \frac{1}{2} \bar{v}_n^T \Sigma_k^{\alpha_n - 1} \bar{v}_n] \tag{4.2}$$

where $\bar{v}_n = v_n - \mu_k^{\alpha_n}$. The representative colour v_n for a voxel should be chosen carefully. The average colour of a voxel is not a good representative as we initialize the GMMs based on pixel colours. The average colour, which is a mixture of several colours, might not lie close to any GMM, despite being a foreground or a background pixel. The problem is intensified in the case of boundary voxels, where the average colour would be a mixture of the foreground and background colours. Our solution is biased towards segmenting more voxels as foreground than background, which would be difficult in case of average colour. To this end, 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 (4.2). 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.

The smoothness terms V_1 and V_2 are also similar to the ones used in GrabCut, defined as: $V(\underline{\nu}) = \sum_{\nu_p, \nu_q \in \underline{\nu}} \delta_{pq} \cdot V_{pq}$, where δ_{pq} is 1 when ν_p and ν_q are neighbours and 0 otherwise, and

$$V_{pq} = \exp^{-\beta \|v_p - v_q\|^2}, \tag{4.3}$$

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 [33].

A mincut on the above graph minimizes the energy *E* efficiently [7, 22]. A rigid but blocky (because of voxelation) outline of the human is obtained after segmentation. Initialization of foreground and background seeds is done by performing GrabCut [33] on the first frame that contains the human. The foreground and background GMMs are also initialized in this process.

4.3 De-identification

After the segmentation of the person, the de-identification transformation is applied on the human being present. We explore two de-identification transformations: $i)$ exponential blur of pixels of the voxel, and $ii)$ line integral convolution (LIC). We explore these transformations in isolation as well as in different combinations, and evaluate the performance of each of these.

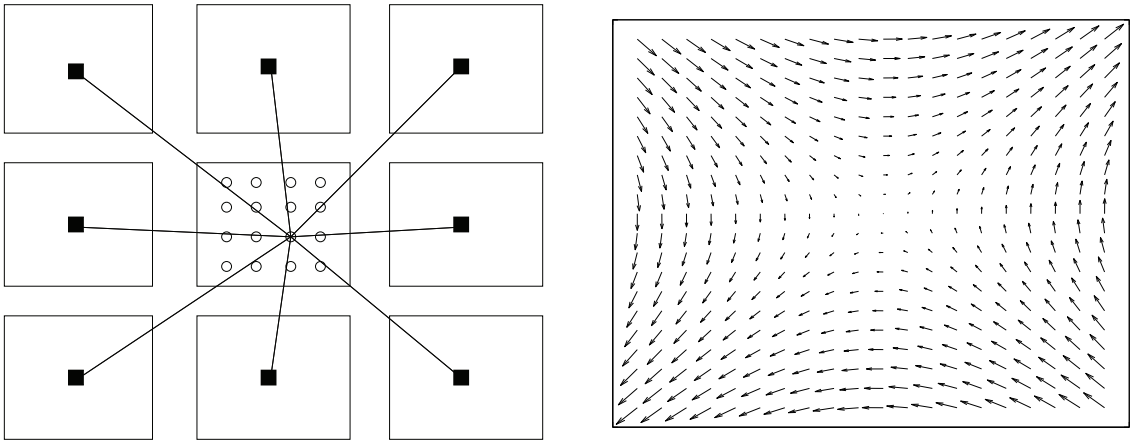
In exponential blur, all neighbouring voxels of a foreground voxel within the distance a participate in de-identification. The parameter a controls the amount of de-identification; more the value of a, more is the de-identification. Typically a lies between 1 and 5. The output colour for each pixel in a foreground voxel is a weighted combination of its neighbouring voxels' average colours. Each voxel is weighted based on the pixel's distance from the center of that voxel. If ν_i is a foreground voxel and ν_p is its neighbouring voxel, the weights corresponding to the $(l, m, n)^{th}$ pixel of ν_i can be calculated $\forall \nu_p \in \Gamma_i$ as:

$$\gamma(l, m, n) = e^{-\frac{d^2_{(l,m,n),\nu_p}}{8a^2}}, \tag{4.4}$$

where Γ_i is the set of voxels which lie within distance a from ν_i, and $d_{(l,m,n),v_p}$ is the distance of the $(l, m, n)^{th}$ pixel of ν_i from the voxel center ν_p.

