2011 International Conference on Document Analysis and Recognition

# Preservative License Plate De-identification for Privacy Protection

Liang Du Haibin Ling
*Center for Data Analytics & Biomedical Informatics, Department of Computer & Information Science*
*Temple University, Philadelphia, PA, USA*
{*liang.du, hbling*}*@temple.edu*

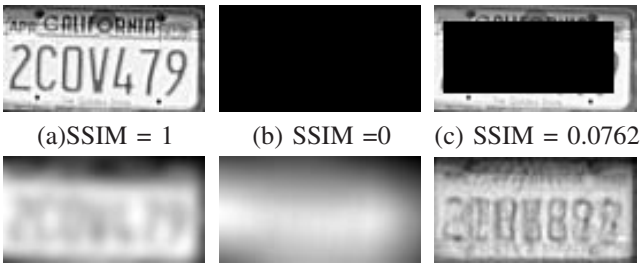
***Abstract*—Advances in imaging devices and web technologies have brought dramatic improvements in collecting, storing, and sharing images. The leakage of privacy information in the process becomes an important issue that has started drawing attention from both academia and industry. In this work, we study the problem of privacy preserving with focus on license plate number protecting in imagery. Specifically, we present a novel method for de-identifying license plate images with the least degradation in image visual quality for privacy protection. Unlike previous de-identification methods that pay little attention to the image quality preservation, our method, named *inhomogeneous principal component blur* (IPCB), adaptively blurs different pixels of a license plate by taking into account the prior distribution of sensitive information. We tested the proposed method on a public dataset in comparison with several popular de-identification methods. The evaluation shows that our method successfully de-identified the privacy information with the least damage of image quality when compared with several other solutions.**

***Keywords*-license plate de-identification, privacy protection, principal component analysis**

## I. Introduction

Advances in imaging devices and information technologies have brought dramatic improvements in collecting, storing, and sharing images. This makes exchange of information more convenient among different individuals and agencies. Web services like Google Streetview can systematically gather and share large scale images of public spaces. However, during this process, images of the public spaces in their original forms, typically contain sensitive information about individuals such as car license numbers on license plates. Publishing such data may violate individual privacy and bring security issues. Consequently, de-identification of such images is required [4].

Recently, image de-identification for privacy protection has drawn attentions from the academic community. Previous studies on visual privacy protection mainly focus on the fields of face de-identification and privacy-preserving surveillance. In [8], Newton et al. proposed a method called $k$-same for face de-identification in images and videos. Face features are extracted and averaged from $k$ nearest neighbors so as to ensure that there cannot exist any face recognition software for which a subject's $k$-samed image can be correctly recognized better than $1/k$ probability. In [1], Agrawal et al. proposed a person de-identification method in videos. In [4], Frome et al. presented a system for automatic detection and de-identification of faces and license plates in images. Both [1] and [4] focus on the detection of to-be de-identified subjects, while the de-identifications are simply done by blurring the detected locations. Chen et al. [3] studied the privacy preserving problem in the context of health care related surveillance. Chan et al. [2] proposed a method for counting peoples without explicit human detection. Upmanyu et al. [12] designed a secure video sharing system inspired by the Chinese Remainder Theorem that splits each frame into a set of random images. Schiff et al. [10] use markers in surveillance videos to detect persons whose identities are sensitive and thereafter hide such information by masking. A survey of privacy preserving video surveillance is given in [9]. The study of privacy protection is by no means limited to visual data. There are in fact pioneer works in the domain of data mining, data publishing, networking, etc. A survey can be found in [5].




(a)SSIM = 1 (b) SSIM =0 (c) SSIM = 0.0762
(d) SSIM = 0.1059 (e) SSIM = 0.2630 (f) SSIM =0.3904

Figure 1. Original and de-identified license plates with SSIM (Sec. II-B). (a) Original license plate without de-identification; (b)De-identification using a black mask to cover the plate; (c)De-identification using a black mask to cover sensitive area; (d)De-identification using disk blur with length 4; (e)De-identification using disk blur with filter radius 20; (f) The proposed de-identification using privacy prior based sparse representation.

