FOREGROUND AND SCENE STRUCTURE PRESERVED VISUAL PRIVACY PROTECTION USING DEPTH INFORMATION

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE in Computer Science

---

by
Semir Elezovikj
May 2014

Thesis Approvals:

Dr. Haibin Ling, Thesis Advisor, Department of Computer and Information Sciences

ABSTRACT

We propose the use of depth-information to protect privacy in person-aware visual systems while preserving important foreground subjects and scene structures. We aim to preserve the identity of foreground subjects while hiding superfluous details in the background that may contain sensitive information. We achieve this goal by using depth information and relevant human detection mechanisms provided by the Kinect sensor. In particular, for an input color and depth image pair, we first create a sensitivity map which favors background regions (where privacy should be preserved) and low depth-gradient pixels (which often relates a lot to scene structure but little to identity). We then combine this per-pixel sensitivity map with an inhomogeneous image obscuration process for privacy protection. We tested the proposed method using data involving different scenarios including various illumination conditions, various number of subjects, different context, etc. The experiments demonstrate the quality of preserving the identity of humans and edges obtained from the depth information while obscuring privacy intrusive information in the background.

Dedicated to Ika Dalip, my grandmother

"And if I come back one day

Take me as a veil to your eyelashes

Cover my bones with the grass

Blessed by your footsteps"

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my thesis advisor, my mentor, Professor Haibin Ling for his constant and untiring support and guidance. He is the primary source for getting my questions answered and I thank him for all the insightful discussions and advice.

I would like to thank my wife Ferria for her unconditional and unwavering patience and support. Her love and understanding were really encouraging and instrumental in this experience. Thank you for gracing my life with your presence.

Finally, I would like to thank my parents, Muhamed and Sabrije, for their hard-work and sacrifice to help me realize my own potential. You two have been pillars of support and wisdom throughout my whole life. Your vision, strength and belief were always there to raise me when I was weary.

TABLE OF CONTENTS

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Privacy protection in person-aware visual systems is an ill-posed problem: the attitude towards privacy is different with different people. With the increasing technological capability of cameras in recent years, video has become commonplace. The quality and resolution of images has improved to the point where it is possible to zoom in enough to get details on potentially private information. A camera with a mere 60-times optical zoom lens, can read the wording on a cigarette packet at 100 yards [13]**.** This has led to glaring privacy-related concerns that have been expressed over the development and deployment of these technologies. Pervasive video surveillance systems inherently eliminate video anonymity and allow for jeopardizing identities and activities of people [8]. For example, the Google Street View service can potentially disclose much more information than it intends to: license plates, house numbers, information of potentially confidential information or people engaging in private, confidential activities. Even though those images are taken from public property, the technological advancements of recent cameras with high resolution and magnification capabilities contribute to capturing privacy intrusive data from private properties.

Google's video service YouTube has become a popular destination for videos of protest and civil disobedience as Google has made an active step towards "helping activists sidestep autocratic regimes"[12]. To address the issues of privacy invasion and disclosure of sensitive data**,** Google provided the "Blur All Faces" option, which enables users to automatically obscure the identity of all people in the video and they also have the option to permanently delete the original, unblurred version of the video from

Google's servers. This helps human rights activists and campaigners to appear anonymously and be protected against incrimination from authoritarian regimes if they were to be exposed on YouTube videos.

Due to its importance, privacy protection in visual data has been attracting a large amount of research, especially from the multimedia and computer vision communities. A lot of systems and algorithms have been developed in the context of visual surveillance [11], such as [4, 3, 10, 6] etc. Another group of studies focus on designing privacy obscuring algorithms that at the same time preserve data utility as much as possible, such as de-identification for human body [1], car license plate [5], faces [7, 2], etc.

In this study we focus on achieving privacy protection in person-aware visual systems. The specific task can vary depending on what information from the original image content we want to privatize. We can obscure the background of a scene and maintain the identity of a person in order to protect the privacy of the surroundings. An example of this is a video conference where the video participants can see plenty of the surroundings with the large field of view offered by modern video conferencing cameras. Privacy in this scenario would mean obfuscating the background. The reason we would want to do this is to protect potentially sensitive information or simply not wanting to disclose the environment. Another type of privacy would be the privacy where the identity of the people is protected and the background, non-human region can be left unobfuscated. The framework proposed in this study is flexible enough to accommodate for multiple privacy scenarios. Chapter 2 covers a comprehensive listing of potential privacy scenarios that we can explore.

In particular, we aim to privatize potential privacy-intrusive information in the surroundings of the foreground subject of interest. Intuitively, we need to distinguish two kinds of information from private ones:

1. Humans in the foreground region who need to remain clear in the visual communication tasks (e.g., video conferences)
2. Geometric structure information (e.g. depth edges) that relates a lot to scene understanding but little to identity.

Towards the above goal, we make two main contributions in this study:

- *Sensitive information inference via depth information.*

  We propose to use depth information to estimate a sensitivity map to measure per-pixel privacy sensitivity. With the help of modern depth sensor, we can identify the image regions for foreground subjects. In addition, we can also derive the depth edges from the gradient magnitude of the depth map. These two strategies are integrated into the estimation of the sensitivity map.

- *Preservative privacy protection using inhomogeneous obscuration*.

  We model the privatization with an inhomogeneous diffusion process. Specifically, we adopt the Perona-Malik diffusion [9] framework and associate the sensitivity map with the rate of diffusion factor. Intuitively, the more sensitive a pixel is, the larger diffusion kernel is applied to obscure it.