The weights γ have certain inherent properties. The distance $d_{(l,m,n),v_p}$ depends only on l, m, n and the relative position of v_p with respect to the current voxel (Figure 4.2). Hence, once the value of a is fixed, the weight vector (of size Na^3) is fixed. Because this weight vector is same for every voxel, it can be pre-computed once and used for every voxel. Moreover, the distance $d_{(l,m,n),v_p}$, and hence the weight vector $\gamma(l, m, n)$, vary smoothly within a voxel and across two voxels. This prevents abrupt changes in colour at voxel boundaries. Also, because the weight corresponding to a distant voxel is low compared to a nearby voxel, the voxels at distance a will have less contribution to a pixel's colour. Hence, the colour of pixels on the either side of voxel boundaries changes smoothly, as only voxels at distance a are added or removed from their active neighbourhood. This kind of smooth temporal blurring of the space-time boundaries aims to remove any gait information of the individual.

The second de-identification transformation is based on line integral convolution (LIC). LIC is used for imaging vector fields [9] on a texture. A long and narrow filter kernel is generated for each vector in the field whose direction is tangential to that of the vector and length is $2L$. L lies typically between 2 and 20. The bounding box around the human is mapped one-to-one onto the vector field. The pixels



(a) (b)

Figure 4.2 (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.

within the bounding box and under the filter kernel are summed, normalized and placed in an output pixel image for the corresponding position. This process is repeated for all foreground pixels obtained after segmentation. LIC distorts the boundaries of the person which tends to obfuscate silhouettes. Different vector fields can be used for achieving different effects. We used a saddle shaped vector field (Figure 4.2) for our experiments. The amount of de-identification is controlled by the line length parameter, L, of the convolution filter.

When used in isolation, blur is more effective to hide gait and facial features, while LIC distorts the silhouettes more. Hence, we tried a combination of these two transformations where we perform LIC on the voxels followed by a voxel based exponential blur. To make identification based on the colour of face and clothes difficult, intensity space compression (ISC) was additionally tried as a subsequent step. 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 context information. The results are presented in Figures 4.3 and 4.4.

4.4 Randomization

Reversing the de-identification transformation is the most obvious line of attack. The transformation should, thus, be irreversible. We use a blurring involving several neighbouring voxels in space and time to prevent direct reversal. However, an indirect reversal approach is to estimate the blurring function from the de-identified frames and then get the original frames back using reconstruction and comparison. Frames of the de-identified video may also 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. However, these techniques assume that the point spread function (PSF) of the camera (or the blurring function) which results in the low resolution image is the same at every pixel of the image. The most intuitive method to prevent this kind of attack is to make the blurring function random which would make the estimation impossible. The key is to randomize the function in such a way that does not adversely affect the image quality and smoothness.

The easiest way to thwart a reversal attack using them is to randomize the blurring function at every pixel. This trivial adjustment makes estimation of the blurring function impossible, and hence direct comparison based reconstruction techniques will not work. Instead of making the whole blurring function random at every pixel which would result in non-smooth, low quality and blocky images, we make use of a separate randomization layer as the final step. This is achieved by using a blurring kernel (one out of a fixed pool of N kernels), chosen randomly for every pixel. The pool contains low pass filters of frequencies and construction slightly different from each other. This blurring is thus sufficiently random, but not so much to introduce sharp lines in the output image. Similar effect could be achieved by adding a small random value to the blurring weight corresponding to each pixel in the previous step. However, the resulting kernel will not be consistent with the notion of an ideal blurring kernel where the weights fall off consistently with respect to distance, and might introduce discontinuities around the boundaries of two voxels.

4.5 Experimental Results

We implemented the above system and conducted the experiments on standard data sets like CAVIAR, BEHAVE, etc., and on our own that provide more clearly visible individuals in videos. We divide the video into $N = 4 \times 4 \times 2$ sized voxels. The parameter m, which decides the representative colour v_n of a voxel used in defining the data term in Equation (4.2), 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. Different parameters were tried for each of the de-identification transformations; $a = 2$ and 4 for exponential blur, $L = 10$ and 20 for LIC on pixels, and $vL = 2$ and 5 for LIC on voxels. $L = 20$ in pixel space is equivalent to $vL = 5$ in voxel space as 5 voxels cover 20 pixels in one dimension. Similar comparisons can be made between $L = 10$ and $vL = 2$.