Privacy information, by definition, is the knowledge that can be used to implicitly or explicitly reveal the identity of an individual [5]. For example, license plates that frequently appear in images of public spaces, are considered as containing sensitive information from the privacy point of view. However, not all information in the license plate region is privacy. Therefore, arbitrarily blurring or blocking some regions in the image can cause unwanted degradation of the original image. For example, in Fig.1(b), a black mask is used to cover the license plate. Such a solution, despite

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.100

protecting privacy, brings very unpleasant artifacts to image, which has a low *Structural Similarity Index* SSIM (defined in (2)) value. The method in (c), which only covers the image areas containing license numbers, performs slightly better than (b) in terms of SSIM, but still suffers from fairly unpleasant artifacts.

In this paper we propose a preservative de-identification method to balance privacy protection and quality preservation, with application to license plates. Unlike many previous studies that destroy information homogeneously in the region to be protected, we treat different pixels differently according to their importance. Intuitively, not all pixels inside a license plate are equally critical in terms of privacy information. For example, the state name is usually less sensitive than licence numbers. Integrating such importance prior with the statistical analysis, we propose an *inhomogeneous principal component blur* (IPCB) approach for license plate de-identification. We applied the proposed method to a public license plate dataset with a quantitative image quality measure. The effectiveness of IPCB is clearly demonstrated in comparison of several previously proposed methods.

In the rest of the paper, we first formulate our problem and introduce the image quality criterion used in our study. Then, in Section III, the proposed method for license plate de-identification is described. After that, the experimental validation of the proposed method on public dataset is presented in Section IV. Finally, we give discussion and conclusions in Section V.

## II. Problem Formulation

### A. Privacy Protection in License Plate Images

License plates (LP) exist in many images and videos of public spaces. The patches of license plates in images can be linked to information of a particular individual. Consequently, it is often desired to hide such information in public shared imagery [4]. Unlike in the previous work where the focus is on detecting LPs (and then simple blurring or blocking is applied), we focus on an effective and preservative de-identification. For this reason, we assume that an LP patch is already segmented and our algorithm is applied to such image directly.

Assume that $I \in \mathbb{R}^d$ (images are reshaped to vectors for convenience) is the original image which contains privacy information, and $\hat{I}$ is the de-identified image through de-identification function $\mathcal{D}$,

$$\hat{I} = \mathcal{D}(I) \ .$$

Privacy information can reveal or link to people's identity and we denote the identification process as $\mathcal{I}(I) : \mathbb{I} \to \mathbb{L}$, where $\mathbb{I}$ is the set of LPs and $\mathbb{L}$ the set of potential identities (license IDs). We have the following definition:

*Definition 1:* A de-identification function $\mathcal{D}$ is said to successfully de-identifying an LP image $I$ if and only if $\mathcal{I}(I) \neq \mathcal{I}(\mathcal{D}(I))$.

The identification function $\mathcal{I}$, in theory, can be either algorithms or human observers. In this study, for simplicity, we visually inspect the de-identification results. For example, in Fig.1, (c) will be considered as an unsuccessful de-identification since license numbers still can be recognized for the blurred image. To evaluate the effectiveness of a de-identification function $\mathcal{D}$, the average de-identification rate (DR) over the dataset $\mathbb{I}$ is needed:

$$\mathrm{DR} = \frac{|\{I \in \mathbb{I} : \mathcal{I}(I) \neq \mathcal{I}(\mathcal{D}(I))\}|}{|\mathbb{I}|} \tag{1}$$

where $|.|$ indicates the cardinality. In our experiments, a license plate is regarded as successfully de-identified if user cannot correctly recognize the at least three license numbers for the de-identified license plate. It is possible that one can recognize the car by appearance or the surroundings and link it to the identity of the owner. In this work, same as in [8], the inference attack of privacy is not taken into account.

### B. Preservative De-identification

A naive solution to achieve a high de-identification rate is to purely block the sensitive region, in our case, an LP. However, a successful de-identification does not necessarily lead to a good de-identification because it may also pollute unsensitive data in the original image (e.g. see Fig.1(b)). In theory, an optimal de-identification algorithm should keep unsensitive information untouched and modify only the privacy part of the data.