The flowchart of the proposed method is summarized in Figure 1. The foreground map $R(x)$ is constructed from the player segmentation data provided by the Kinect's

Depth Stream . The foreground map is a bitmap where each pixel signifies the index of the person in the field-of-view (up to 6 human figures can be identified). The depth map $D(x)$ is a map of distances to the nearest object (in millimeters) for each pixel. The depth image in our study has a resolution of 640x480 pixels. Using the depth map $D(x)$ and the foreground map $R(x)$ we create a sensitivity map $P(x)$ which denotes the level of privacy/sensitivity for each pixel. The $P(x)$ defines how much obfuscation should be applied to the original image $I(x)$ to each pixel.

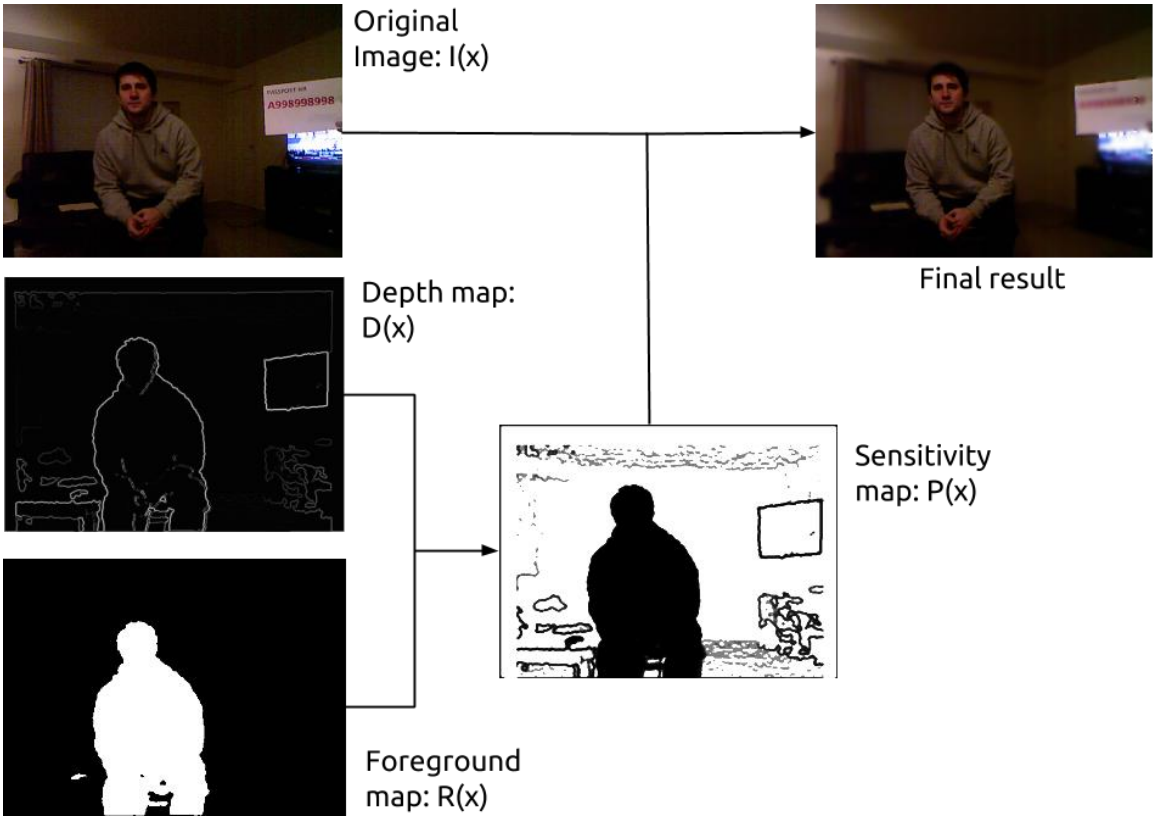


Figure 1. *The flowchart of the proposed method*

Note that, while using person-aware visual communication as the application context, our technology can be easily extended to other application scenarios, since both the sensitivity map and the inhomogeneous obscuration techniques are general and easy to accommodate different requirements of different privacy definitions.

CHAPTER 2

PRIVACY SCENARIOS

Person-aware visual communication is the primary application context that we explore in our research, however, one can easily see how the proposed framework can be easily extended to handle different application contexts as well. In this section, we are going to explore other privacy scenarios where our framework can be used, and for each privacy scenario we are going to specify which regions are considered to be potentially privacy-intrusive.

- *Video Conferencing*

  This is the primary scenario that we cover in our research. The non-private regions in this scenario are considered to be the humans in the scene and edges of detected objects. Edges of objects contribute a negligent amount of privacy protection if they were to be obfuscated. Video conferencing systems are widely adopted in major corporations, hospitals and universities for interactive meetings among participants from distant locations. This medium offers an evident advantage, however it raises some serious concerns regarding privacy. A scenario where doctors are having a video conference and confidential documents are lying around on the table is obviously violating the privacy rights to the patients whose names are on those documents. The system proposed in this paper will remove that privacy concern by obfuscating heavily intra-region. The doctors, in the

scenario we just mentioned, would still be visible and carrying on their video conference with the removed risk of exposing private documents.

- *Interactive Human Privatization*

  In this type of scenario, we mark the humans that we want to be private and we obscure their identities. Humans on the scene that were unmarked will be considered as part of the non-private region. The Kinect device can identify up to 6 human figures in a segmentation map. In this privacy scenario, we can specify the indices of human figures on the scene that we want be obfuscated. In Figure 2, even though there are 2 human figures on the scene, we interactively select only 1 human figure to be included in the foreground map $R(x)$. As long as the subjects on the scene do not leave the depth sensor's field of view the "player indices" are guaranteed to stay the same for the same player. If one of the "players" leaves the scene and comes back later, a random player index will be assigned to the player regardless of their previous "player index".