Our implementation is not real time currently. It takes about 10 to 12 seconds on an average to completely process and obtain results on a block of size 4 voxel planes on an Intel 2.4 GHz processor with 1 GB RAM. The tracking module takes about 8 to 10% of the running time. Graph cut in itself takes only about 2 to 3% of the total time to run. The de-identification and randomization modules together take over 12% of the time. The rest of the time is spent in voxelizing the video, calculating the energy functions for t-edges and n-edges of the graph, etc. The inherent parallelism of many of these modules may be explored for a real-time implementation on the GPU. However, we do not address the real-time issue in this work.

Visual results in selected frames are shown in Figures 4.3, 4.4 and 4.5. Figure 4.3 shows the output of different de-identification transformations on a single frame from different videos. 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, Blur-4 and LIC-20 perform better than Blur-2 and LIC-10 in masking the identity of people. However the output of LIC-20 sometimes looks unnatural and ghost-like. The combination of LIC and Blur works better than either by itself; the user-study conducted on the videos conforms with the statement and is discussed in the next section. The effect of changing the parameters of the transformations can be seen in the figures. The intensity space compression (ISC), as shown in Figures 4.3 and 4.4, can remove colour dominated information such as race, but can accentuate the body structure of the person. Figure 4.5 shows frames of de-identified videos in which people are performing different activities. As can be seen, the activity is recognizable but the person is not, which is the underlying goal of de-identification. More results can be seen in the video available on *http://cvit.iiit.ac.in/projects/de-id/index.html*.

4.5.1 Algorithmic Evaluation

To gauge the robustness of our system against algorithmic recognition techniques, we tested the de-identified videos on a standard face detector and a human detector, which are used as the first step by most recognition algorithms. Since recognition requires more intricate feature information than detection, we show failure of state-of-the-art detection algorithms as proof of robustness of our algorithm against computer vision algorithms. We used OpenCV's implementation of the Viola-Jones face detection algorithm [38] for face detection and the HOG based human detector for person detection [15]. On a total of 24 de-identified videos, and 6110 frames in which a person was present, the face detector resulted in 0.2% hits and the human detector resulted in 56.2% hits, on an average. Table 4.1 summarizes the output for different transformation combinations. An increase in the de-identification transformation parameter reduces the number of hits, as expected. However, when the detectors were tested on clear

	Percentage of Success	
Algorithm, Parameter	*Human Detection*	*Face Detection*
Blur, $a = 2$	89.7	1.0
Blur, $a = 4$	56.1	0
LIC, $L = 10$	73.6	0.3
LIC, $L = 20$	22.4	0
$vL = 2, a = 2$	64.4	0
$vL = 2, a = 4$	59.3	0
$vL = 5, a = 2$	44.9	0
$vL = 5, a = 4$	38.9	0

Table 4.1 Percentage of correct answers for the face and human detectors.

videos, we get 97.2% and 7.8% hit rates in the case of person detector and face detector respectively[1]. A fall in the hit rate in de-identified videos, especially of the face detector, can be taken as a confirmation that our system is robust against recognition algorithms, as the fine details which are the requirement of any recognition algorithm are removed from the videos. The person detector worked in more than half the cases on an average, which is acceptable, as it only only indicates that a human is present in the video.

[1]The videos contained people at a large distance from the camera, as in surveillance, and frontal and profile faces. All these explain the low hit rate in the case of face detector.

Figure 4.3 The first row shows the clear frames. The next five rows show the output of Blur-2, Blur-4, LIC-10, LIC-20, and Blur-2 followed by an intensity space compression, in that order.

(a) Clear frames (b) Result of Blur-4 (c) Clear frames (d) Result of Blur-4 followed by intensity space compression

Figure 4.4 Results on two different videos. The clear frames are shown in the odd columns while corresponding de-identified frames in the even columns.

(a) Talks on the phone (LIC-10)

(b) Waves at the camera (Blur-2)

(c) Kicks in the air (vLIC-2, Blur-2)

Figure 4.5 De-identified frames showing people performing different activities; the activity is recognizable but the person is not.

Chapter 5

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, two user studies were conducted to test the usefulness of the system. An effective de-identification algorithm should preserve the action in a video, while removing the identity of the actor completely. The user studies were aimed to bring out the trade-off between privacy and context in a video de-identified by different parameters. Another purpose of these studies was to weigh the importance of gait as a feature for recognition.