*Definition 2:* An *optimal de-identification* $\mathcal{D}^*$ of an LP image $I$ de-identifies the image at the minimum cost of image visual quality degradation, i.e.

$$\mathrm{sim}(I, \mathcal{D}^*(I)) = \max_{\mathcal{D}:\mathcal{I}(I)\neq\mathcal{I}(\mathcal{D}(I))} \mathrm{sim}(I, \mathcal{D}(I)) \ ,$$

for some image similarity measure $\mathrm{sim}(.,.)$.

To quantify perceptual image quality after de-identification, we need an image quality assessment $\mathrm{sim}(.,.)$. There are two popular choices: the mean squared error (MSE), which is computed by averaging the squared intensity differences of distorted and reference image pixels, and the related quantity of peak signal-to-noise ratio (PSNR). These two metrics are widely used because they are simple to calculate, have clear physical meanings, and are mathematically convenient in the context of optimization. However, they do not match very well to the perceived visual quality by humans [6]. Instead, we propose using the *Structural Similarity Index* (SSIM) [14] as an objective measurement for assessing perceptual image quality degradation after de-identification. Different from the traditional MSE based image quality measurement, SSIM is based on the degradation of structural information that human visual perception is highly adapted for.

This property is desirable for measuring visual quality degradation of images after de-identification.

SSIM consists of three components including luminance similarity $l(\mathbf{x}, \mathbf{y})$, contrast similarity $c(\mathbf{x}, \mathbf{y})$ and structural similarity $s(\mathbf{x}, \mathbf{y})$:

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y})^{\alpha} \cdot c(\mathbf{x}, \mathbf{y})^{\beta} \cdot s(\mathbf{x}, \mathbf{y})^{\gamma} \ , \qquad (2)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the reference image and the distorted image respectively; and

$$\begin{aligned} l(\mathbf{x}, \mathbf{y}) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \ , \\ c(\mathbf{x}, \mathbf{y}) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \ , \\ s(\mathbf{x}, \mathbf{y}) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \ , \end{aligned}$$

where $\mu_x$ and $\mu_y$ are the mean intensities of $\mathbf{x}$ and $\mathbf{y}$ respectively; $\sigma_x$, $\sigma_y$ and $\sigma_{xy}$ are standard deviations and the correlation coefficient of $\mathbf{x}$ and $\mathbf{y}$ respectively; and $C_1,\ C_2,\ C_3$ are constants. For a detailed explanation of SSIM, one can refer to [14].

There is a tradeoff between image quality (SSIM) and de-identification rate. For example, if we cover the license plate with a black mask, this kind of de-identification will be considered as a successful de-identification. However, the image quality after de-identification will be very poor. On the other hand, if we leave the license plate untouched, the image quality after this kind of 'de-identification' will be very high, however, the privacy is not preserved and it is not a successful de-identification.

## III. Inhomogeneous Principal Component Blur for Privacy Protection

The goal of this work is to de-identify license plates in images. Naturally, license plate detection should be regarded as the first procedure. However, since our focus is to develop an optimal de-identification of license plate and license plate detection is largely seen as a solved problem, the license plate detection is out of the scope of this paper.

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. It has many successful applications in image processing and computer vision. By reconstructing the original data with a limited number of eigenvectors, we can get a sparse representation of the original data. The basic idea of our de-identification method is that by representing the license plate by different number of eigenvectors according to their privacy prior, we can obtain the de-identification with the minimum cost of image degradation.

In the training phase, we follow the framework of PCA for face recognition [11]. For the sake of self completeness, we briefly repeat it here. Specifically, for an LP image set $\mathbb{I} = \{I_1, I_2, \ldots, I_T\}$, the mean plate is defined by $\bar{I} = \frac{1}{T}\sum_{i=1}^{T} I_i$. Then, by subtracting $\bar{I}$ we get a zero-mean image set $\{J_i = I_i - \bar{I} : i = 1, \ldots, T\}$, and form a zero mean matrix $A = [J_1 J_2 \ldots J_T] \in \mathbb{R}^{d \times T}$. By solving the eigenvectors of covariance matrix $C = AA^T$, a set of components (eigenvectors) $X_1, X_2, ..., X_m$ can be obtained and used for license plate representation in the following steps.