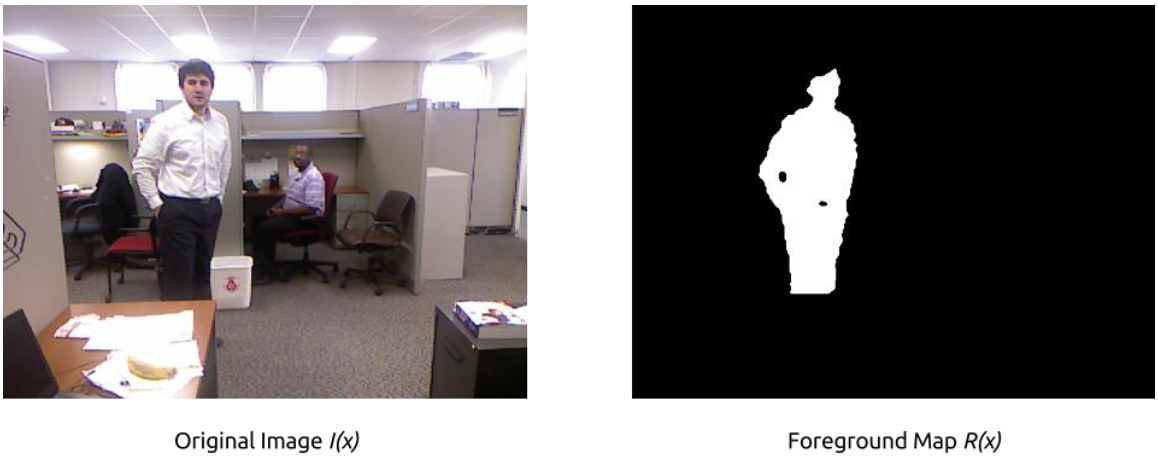


Figure 2. *Interactive human privatization*

- *Interactive Document Privatization*

  In this type of scenario, we mark the documents that we want to be private and obscure them so that they are not visible.

- *Street View*

  Achieve privacy with technologies that provide panoramic views from positions along streets in the world. Google for example, allows users themselves to flag inappropriate or sensitive imagery for Google to review and remove, however there is a window where potentially intrusive information is available to the public. The private sections in this type of scenario are: license plate numbers, street numbers, and identity of humans. Special accent must be put on people engaging in activities visible from public property in which they do not wish to be seen publicly. Figure 3 is an example of this type of scenario. The license plate of the van and the 2 human figures on the scene are obfuscated.

Figure 3. *Street view privacy scenario*

- *Document Preservation*

  An example for this scenario is an art gallery. In this scenario, the humans on the scene are privatized – we simply do not want to show the identity of viewers. The documents, in this case, the art pieces are visible on the scene.

CHAPTER 3

VISUAL PRIVACY PROTECTION USING DEPTH INFORMATION

3.1 Data Preparation

To get simultaneously depth and color images, the Microsoft Kinect Sensor is used in our study. Kinect provides a depth image camera and a functionality for real-time human-background disambiguation, allowing us to separate humans from the background in a cluttered environment with varying lightning conditions and camera angles. We start by retrieving a color image frame and a corresponding depth frame from the Kinect Sensor, both in the default resolution of 640X480 at 30 frames per second. The SDK provides methods for aligning the color frame and the depth frame. The SDK also performs human detection and returns the result as a player index map, where each pixel has an associated player index or 0 for non-player. Using the player indices, we are able to separate the background objects from human objects.

After some simple calibration steps using information provided by the sensor, we finally have three inputs from the Kinect sensor: color image $I(x)$, depth map $D(x)$ and player index map $Q(x)$.

3.2 Sensitivity Map

The sensitivity map, as described previously, should capture information from two sources: the foreground information and the privacy-irrelevant structure information.

We derive the foreground information from the player index map by treating all players (within a certain depth threshold) as foreground. This way, we create a binary region-of interest (ROI) map $R(x)$ defined as:

$$R(x) = \begin{cases} 1 & if\ Q(x) > 0\ and\ D(x) < \tau\, max_x\big(D(x)\big) \\ 0 & otherwise \end{cases}$$

where τ is threshold set to 0.5. For scene structure information, we measure it by measuring the depth change, which is captured by using the magnitude of the depth gradient, i.e. $\|\nabla D(x)\|$. Note that, theoretically, second order derivatives also reflect depth discontinuity. In practive, we find that the first order gradient performs very well already and therefore discard the second order statistics for computational efficiency. Combinging the two components, we define the sensitivity map $P(x)$, as follows:

$$P(x) \propto f\left(\frac{1 - R(x)}{\|\nabla D(x)\| + \varepsilon}\right)$$

where ε is a small positive number to avoid underflow, $f(.)$ is a sigmoid-like function to avoid extreme values. Figure 2 shows an example of a sensitivity map.

3.3 Inhomogeneous Privacy Obscuration

The idea to use sensitivity map for privacy protection is to encourage large obscurity in the areas of large sensitivity, while keep non-sensitive regions as un-altered as possible. We implement the idea using the inhomogeneous diffusion framework, and use the sensitivity map to guide the diffusive factor. In particular, we formulate the problem as the following diffusion process:

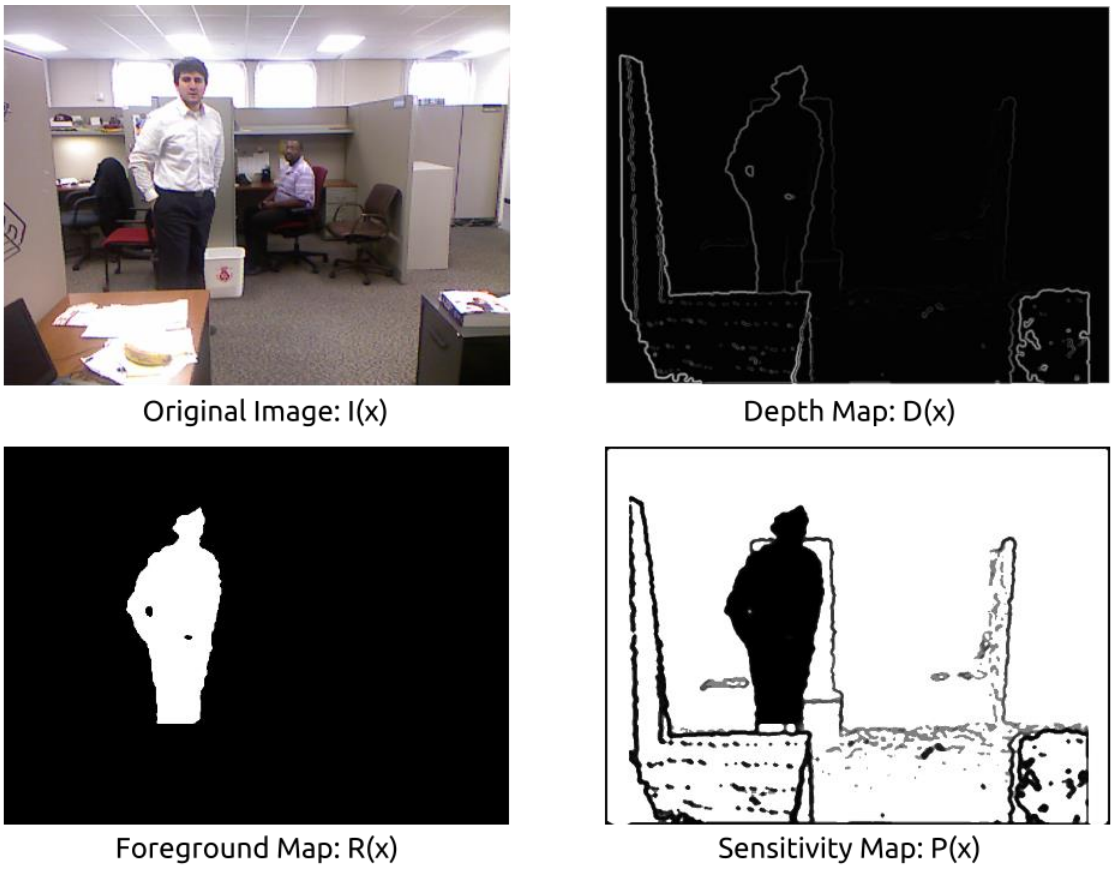


Figure 4. *Example of a sensitivity map*

$$\partial_t I(x,t) = div\left(g\big(I(x,t), P(x)\big)\,\nabla I(x,t)\right), t > 0$$

where $g(.)$ defines the diffusivity and t is the "artificial" time step for the diffusion process.

We base our implementation on the classical Perona-Malik diffusion [9] on each of the three channels of the color image (red, green, blue) in combination with the sensitivity map. In [9], two flux functions which will help limit the diffusion process to contiguous homogeneous regions and not cross region boundaries are introduced:

$$c(\|\nabla I\| = \exp\ (-\|\nabla I\|^2\ /K)$$

$$c(\|\nabla I\| = (1 + \|\nabla I\|^2/K)^{-1}$$

The first flux equation prefers high-contrast edges over low-contrast ones. The second flux function prefers wide regions over narrow ones. In our study, we are using the first flux function since it empirically demonstrates better performance. To combine the flux function with the sensitivity map $P(x)$, we define our sensitivity function as:

$$g\big(I(x,t), P(x)\big) = c(\|\nabla I\|)P(x)$$

For implementation, the original P-M diffusion iteratively updates the image and at iteration $t+1$ the numerical solution is:

$$I_{i,j}^{t+1} = I_{i,j}^{t} + \lambda\,[c_N\,.\nabla_N I + c_S\,.\nabla_S I + c_E\,.\nabla_E I + c_W\,.\nabla_W I]_{i.j}^{t}$$

where N, S, W and E correspond to the pixel above, below, to the left and to the right of the pixel that we are currently processing. The update parameter λ affects how much smoothing happens in each iteration. The algorithm can be stable if λ is in the range of $[0, 0.25]$ as suggested in [9].

To combine the sensitivity map $P(x)$, we revise the iteration as following:

$$I_{i,j}^{t+1} = I_{i,j}^{t} + \lambda\,P_{i.j}^{t}[c_N\,.\nabla_N I + c_S\,.\nabla_S I + c_E\,.\nabla_E I + c_W\,.\nabla_W I]_{i.j}^{t}$$

The resulting Perona-Malik diffusion of the original color image will have a smoothing factor of 0 when the pixel has a player index other than 0 associated with it, and a small value at the edges calculated based on the gradient of the distance map $D(x)$. Looking at the sensitivity map, we can infer knowledge about the granularity of regions. This way we preserve semantically meaningful boundaries, and we use those boundaries to achieve intra-region smoothing in homogeneous regions, succeeding in obscuring

potentially sensitive information in the background. In order to achieve high level of privacy and completely obscure the background we use 30 iterations and the smoothing factor in each iteration, λ, is set to 0.25, which when combined with $P(x)$ will be in the range 0 and 0.25.