The videos used in these studies were our own, taken in different plausible settings, featuring students of our institute. Preliminary results of an early user study on standard data sets like CAVIAR, BEHAVE, etc., revealed that these data sets are not challenging enough for studies on de-identification. In BEHAVE, the scene is captured by a distant camera placed high up looking down the road. The actors' faces are very small and not recognizable. In CAVIAR, the faces are recognizable when the actors came close to the camera. However, for unfamiliar people, the only cue available to identify these actors (even in most clear videos) is the colour of their clothes. Since we did not have access to an image of these actors other than that in the video itself, the user study on these de-identified videos necessarily meant matching the colour of clothes in the candidate images and videos. Clearly that does not provide a correct assessment of the technique employed. To prove that the users were indeed using the clothes' colour information and nothing else to match the actors, we identified some cases where the actors appeared in two different videos in CAVIAR wearing different clothes. For such actors, we showed the users the candidate image of the actor in one video and de-identified video of the other. In almost all the cases, the users couldn't identify the actor correctly, while their accuracy was very high for the cases when the template image was taken from the same video. Hence, there was a need to create our own data set for the evaluation of our method, in which faces are clearly visible for even unfamiliar people to recognize. We could also take different images and videos of our actors in different clothing and scenarios to conduct a detailed user study as explained below.

Figure 5.1 Screenshot of the portal used for the user study for Identification

The individuals in our data set were asked to perform actions like waving hand (as a gesture to greet), talking on the phone, turning head left or right, carrying a bag, eating or drinking, etc. The user study gauged identification of the person and the action performed. Clear videos were also used as examples to enable the learning of gait, silhouette, and other aspects. A demo of our system was put up for evaluation at a technical event at our institute, which was attended by several hundred visitors from outside. The study was conducted on 74 such visitors. The subjects were completely unfamiliar with the individuals appearing in the videos. As shown in Table 5.1, 8 different parameter combinations of the de-identification transformations were included in the study. The study consisted of 8 sets of 6 videos each. Half the videos in each set was de-identified using one combination and was used for the *identification* experiment. The other half was de-identified using another combination and was used for the *search* experiment. In this manner, all the 8 parameter combinations were covered for both the experiments (identification and search) in these 8 sets. The people taking the study were also divided into 8 sets (named A to H), and each user took the study on identification and search in one set. The

Figure 5.2 Screenshot of the portal used for the user study for Search

users were shown a randomly chosen video for identification and another for search. This was to ensure that the outcome of the experiment is not affected by the type of videos used for the purpose.

For the identification experiment, the users were asked to match the individual in the de-identified video against a pool of 20 candidate photographs (face and upper body only) shown in clear (Figure 5). They were also asked to select the action in the video. Next, the users were shown clear videos of those 20 candidates from which they could learn their walking style, posture, etc. These videos were taken in a different setting and in different clothes than the de-identified videos to ensure there is no unnecessary learning from the background, context, etc. The users could go back and change their previous answer. Similarly, for the search experiment, the users were asked to search for an individual in a pool of 8 de-identified videos (Figure 5). They were first shown a clear image of the person and were asked to

find him/her. Then they were shown a clear video of the same person, and were given an option to go back and change their answer. All their answers were recorded and are summarized in Table 5.1.

The numbers in Table 5.1 represent the correct identifications and searches by the users. The alphabet in the parentheses represents the set of people who took that particular experiment. The first column of the table represents the different algorithms and their parameters used. The next two columns are for different tests (identification and search from images, and then from videos). The last column shows the number of times the activity was correctly recognized by the users for a particular parameter.

The study can be divided into three categories for the sake of analysis. One category deals with the effect of a certain de-identification algorithm and parameter on the recognition ability of the users. Another category compares the improvement in performance of the users due to learning the gait and silhouette, in identification and search. The third category analyzes the ability of the users to recognize the activity for different parameters. The user study results are mostly as expected. Individuals with very special walking styles or body structures had much better recognition. The users could recognize the activity in the de-identified video in most cases for all parameters, at an average of about 80%. The impact of parameters is also as expected. The trade-off between privacy and context in the de-identified videos is apparent from the results. As the parameter controlling the amount of de-identification increases, the percentage of correct answers decreases. This necessarily means that as the actors became less identifiable, the video started losing the context and detail, as expected. This is almost always true, except in few cases, as explained later.

There are a few observations to be made from the study.

1. In general, search is easier than identification, as it is easier to learn about one person in search than about all possible candidates in identification. Hence, very few people changed their answers in the case of identification when they were shown clear videos after images to learn the gait. It also makes sense intuitively as verification is easier than identification.