Given a license plate $I$, it can be reconstructed using the components as $I = \bar{I} + \sum_{i=0}^{m} c_i X_i$, where $c_i$ is the projection coefficient of $I$ on the $i$-th component. By using limited number of components in reconstruction, we can get a blurred version of $I$ where some details are discarded. This idea can be used for de-identification [7]. Specifically, denoting the procedure as $\mathcal{D}_{pca}$, we have

$$\hat{I} = \mathcal{D}_{pca}(I) = \bar{I} + \sum_{j=0}^{k} c_j X_j \ , \qquad (3)$$

where $k$ is the number of components chosen for preserving global appearance information and usually $k \ll m$. The method $\mathcal{D}_{pca}$, compared with blur-based de-identification methods (e.g., Gaussian and motion blurs), effectively preserves sensitive information while causes less artifacts. However, as a global solution, it does not distinguish pixels in $I$, which in practice often have various level of sensitive information. In the following we extend the solution to inhomogeneously address the problem.

First, we denote the distribution of sensitive information in LP images using a probability map $\mathbf{P} \in \mathbb{R}^d$, such that $\mathbf{P}(i)$ denotes the probability that pixel $i$ belongs to a license number or letter. In practice, $\mathbf{P}$ can be estimated from set $\mathbb{I}$. Then, for each pixel $I(i)$ in image $I$, $i = 1, \ldots, d$, it is de-identified individually using $k_i$ components such that

$$k_i = \lfloor k \times e^{-\mathbf{P}(i)} \rfloor \qquad (4)$$

depends on $\mathbf{P}(i)$. Consequently, our proposed method $\mathcal{D}_{ipcb}$, namely *inhomogeneous principal component blur* (IPCB), has the following form,

$$\hat{I}(i) = \mathcal{D}_{ipcb}(I)(i) = \bar{I}(i) + \sum_{j=0}^{k_i} c_j X_j(i) \ , \ i = 1, \ldots, d, \qquad (5)$$

where $(i)$ means the $i$-th element (pixel) in the corresponding vectors. Compared with Eqn. (3), the proposed method in Eqn. (5) allows treating each pixels individually. This way, the information in the insensitive region can be preserved. $\mathcal{D}_{pca}$ can also be viewed as a special case of $\mathcal{D}_{ipcb}$ with a uniform distribution $\mathbf{P}$.

Fig. 2 shows the schematic diagram of the proposed method. By using the prior based reconstruction of license plate, we can make less modification of the original data while de-identifying the privacy information. For instance, if a pixel in the license plate is very unlikely to contain

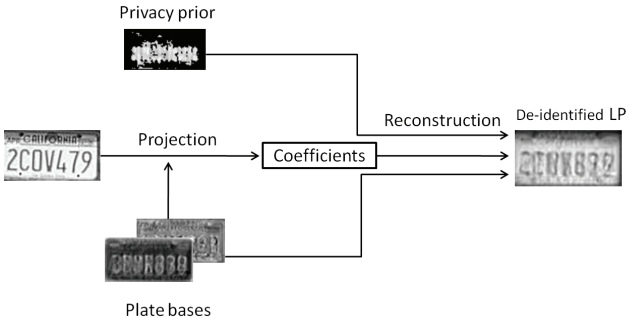



Figure 2. Preservative privacy protection through inhomogeneous principal component blur.

license numbers, e.g. in the boundary of the plate, the prior for it will be almost 0 and the number of reconstruction eigenvectors is inversely proportional to it according to (4) and (5). This prevents unnecessary modification of images, which is desirable for an optimal de-identification according to Definition 2.

In addition, from a cryptography point of view, the proposed method has superiority over traditional blur-based de-identification. There are numerous de-blurring and super-resolution methods, which jeopardize the reliability of blur-based de-identification. On the contrary, the proposed method is a sparse representation of the original license plate. A de-identified plate cannot be recovered without knowing the coefficients of each principal components.