CHAPTER 4

EXPERIMENTAL RESULTS

We use different scenarios to evaluate our approach to visual privacy protection. The proposed algorithm in this paper has been tested in various conditions, e.g., having multiple people on the scene, the person of interest being turned with the back towards the camera, scenarios where the people of interest are seating or standing as well as testing scenarios under different angles. The criteria for evaluating the quality of privacy boils down to the following two:

1. Level of protection of sensitive information
2. Level of clarity of information that is not considered sensitive (preservation of visual meaning of the scene)

Figure 5 shows a scenario with multiple people. This is an example of the first privacy scenario given in Chapter 2: *Video Conferencing.* We have two human figures on the scene, and in this scenario we want to leave the human figures unobscured. The people of interest are not necessarily in the ”foreground” depth-wise since there are objects that are closer. We can see from the sensitivity map that the foreground region that we want to maintain $R(x)$ is precisely identified, and that the edges based on depth gradient are properly identified as well. The last image in Figure 5 is the result of Perona-Malik diffusion in combination with the sensitivity map. Compared to Perona-Malik diffusion without using our sensitivity map, we can see that our results preserve humans and maintain the strength of the edges, while smoothing heavily inside the homogeneous regions. The information on the whiteboard is successfully obscured as well. Even though

the image in d) is obscured, we still have enough visual information to tell the meaning of the scene. The clarity of the regions in the image that are not considered private is at a high level and the obscuration level of the private regions is at a point where it is not possible to identify potentially sensitive information that in this case may be found on the whiteboard or on documents lying around on the desk. As we can see from c), even though we have a high level of obscuration of the scene, the clarity of the regions that we want to be preserved is not satisfactory.

Figure 6 shows the performance of our algorithm in a scenario with varying lightning conditions. Given that the semantic boundaries are calculated from the depth information, the boundaries are successfully retrieved regardless of the small variety of color in the scene. The ”player” region is successfully retrieved as well and the image that is protected using the sensitivity map demonstrates how potential sensitive information in the scene is successfully protected. In this case, we have a paper with information about a passport number. The sensitivity map demonstrates that the edges of the paper are preserved, while there is an equal level of intra-region smoothing as the image that has Perona-Malik diffusion performed on the image directly, with no information about the privacy level. Figure 6 is a clear demonstration not obfuscating the edges of detected objects contributes very little to the overall level of privacy protection, but a lot to the visual semantics of the scene. The edges of the TV and the document on the desk in the background are preserved using our framework, while at the same time we can also notice heavy intra-region smoothing on the TV display – it is not possible to identify the contents on the TV display. As we can see in b), the result of applying

Perona-Malik diffusion on the original image without taking into consideration the edges of objects, we destroy a lot of visual cues that help us construct the meaning of the scene. In d) all the potentially privacy-intrusive information is protected, but we also preserved a lot of the original visual cues that are not considered private and help us construct the meaning of the scene.