2. A combination of LIC and Blur is better than these transformations in isolation. While this is true in most cases, more users changed their answers (usually to correct ones) when they were

	From Images		*From Videos*		*Activity*
Algorithm, Parameter	*Identification*	*Search*	*Identification*	*Search*	*Recognition*
Blur, $a = 2$	4(B)	4(A)	4(B)	4(A)	7(B)
Blur, $a = 4$	1(D)	3(C)	0(D)	1(C)	6(D)
LIC, $L = 10$	3(F)	2(E)	4(F)	2(E)	7(F)
LIC, $L = 20$	1(H)	3(G)	1(H)	3(G)	7(H)
$vL = 2, a = 2$	3(C)	3(D)	2(C)	3(D)	11(C)
$vL = 2, a = 4$	0(G)	1(H)	1(G)	4(H)	8(G)
$vL = 5, a = 2$	2(A)	0(B)	2(A)	2(B)	8(A)
$vL = 5, a = 4$	2(E)	0(F)	3(E)	4(F)	7(E)

Table 5.1 Number of correct identifications for *search* and *identification* experiments in the user study. Sets A to H had 9, 8, 11, 9, 9, 8, 10, and 10 users respectively.

shown clear videos for the combinatorial cases. While this might look like an anomaly, it could be because the faces were obscured totally by the combined transformations. Hence, there was an increased reliance on the clear videos for learning the gait of these individuals, which is more pronounced in the case of search than identification for reasons explained earlier.

3. The users fared better in identification from videos than images for all de-identification transformation combinations. The users spent only about 4 to 5 minutes on an average to complete the entire study and may have had only limited have enough time to learn the gait, etc., from the videos. The users did not change their answers when they moved from images to videos in most cases. If they did, they changed their answer to the correct one. However, in some cases, like identification and search in Blur-4 and identification in $vL = 2, a = 2$, some users who gave correct answers from the images changed their answers when they were shown videos. While part of this anomaly could be attributed to the anxiety of people when they are a part of such user studies, most of it stemmed largely from the fact that the videos which were used for the user study contained two men and two women whose height, build and walking styles were similar. Moreover, Blur-4 hides facial features more than any other transformation, and there was more reliance on videos for recognition. Two cases out of three in which the anomaly occurred are from the same set, which means that that particular set of users were more anxious and confused than others and changed their answers to the wrong ones after seeing the videos.

4. As the parameter controlling the amount of de-identification increases, the percentage of correct answers decreases. However, across different algorithms, LIC–10 is more effective than Blur-2. $vL = 2, a = 4$ is similar to $vL = 5, a = 2$. While users perform better in identification on one case, they perform better on another in search. $vL = 5, a = 2$ and $vL = 5, a = 4$ are also similar in performance, with only major difference being in search from videos. A possible explanation for this anomaly is that the set of users who took this particular experiment (F) were good at recognition, as is also apparent from the high numbers corresponding to the other experiment conducted with the same set, LIC-10. Another anomaly occurs between $vL = 2, a = 4$ and $vL = 5, a = 4$. In the case of $vL = 5, a = 4$ under identification, the percentage of correct answers is more than the corresponding figures in $vL = 2, a = 4$. The anomalies can be attributed to different set sizes, difference in the difficulty of de-identified videos across sets, and randomness which is unavoidable in any user study.

5. Gender is very hard to hide. Most users could identify the gender correctly, except a few times. The few anomalies only occurred in cases where high values of parameters for LIC were used. Note that our study was not attempting to hide the gender.

To test the effect of familiarity on recognition ability, another user study was conducted. 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 from the same lab as the individuals appearing in the videos and were quite familiar with them. Others were from different labs and were only casually familiar with these individuals.

	Familiar		*Casually familiar*	
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 5.2 Number of correct identifications in the user study on familiar people.

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 5.2. The numbers in the table represent the correct identifications by the users. Overall correct recognition was fairly high due to the familiarity of the users with the subjects. The users rated the gait or the walking style to be a big give-away. For example, individual 4, for whom the highest recognition was reported (about 80%), has a very unique walking style. For individual 2, only about 20% of the answers were correct because this person has no unique body shape or walking style. The correct answers for this person were only from those sets in which low values of parameters for Blur and LIC were used.

Another user study was conducted on 30 people to capture the users' experience of the processed videos. We showed each user 9 videos, each processed with a different parameter combination (including ISC). The users were asked to rate each video on a scale of $1 - 7$ to specify how natural (or acceptable) they found a particular parameter, where a score of 1 meant very unnatural and unacceptable while 7 meant completely acceptable. The results are shown in Table 5.3. All the parameter combinations scored above 3 on an average, while LIC-20 scored 2.6 and ISC scored only 2.2. Blur scored about 4.5 on an average (with answers ranging from $3 - 7$), which is slightly better than LIC which scored about 3.5 on an average (with answers ranging from $2 - 6$). The average scores of LIC and Blur combinations were between $3 - 6$, with scores decreasing as the parameter values were increased.