## IV. Experiments

To evaluate the effectiveness of the proposed IPCB de-identification method, we apply it to the Caltech car dataset [1] for license plate de-identification. The dataset contains 126 car images in rear view. License plates in the dataset are clear and located roughly in the center of an image. As mentioned before, license plate detection is out of the scope of this work. Thus, all license plates are manually segmented from the original image and convert to gray level image.

We compared our algorithm with several previously commonly de-identification approaches including the disk blur, Gaussian blur, motion blur and PCA-based method (i.e., $\mathcal{D}_{pca}$). The evaluation is qualitative: we tune different methods with different degree of privacy protection and then check their performance when they achieve roughly same SSIMs (i.e., similar levels of image degradation). We observed that IPCB consistently outperforms other approaches. Fig. 3 shows some de-identification examples together with the SSIM values. In the figure, it is clear that in addition to achieving higher SSIMs, IPCB successfully obscures the license IDs.

In order to show the advantage of IPCB in preserving image quality while preserving privacy in context, we collected some license plates with different sizes and locations in images. Fig. 4 shows an example of such de-identified image. From the figure, we see that the degradation in the IPCB de-identified image is barely noticeable compared with the results by other approaches.

[1] http://www.vision.caltech.edu/Image_Datasets/cars_markus/

## V. Discussions and Conclusions

In this paper, we proposed a method for license plate (LP) de-identification. To protect the privacy information at minimum cost of image degradation, a novel method named inhomogeneous principal component blur (IPCB) is developed. IPCB takes into account the spatial distribution of privacy information to reduce the degradation in less privacy-sensitive regions. The experiments on a public car dataset show clearly the efficacy of the proposed method in comparison with several previously proposed methods. In the future, we plan to improve the privacy layout prior by using large LP datasets.

## Acknowledgment

The authors would like to thank Markus Weber for the Caltech Car dataset. This work is supported in part by NSF Grant IIS-1049032.

## References

[1] P. Agrawal and P. J. Narayanan. “Person De-identification in Videos”, in *ACCV*, 2009.
[2] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. “Privacy preserving crowd monitoring: Counting people without people models or tracking”, in *CVPR*, 2008.
[3] 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.
[4] 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.
[5] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. “Privacy-preserving data publishing: A survey of recent developments”, ACM Comput. Surv. 42(4), 2010.
[6] B. Girod. “What’s wrong with mean-squared error?”, in *Digital images and human vision*, A. B. Watson (Ed.). MIT Press, Cambridge, MA, USA ,207-220,1993.
[7] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. “Semi-supervised learning of multi-factor models for face de-identification”, in *CVPR*, 2008.
[8] E. M. Newton, L. Sweeney. “Bradley Malin: Preserving Privacy by De-Identifying Face Images”, *IEEE Trans. Knowl. Data Eng.* 17(2): 232-243 ,2005.
[9] A. Senior. “Protecting Privacy in Video Surveillance”, in *Privacy Protection in a Video Surveillance System*, 35– 47, 2009.
[10] J. Schiff, M. Meingast, D. K. Mulligan, S. Sastry, and K. Goldberg. “ Respectful Cameras: Detecting Visual Markers in Real-Time to Address Privacy Concerns”, in *IROS*, 2007.
[11] M.A. Turk, A.P. Pentland. “Face recognition using eigenfaces”, in *CVPR*, 1991.
[12] M. Upmanyu, A. Namboodiri, K. Srinathan, and C. Jawahar. “Efficient privacy preserving video surveillance”, in *ICCV*, 2009.


Figure 3. Results of different de-identification methods with its SSIM below each image.

(a) Black mask covering (b) Motion blur (c) IPCB (proposed)

Figure 4. An example of de-identified street view using different methods.

[13] T. Winkler and B. Rinner. "A systematic approach towards user-centric privacy and security for smart camera networks", in *ICDSC*, 2010.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity", *IEEE Trans. on Image Processing*, 13(4): 600-612, 2004.