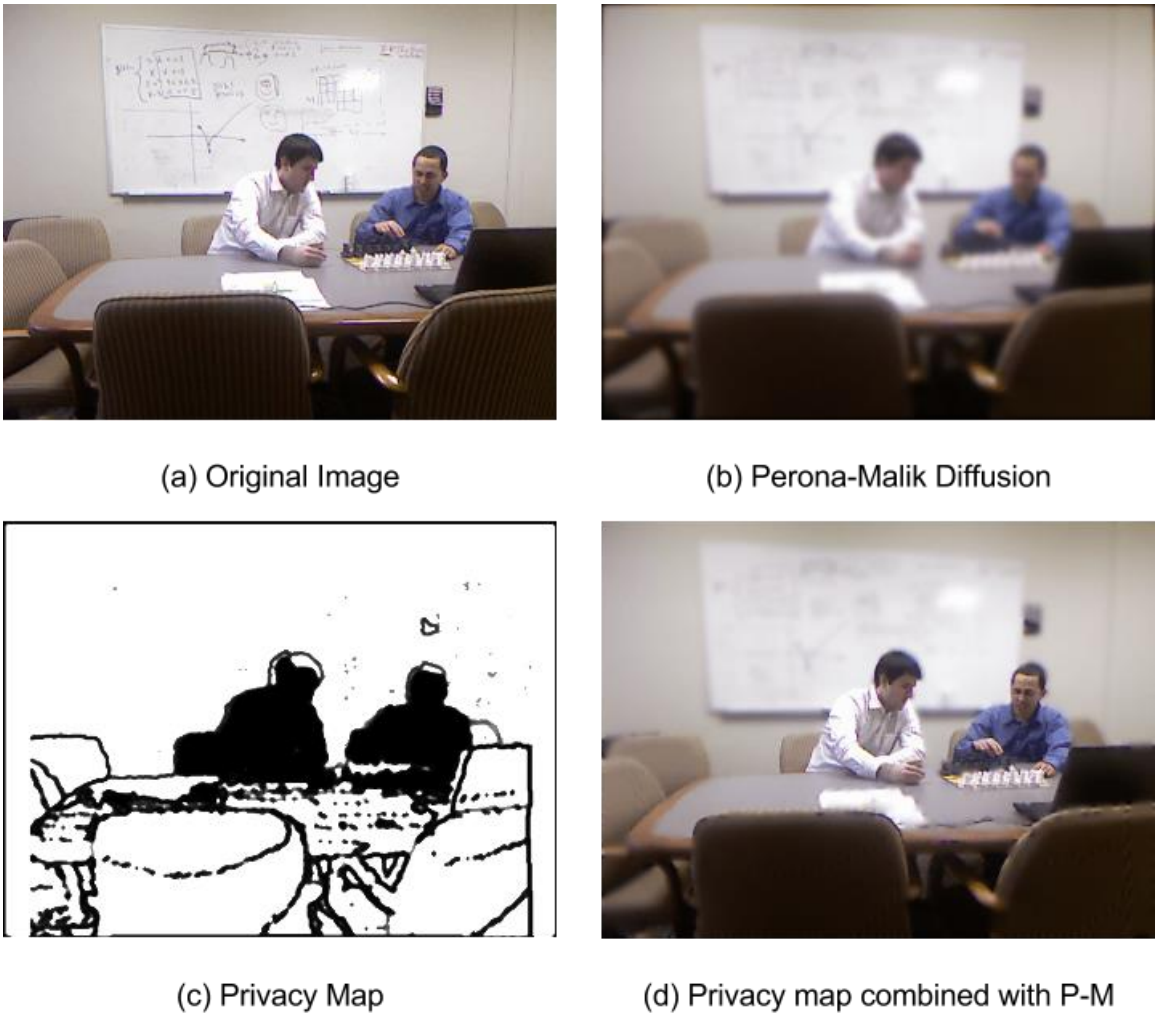


Figure 5. *Video conferencing privacy scenario*

Figure 7 shows a user standing sideways and we can see that the sensitiviy map is accurately calculated helping