Algorithm, Parameter	*Naturalness*
Blur, $a = 2$	5.2
Blur, $a = 4$	3.5
LIC, $L = 10$	3.9
LIC, $L = 20$	2.6
$vL = 2, a = 2$	5.2
$vL = 2, a = 4$	3.9
$vL = 5, a = 2$	3.8
$vL = 5, a = 4$	3.0
ISC	2.2

Table 5.3 Human experience scores on a scale of 1 (low) to 7 (high).

The difference in the naturalness scores of LIC and Blur was not significant enough to affect the choice between these two algorithms.

5.1 Limitations

Our algorithm consists of many modules and the results are sensitive to proper functioning of all the modules involved. It is necessary for each module to function perfectly for our de-identification to work. Failure to do so in even one frame can jeopardize privacy. Each module has its own limitations. The tracking module misses the extended arms, feet, hair, or even the face sometimes which might compromise privacy. The segmentation module is largely dependent on colour, and gives errors when the background is cluttered or the background and foreground have the same colour around the segmentation boundary (Figure 5.3). The seeds for segmentation and the GMMs depend on the success of GrabCut, which is a very crucial step. Also, a miss by the HOG detector will certainly prove fatal for the de-identification process.

(a) (b)

(c) (d)

Figure 5.3 Segmentation result on a video with dynamic background.

5.2 Discussion

The results of the user study conform to our expectations. The expected trade-off between privacy and context in the de-identified videos is apparent from the results. As the parameter controlling the amount of de-identification increases, the percentage of correct answers decreases usually. That is, as the actors became less identifiable, the video started losing the context and detail. The results also confirm that verification or searching a person in de-identified videos is easier than identification of a person in a de-identified video. The 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. Hence, people familiar with the subjects in the videos can identify them in the de-identified videos with more accuracy than unfamiliar people.

An interesting observation to be made from the user study is that most common people use facial features primarily while identifying a person. But only when these features are obscured completely, there is increased reliance on other features such as body shape, gait, etc., for identification. Another important observation is that gender is very hard to hide, although we were not attempting to hide the gender in this particular work. The transformations and combinations that worked well against manual identification could not fare well against gender recognition.

Our user study confirms 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. The studies also suggest that an action can be recognized with more accuracy than the person performing that action in a de-identified video, for all the combinations and parameters of a de-identification transformation.

Chapter 6

Conclusions and Future Work

Video surveillance has become a common feature of our everyday life these days. While video surveillance is an important tool to prevent, detect and reduce crime and disorder, it is widely used to monitor people's conduct, detect their presence in an area and/or possibly study their actions too. The public expects it to be used responsibly complying with the data protection norms. The access to these videos is expected to be limited to authorized persons only. The usage of these videos is assumed to be fair and for everyone's advantage. However, de-identification is strongly recommended to take care of potential viewing by non-authorized people. Since de-identification only aims to remove the identity of the people in a video without compromising on the action or context, in cases where it suffices to see the action being performed in the video, an irrevocable de-identification is recommended. For other purposes, a revocable de-identification with a strong private key for decryption should suffice.

Apart from video surveillance where people are usually aware that they are "being watched", technologies like Google Street View, EveryScape, Mapjack, etc., also pose an unintentional threat to an ordinary individual's privacy. Individuals appear in these videos purely unintentionally and there is no need to know their identities. All individuals should therefore be de-identified irrevocably.

We outlined the different scenarios where de-identification is necessary and the issues brought out by those. First of all, it is important to assess the purpose of video surveillance and the expected benefits before it is installed in an organization. For instance, a surveillance system which relies on the streaming video for taking an immediate action can be encrypted using a visual domain de-identification technique. However, if it suffices to store the video, and later view the contents in case a mishap has occurred, a data-encryption technique can be employed at the camera itself and the data can be stored in a safe place. The actual data can be retrieved later when needed. In surveillance scenarios, it is best to not leave the privacy protection technologies to be employed (or requested) by the consumers at their will, but incorporate them into the video surveillance systems from the beginning based on their implied use. It is better to err on the side of caution than to compromise the privacy of people.