achieve good privacy protection when combined with Perona-Malik diffusion. The same is valid for Figure 8 as well.

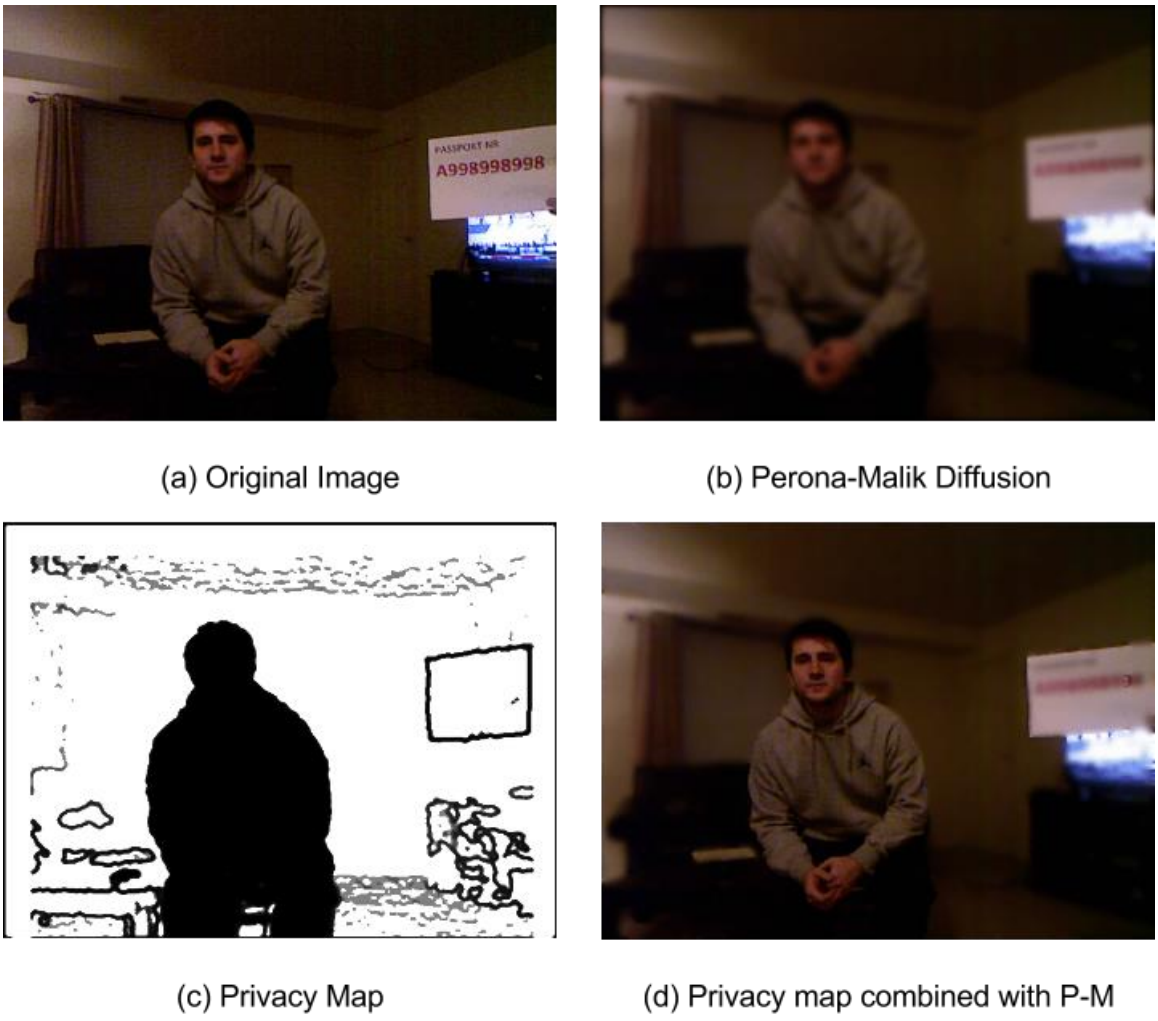
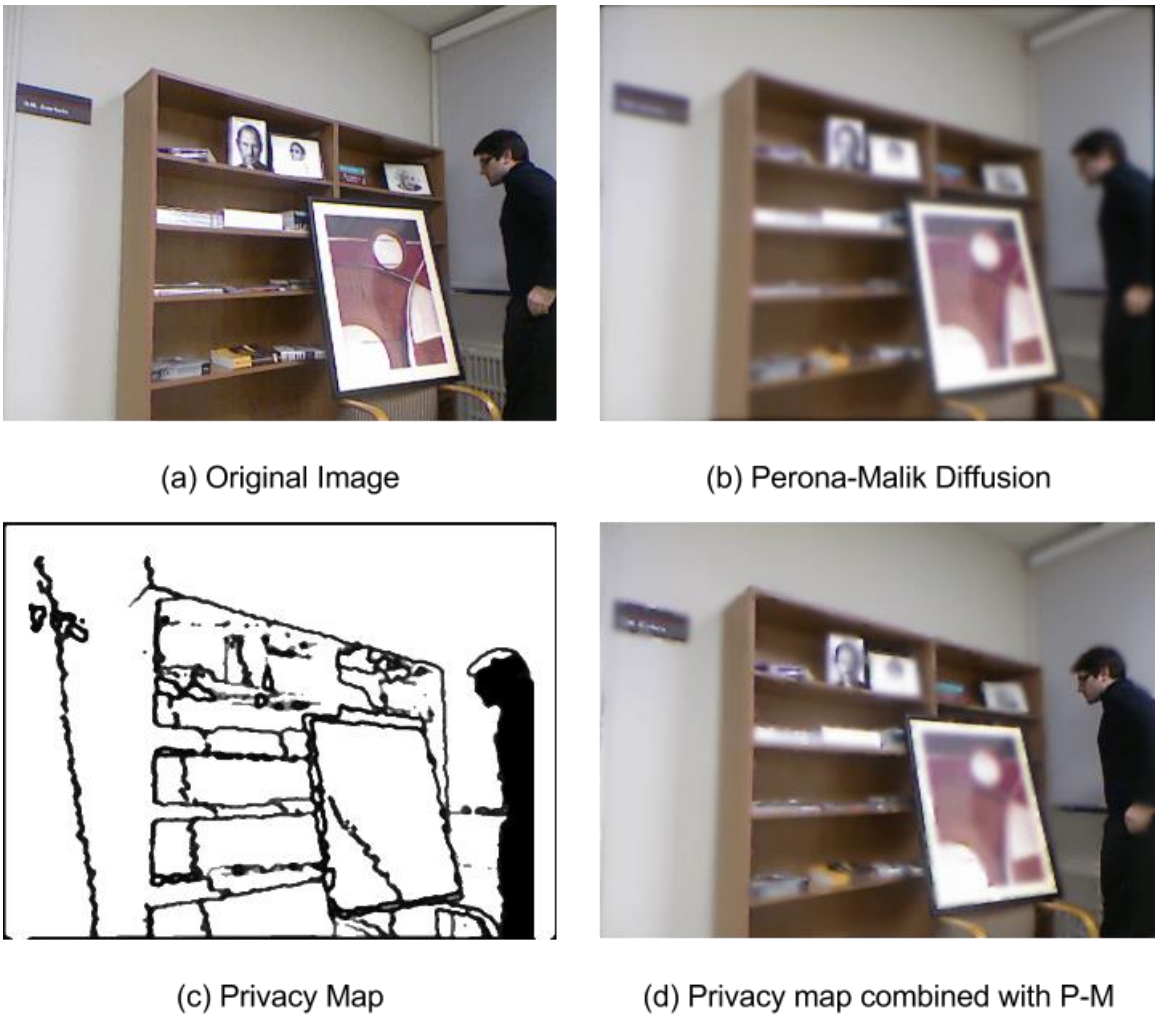