Second, it is important to ensure that clear videos are not accessible to anyone other than properly authorized personnel. Recorded videos should be stored in a way that maintains the integrity of the

videos. This is to ensure that the rights of the subjects are protected and that the untransformed videos can be viewed if the situation demands. To avoid the risk of clear videos being viewed by a third party, the necessary de-identification must be performed at the camera itself. Only the de-identified video must be transmitted or recorded. The de-identification could be reversible and the parameters needed for reversing the transformation should be saved along with the video using sufficiently strong encryption. However, a reversible transformation poses the risk of being attacked. We recommend an irreversible de-identification transformation to be applied to the videos at the camera itself. There is a need to store the original video, with sufficiently hard encryption, along with the de-identified video. The required keys for decryption is available only with authorized persons. This needs additional storage space, which can be reduced by saving only the portions that contain humans.

An ideal de-identification system should aim to obfuscate all identifiable features in a video. Face plays the most important role in both, automatic and manual identification. A de-identification algorithm must obscure faces beyond recognition. However, other features like gait, silhouette, body posture, etc., are unique to a person and can help in recognition. Humans exploit this information effectively and algorithmic identification schemes using body silhouette and gait have been developed with some success in the past. Hence face de-identification is not enough when it comes to privacy protection of people. Other identifiable features like gait, silhouette, etc., must also be obscured.

For any de-identification method to work, it is necessary to maintain the usability of the videos. Excess obscuration makes video surveillance meaningless. Context preservation is important for the usability of the videos captured. The de-identification transformation should not make the video meaningless by removing all context from it, or render it unviewable by distorting the naturalness. Preliminary operations like motion detection, object tracking, action analysis, etc., should be possible in a de-identified video. A person watching the live streaming of such a video should be able to analyze the subject's behaviour. Actions such as the subject entering a certain area of a camera's field of view, or picking up an object, or abandoning the luggage, or even acts of aggression should be identifiable. An unusual behaviour may be indicative of a security threat and an alarm could be raised well in time to avoid a mishap.

While designing a de-identification system, it should be kept in mind that a method that looks convincing to the human eye might be not be good against automatic methods. A robust de-identification transformation should work well against both, automatic and manual evaluation. It should also be robust against subversion attacks. Above all, for a de-identification system to be suitable for a surveillance scenario, it should be automatic. Additionally, it can have a variable control parameter, which could either be controlled by a manual feedback system, or an automatic action-context recognition system.

In this thesis, 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 sufficient familiarity with the subjects and the user. Blurring is a good way to hide the identity if gait is not involved. The trade-off between privacy and context in the de-identified videos is evident from the user studies. The results showed that as the parameter controlling the amount of de-identification increased, the actors became less identifiable (more privacy), while the video started losing the context and detail. The studies also suggest that an action can be recognized with more accuracy than the actor in a de-identified video, which is the guiding principle of de-identification.

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. A de-identification system would benefit from a framework that incorporates a person's personal sense of privacy into the system. However, usage of such a system would be limited to closed group (society) surveillance scenarios only. An ideal de-identification transformation should work in real-time, either on a CPU or a GPU. The inherent parallelism of many of the modules of our system may be explored for a real-time implementation on the GPU. However, we do not address the real-time issue in this work. Also, a context recognition system can ensure the usability of the videos stored. The feedback from such a system can control the amount of privacy provided based on the action and context before storing the videos. The possibility of a feedback system that is based on the balance between the privacy and information content in a de-identified video is another thing that a de-identification system would benefit from and needs to be explored in the future.

Related Publications

1. **Journal:** Prachi Agrawal and P. J. Narayanan, "Person De-identification in Videos", *The IEEE Transactions on Circuits and Systems for Video Technology* (**TCSVT**), 2010.

2. **Conference:** Prachi Agrawal and P. J. Narayanan, "Person De-identification in Videos", *The Ninth Asian Conference on Computer Vision* (**ACCV 2009**), Xi'an, China, Sep. 2009.

Bibliography

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR (1)*, pages 798–805, 2006.

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. M. Drucker, A. Colburn, B. Curless, D. Salesin, and M. F. Cohen. Interactive digital photomontage. *ACM Trans. Graph.*, 23(3):294–302, 2004.

[3] S. Avidan and M. Butman. Blind vision. In A. Leonardis, H. Bischof, and A. Pinz, editors, *ECCV (3)*, volume 3953 of *Lecture Notes in Computer Science*, pages 1–13. Springer, 2006.

[4] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):1–8, 2008.

[5] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. *Comput. Graph. Forum*, 23(3):669–676, 2004.

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[7] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, pages 105–112, 2001.