Figure 6. *Varying the lightning conditions*

Figure 7 and Figure 8 also demonstrate the performance of our algorithm from different distances. In Figure 7, the player is near the edge of the image and further away

in the scene. We can see from the sensitivity map that the edges are successfully detected and the player region as well. The human figure constitutes the foreground map $R(x)$ and we can see from d) that the "player" region in the foreground map is left unaltered. In addition, the edges from the books on the shelf and the painting in front of the shelf have a low privacy index associated with them leading to the edges being unaltered and the region within the object being altered heavily.



(a) Original Image

(b) Perona-Malik Diffusion

(c) Privacy Map

(d) Privacy map combined with P-M

Figure 7. *Subject further away from the sensor*

The resulting privacy protected image is a good example of an image that is successful at protecting the privacy, but also maintaining the visual semantics of the scene. Figure 7 demonstrates that the player is successfully retained even after being further away in the scene turned sideways towards the sensor.

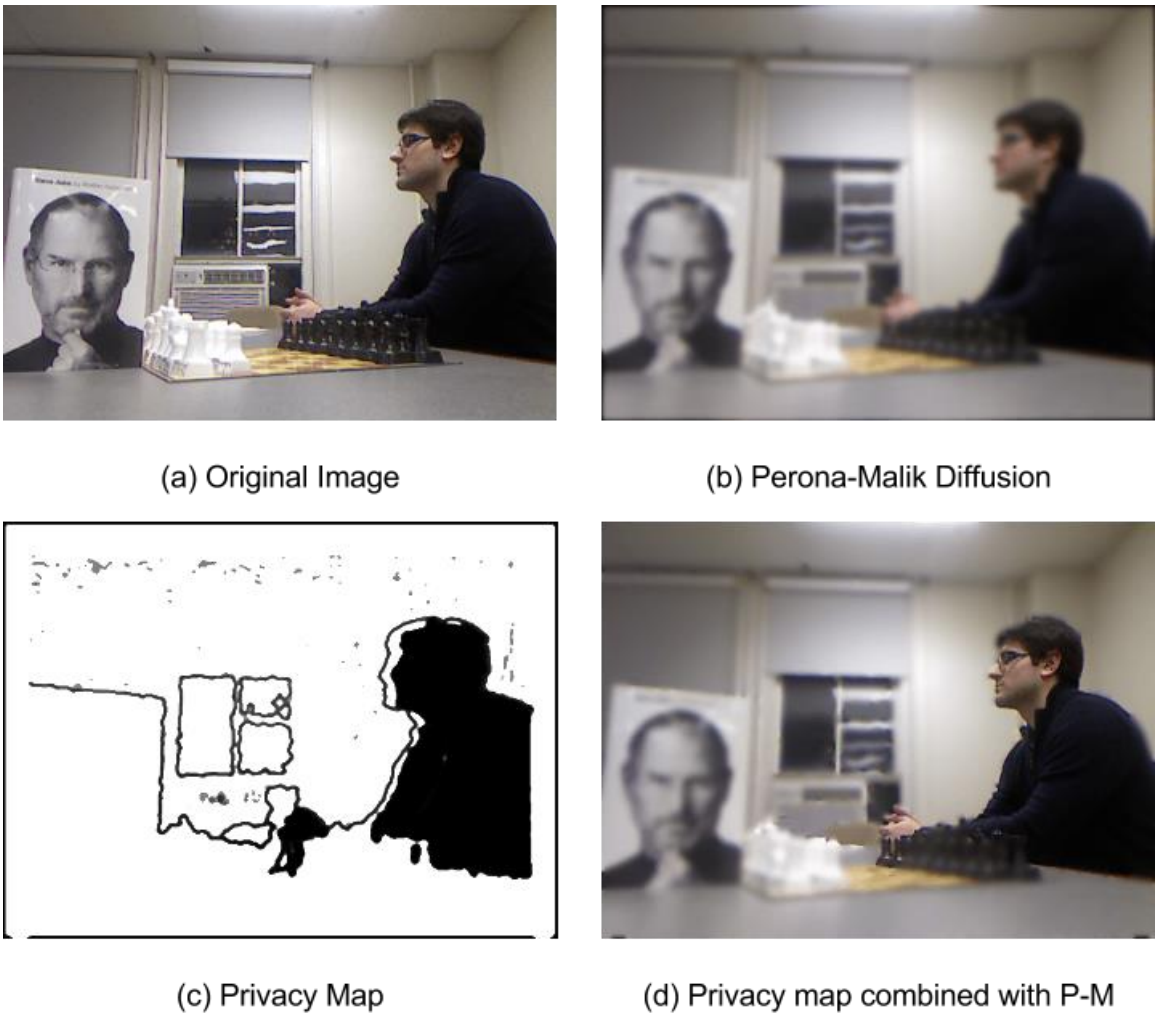


Figure 8. *Foreground region with subject being seated*

Figure 8 is a good example of demonstrating that even though the human figure is seated and only half of the human figure is visible on the scene, it is still identified as a

human figure by the Kinect sensor – leading to the whole “human” region having a minimum privacy index and being left unobfuscated. The edges of the book on the desk are successfully detected and left unaltered. We can see from a) that there is writing on the book cover however that can not be told from the book in d).

CHAPTER 5

CONCLUSION

In this study we presented a new privacy protection method for visual data using depth information acquired from depth sensors like the Kinect. We presented the potential privacy intrusion by various video surveillance systems and the importance for privacy protection in such systems. The depth information helps to distinguish non-private foreground regions and scene structures, through human detection and depth gradients. Such informationis used to create a sensitivity map to define per-pixel privacy levels. The map is then combined with an inhomogeneous diffusion process to achieve privacy protection. We evaluated the proposed approach using a set of RGBD images containing various image conditions and scenarios. The results show that our method successfully obscures potentially privacy-intrusive information, while at the same time maintaining the non-private visual features.

Further work can be performed in systematic evaluation of the quality of privacy of the proposed system and also evaluation of the level of clarity of information that is considered non-sensitive. We have identified the need for preserving visual features that are not considered private in order to construct the meaning of the scene. Different systems require that potentially private areas be detected dynamically in a real-time video system. We can also test with different depth sensors and explore more privacy scenarios. The main purpose of this project is to make it clear that once we have obtained a sensitivity map, we can use different detail-reduction techniques to achieve stronger levels of privacy. At the same time, the sensitivity map can help us distinguish which parts of the scene are not private, so that they can contribute to the visual meaning of the scene. We can also try different obfuscation techniques to achieve a level of privacy

protection to the point where the sensitive information is protected and the non-sensitive information is as preserved as possible.

REFERENCES CITED

1. P. Agrawal and P. Narayanan, "Person de-identification in videos," IEEE T. CSVT, 21(3):299–310, 2011.
2. D. Bitouk, N. Kumar, S. Dhillon, P. N. Belhumeur, and S. K. Nayar, "Face Swapping: Automatically Replacing Faces in Photographs," ACM SIGGRAPH, 2008.
3. T. E. Boult, "PICO: Privacy through invertible cryptographic obscuration," in Computer Vision for Interactive and Intelligent Environment, 2005.
4. Y. Chang, R. Yan, D. Chen, and J. Yang, "People identification with limited labels in privacy-protected video," ICME, 2006.
5. L. Du and H. Ling, "Preservative license plate de-identification for privacy protection," ICDAR, 2011.
6. 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," ICCV, 2009.
7. R. Gross, L. Sweeney, F. D. la Torre, and S. Baker, "Semi-supervised learning of multi-factor models for face de-identification," CVPR, 2008.
8. Martnez-Ballest, Antoni, "Towards a trustworthy privacy in pervasive video surveillance systems", Pervasive Computing and Communications Workshops, 2012.
9. P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion", IEEE T. PAMI, 12(7):629-639, 1990.

10. 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," IROS, 2007.
11. A. Senior, Protecting Privacy in Video Surveillance, in Privacy Protection in a Video Surveillance System, 2009.
12. Guardian Unlimited, "Google introduces face-blurring to protect protesters on YouTube", 2012.
13. Christopher Slobogin, Public Policy: Camera Surveillance of Public Places and the Right to Anonymity, 72 Miss. L.J. 213, 222 (2002)