[8] M. Boyle, C. Edwards, and S. Greenberg. The effects of filtered video on awareness and privacy. In *CSCW*, pages 1–10, 2000.

[9] B. Cabral and L. C. Leedom. Imaging vector fields using line integral convolution. In *SIGGRAPH '93*, pages 263–270, 1993.

[10] D. Chen, Y. Chang, R. Yan, and J. Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007(1):107–107, 2007.

[11] S. C. S. Cheung, J. K. Paruchuri, and T. P. Nguyen. Managing privacy data in pervasive camera networks. In *ICIP*, pages 1676–1679, 2008.

[12] Y.-Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *CVPR (2)*, pages 264–271, 2001.

[13] R. T. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *Proceedings of IEEE Conference on Face and Gesture Recognition*, pages 351–356, 2002.

[14] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, pages 2142–2149, 2000.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[16] F. Dufaux and T. Ebrahimi. Scrambling for privacy protection in video surveillance systems. *IEEE Trans. Circuits and Systems for Video Technology*, 18(8):1168–1174, 2008.

[17] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *ICCV*, 2009.

[18] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies*, pages 227–242, 2005.

[19] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *CVPRW '06*, page 161, 2006.

[20] R. Gross, L. Sweeney, F. D. la Torre, and S. Baker. Semi-supervised learning of multi-factor models for face de-identification. In *CVPR*. IEEE Computer Society, 2008.

[21] N. Hu, S.-C. S. Cheung, and T. Nguyen. Secure image filtering. In *ICIP*, pages 1553–1556. IEEE, 2006.

[22] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.

[23] T. Koshimizu, T. Toriyama, and N. Babaguchi. Factors on the sense of privacy in video surveillance. In *CARPE '06*, pages 35–44, 2006.

[24] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[25] G. Li, Y. Ito, X. Yu, N. Nitta, and N. Babaguchi. Recoverable privacy protection for video content distribution. *EURASIP Journal on Information Security*, vol. 2009, Article ID 293031, 2009.

[26] Y. Li, J. Sun, and H.-Y. Shum. Video object cut and paste. *ACM Trans. Graph.*, 24(3):595–600, 2005.

[27] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *PAMI*, 28(7):1052–1062, 2006.

[28] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, 2006.

[29] E. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying facial images. *IEEE Transactions on Knowledge and Data Engineering*, 17:232–243, 2003.

[30] S. Park and M. Trivedi. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. In *AVSBS05*, pages 171–176, 2005.

[31] P. J. Phillips. Privacy operating characteristic for privacy protection in surveillance applications. In T. Kanade, A. K. Jain, and N. K. Ratha, editors, *AVBPA*, volume 3546 of *Lecture Notes in Computer Science*, pages 869–878. Springer, 2005.

[32] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, pages 824–831, 2005.

[33] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[34] J. Schiff, M. Meingast, D. K. Mulligan, S. Sastry, and K. Y. Goldberg. Respectful cameras: detecting visual markers in real-time to address privacy concerns. In *IROS*, pages 971–978. IEEE, 2007.

[35] A. W. Senior. Privacy enablement in a surveillance system. In *ICIP*, pages 1680–1683, 2008.

[36] A. W. Senior, S. Pankanti, A. Hampapur, L. M. G. Brown, Y. li Tian, A. Ekin, J. H. Connell, C.-F. Shu, and M. Lu. Enabling video privacy through computer vision. *IEEE Security & Privacy*, 3(3):50–57, 2005.

[37] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Unified crowd segmentation. In *ECCV (4)*, pages 691–704, 2008.

[38] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001.

[39] H. Wactlar, S. Stevens, and T. Ng. Enabling personal privacy protection preferences in collaborative video observation. In *NSF Award Abstract 0534625*.

[40] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The active badge location system. *ACM Trans. Inf. Syst.*, pages 91–102, 1992.

[41] J.-H. Yoo, D. Hwang, and M. S. Nixon. Gender classification in human gait using support vector machine. In *ACIVS*, pages 138–145, 2005.

[42] X. Yu, K. Chinomi, T. Koshimizu, N. Nitta, Y. Ito, and N. Babaguchi. Privacy protecting visual processing for secure video surveillance. In *ICIP*, pages 1672–1675, 2008.

[43] W. Zhang, S. C. S. Cheung, and M. Chen. Hiding privacy information in video surveillance system. In *ICIP (3)*, pages 868–871, 2005.

[44] Z. Zivkovic and B. J. A. Kröse. An em-like algorithm for color-histogram-based object tracking. In *CVPR (1)*, pages 798–803, 2004.